Quantized GAN for Complex Music Generation from Dance Videos

Ye Zhu
Illinois Institute of Technology

Kyle Olszewski
Snap Inc.

Yu Wu
Princeton University

Panos Achlioptas
Snap Inc.

Menglei Chai
Snap Inc.

Jian Ren
Snap Inc.

Yan Yan
Illinois Institute of Technology

Sergey Tulyakov
Snap Inc.

Abstract

We present Dance2Music-GAN (D2M-GAN), a novel adversarial multi-modal framework that generates complex musical samples conditioned on dance videos. Our proposed framework takes dance video frames and human body motions as input, and learns to generate music samples that plausibly accompany the corresponding input. Unlike most existing conditional music generation works that generate specific types of mono-instrumental sounds using symbolic audio representations (e.g., MIDI), and that heavily rely on pre-defined musical synthesizers, in this work we generate dance music in complex styles (e.g., pop, breakdancing, etc.) by employing a Vector Quantized (VQ) audio representation, and leverage both its generality and high abstraction capacity of its symbolic and continuous counterparts. By performing an extensive set of experiments on multiple datasets, and following a comprehensive evaluation protocol, we assess the generative qualities of our proposal against alternatives. The attained quantitative results, which measure the music quality and consistency, beats correspondence, and music diversity, clearly demonstrate the effectiveness of our proposed method.[1]

1. Introduction

Although seemingly intuitive, music generation from dance videos has been a challenging task compared to its counterpart in the inverse direction (i.e., dance generation from music) due to two main reasons. First, typical audio music signals are high-dimensional and require sophisticated temporal correlations for overall coherence [2, 10]. For example, CD-quality audio has a typical sampling rate of 44.1 kHz, resulting in over 2.5 million data points ("dimensions") for a one-minute musical piece [3]. In contrast, most dance generation works output the relatively low-dimensional motion data in the form of 2D or 3D skeleton keypoint (e.g., displacement for dozens of joints) conditioned on the music [13, 14, 20, 22], which are then rendered into dance sequences and videos. To tackle the challenge of the high dimensionality of audio data, the research studies on music generation from visual input [7, 8, 23] often rely on the low-dimensional intermediate symbolic audio representations (e.g., 1D piano-roll or 2D MIDI). The symbolic representations benefit existing learning frameworks with a more explicit audio-visual correlation mapping and more stable training, as well as widely-established music synthesizers for decoding the intermediate representations. However, such symbolic-based works suffer from the limitations on the flexibility of the generated music, which brings us to the second challenge of dance video conditioned music generation. Specifically, a separately trained model is usually required for each instrument and the generated music is composed with acoustic sounds from a single predefined instrument [5, 7, 17]. Consequently, the typical resulting music is simple and lacks harmony and richness for accompanying real-world dance videos (e.g., you can watch a person dancing hip-hop with such piano-based generated samples in our supplementary videos). These facts make existing conditional music generation works difficult to generalize in complex musical styles and real-world scenarios.

To fill this gap, we propose a novel adversarial multi-modal framework that learns to generate complex musical samples from dance videos via the Vector Quantized audio representations. Inspired by the recent successes of VQ-VAE [3, 16, 19] and VQ-GAN [6], we adopt quantized vectors as our intermediate audio representation, and leverage both their increased abstraction ability compared to continuous raw audio signals, as well as their flexibility of better representing complex real-world music compared to classic symbolic representations. Specifically, our framework takes the visual frames and dance motions as input (Figure **??***), which are encoded and fused to generate the correspond-*

[1] See samples at https://l-yezhu.github.io/D2M-GAN/

ing audio VQ representations. After a lookup process of the generated VQ representations in a learned "codebook", the retrieved codebook entries are decoded back to the raw audio domains using a fine-tuned decoder from JukeBox [3]. Additionally, we deploy a convolution-based backbone and follow a hierarchical structure with two separate abstraction levels (i.e., different hop-lengths) for the audio signals. The higher-level model has a larger hop-length and fewer parameters, resulting in faster inference. In contrast, the lower-level model has a lower abstraction level with smaller hop-length, which enables the generation of music with higher fidelity and better quality.

2. Method

An overview of the architecture of the proposed D2M-GAN is shown in Figure 1. Our approach entails a hierarchical structure with two levels of models that are independently trained with a similar pipeline. For each level, the model consists of four modules: the motion module, the visual module, the VQ module consisting of a VQ generator and the multi-scale discriminators, and the music synthesizer. Our hierarchical structure amplifies the flexibility to choose between the trade-off of the music quality and computational costs according to practical application scenarios. A detailed description of these modules is given below while further architectural details and model-selection-tuning are included in the supplementary.

2.1. Generator

The generator $G = \{G_m, G_v, G_{\mathrm{vq}}\}$ includes the motion module G_m, the visual module G_v, and the principal VQ generator G_{vq} in the VQ module, which takes the fused motion-visual data as input and outputs the desired VQ audio representations.

$$f_{\mathrm{vq}} = G_{\mathrm{vq}}(G_m(x_m), G_v(x_v)) = G(x_m, x_v), \tag{1}$$

where x_m and x_v represent the motion and visual input data, respectively. f_{vq} is the output VQ representations. All these modules are implemented as convolution-based feed-forward networks. For the principal VQ generator, we use leaky rectified activation functions [26] for its hidden layers and a tanh activation for its last layer before output to promote the stability of GAN-based training [18].

It is also worth noting that we find that using batch normalization and the aforementioned activation function designs [15, 18, 21] is crucial for a stable GAN training in our framework. However, the application of the tanh activation will also restrict the output VQ representations within the data range between -1 and $+1$. We choose to scale activation after the last tanh activation by multiplying by a factor σ. The hyper-parameter σ enlarges the data range of VQ output and makes it possible to perform the lookup of pre-learned large-scale codebooks $\mathrm{LookUp}(f'_{\mathrm{vq}})$ with $f'_{\mathrm{vq}} = \sigma f_{\mathrm{vq}}$. Another significant observation regarding the generator's design is using a wide receptive field. Music has long temporal dependencies and correlations compared to images, therefore, the principal VQ generator with a larger receptive field is beneficial for generating music samples with better quality, which is consistent with the findings from previous works [4, 12]. To this end, we design our generator with relatively large kernel sizes in the convolutional layers, and we also add residual blocks with dilations after the convolutional layers. All previously described submodules within our generator G are jointly optimized.

2.2. Multi-Scale Discriminator

Similar to the generator, the discriminator in the D2M-GAN is also expected to capture the long-term dependencies of musical signals encoded in the generated sequence of VQ features. However, different from the generator design that focuses on increasing the receptive fields of the neural networks, we address this problem in the discriminator design by using a multi-scale architecture. The multi-scale discriminator design has been studied in previous works within the field of audio synthesis and generation [11, 12, 25].

The discriminator $D = \{D_1, D_2, D_3\}$ in the VQ module of our D2M-GAN is composed of 3 discriminators that operate on the sequence of generated VQ representations and its downsampled features by a factor of 2 and 4, respectively. Specifically, different from the multi-scale discriminators proposed in previous works that directly take the raw audio as input, we reshape the VQ representations f'_{vq} along the temporal dimension before feeding them into the discriminators, which is also important for D2M-GAN *to reach a stable adversarial training since music is a temporal audio sequence. Finally, we use the window-based objectives [12] (Markovian window-based discriminator analog to image patches in [9]). Instead of learning to distinguish the distributions between two entire sequences, window-based objective learns to classify between distributions of small chunks of VQ sequences to further enhance the overall coherence.*

3. Experiments

3.1. Experimental Setup

Datasets. *We validate the effectiveness of our method by conducting experiments on the AIST++ [14] dataset. The AIST++ dataset [14] is a subset of AIST dataset [24] with 3D motion annotations. We adopt the official cross-modality data splits for training, validation, and testing, where the videos are divided without overlapping musical pieces between the training and the validation/testing sets. The number of videos in each split is 980, 20, and 20, respectively. The videos from this dataset are filmed in pro-*

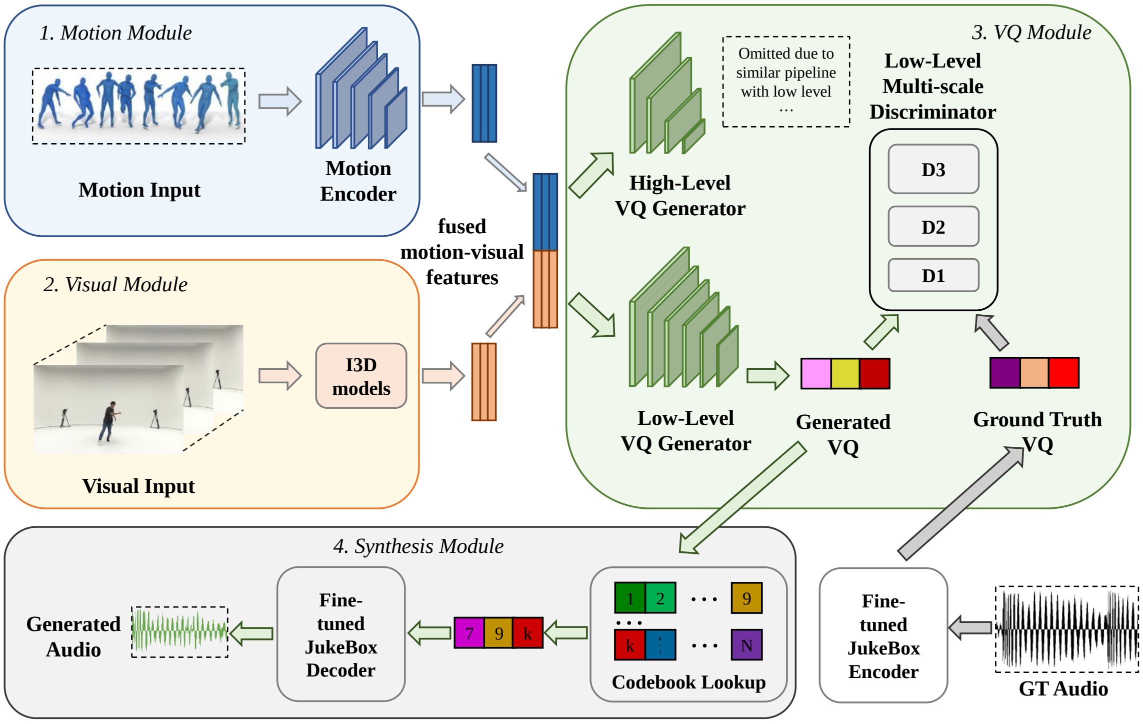

Figure 1. **Overview of the proposed architecture of the *D2M-GAN***. Our model takes the motion and visual data from the dance videos as input and process them with the motion and visual modules, respectively. It then forwards the fused representation containing information from both modalities to ground the generation of audio VQ-based representations with the VQ module. The resulting features are calibrated by a multi-scale GAN-based discriminator and are used to perform a *lookup* in the pre-learned codebook. Last, the retrieved codebook entries are decoded to raw musical samples via by a pre-trained and fine-tuned decoder, responsible for synthesizing music.

Category	Features	Type	Metric	Methods	Scores
Dance-Music	Rhythm	Obj.	Beats Coverage & Beats Hit	Dance2Music [1]	83.5 & 82.4
				Foley Music [7]	74.1 & 69.4
				Ours High-level	88.2 & 84.7
				Ours Low-level	**92.3 & 91.7**
Dance-Music	Genre & Diversity	Obj.	Genre Accuracy	Dance2Music [1]	7.0
				Foley Music [7]	8.1
				Ours High-level	24.4
				Ours Low-level	**26.7**
Dance-Music	Coherence	Subj.	Mean Opinion Scores	Random JukeBox [3]	2.1
				Dance2Music [1]	2.9
				Foley Music [7]	2.8
				Ours High-level	3.5
				Ours Low-level	3.4
				GT	**4.4**
Music	Overall quality	Subj.	Mean Opinion Scores	JukeBox [3]	3.4
				Ours High-level	3.1
				Ours Low-level	3.7
				GT	**4.8**

Table 1. Evaluation protocol and the corresponding results for the experiments on the AIST++ dataset [14]. *Obj.* stands for *Objective*, which means the scores are automatically calculated. *Subj.* stands for *Subjective*, which means the scores are given by human evaluators.

fessional studios with clean backgrounds. There are in total 10 different dance genres and corresponding music styles, which include breakdancing, pop, lock and etc. The number of total songs is 60, with 6 songs for each type of music. We use this dataset for the main experiments and evaluations.

Comparisons. *We compare our proposed method with several baselines. Ground Truth: GT samples are the original music from dance videos. Foley Music [7]: music samples generated using the Foley Music system. Foley Music model generates MIDI musical representations based on keypoints motion data and then converts the MIDI back to raw waveform using a pre-defined MIDI synthesizer. Specifically, the MIDI audio representation is unique for each musical instrument, and therefore the Foley music model can only generate musical samples with mono-instrumental sound. Dance2Music [1]: music samples generated using the online approach proposed in [1]. Similar to [7], the generated music is monotonic in terms of the musical instrument. JukeBox [3]: music samples generated or reconstructed via the JukeBox model.*

We observe in Table 1 that the genre accuracy scores of our D2M-GAN *are considerably higher compared to the competing methods. This is due to the reason that the competing methods rely on MIDI events as audio representations, which require a specific synthesizer for each instrument, and thus can only generate music samples with mono-instrumental sound. In contrast, our generated VQ audio representations can represent complex dance music similar to the input music types, which helps to increase the diversity of the generated music samples. It also makes the generated samples to be more harmonious with the dance videos compared to acoustic instrumental sounds from [1, 7], as shown in the next evaluation protocol for the coherence test.*

Overall Quality. *Finally, we look at the general sound quality of the generated samples by conducting the subjective MOS tests similar to the coherence evaluation, where the human testers are asked to give a score between* 1 *to* 5 *for the general quality of the music samples. During this test, only audio signals are played to the testers. The JukeBox samples are obtained by directly feeding the GT samples as input. The MOS tests show that our* D2M-GAN *is able to generate music sample with plausible sound quality comparable to the JukeBox model.*

4. Conclusion

We propose D2M-GAN *framework for complex music generation from dance videos via the VQ audio representations. Extensive experiments on multiple datasets, and comprehensive evaluations in terms of various musical characteristics prove the effectiveness of our method.*

References

[1] *G. Aggarwal and D. Parikh. Dance2music: Automatic dance-driven music generation.* arXiv preprint arXiv:2107.06252, *2021.* 3, 4

[2] *J.-P. Briot, G. Hadjeres, and F.-D. Pachet.* Deep learning techniques for music generation, *volume 1. Springer, 2020.* 1

[3] *P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever. Jukebox: A generative model for music.* arXiv preprint arXiv:2005.00341, *2020.* 1, 2, 3, 4

[4] *C. Donahue, J. McAuley, and M. Puckette. Adversarial audio synthesis. In* ICLR, *2019.* 2

[5] *H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In* AAAI, *2018.* 1

[6] *P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In* CVPR, *2021.* 1

[7] *C. Gan, D. Huang, P. Chen, J. B. Tenenbaum, and A. Torralba. Foley music: Learning to generate music from videos. In* ECCV, *2020.* 1, 3, 4

[8] *C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck. Music transformer: Generating music with long-term structure. In* ICLR, *2019.* 1

[9] *P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In* CVPR, *2017.* 2

[10] *S. Ji, J. Luo, and X. Yang. A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions.* arXiv preprint arXiv:2011.06801, *2020.* 1

[11] *J. Kong, J. Kim, and J. Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In* NeurIPS, *2020.* 2

[12] *K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In* NeurIPS, *2019.* 2

[13] *H.-Y. Lee, X. Yang, M.-Y. Liu, T.-C. Wang, Y.-D. Lu, M.-H. Yang, and J. Kautz. Dancing to music. In* NeurIPS, *2019.* 1

[14] *R. Li, S. Yang, D. A. Ross, and A. Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In* ICCV, *2021.* 1, 2, 3

[15] *M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet. Are gans created equal? a large-scale study. In* NeurIPS, *2018.* 2

[16] *A. v. d. Oord, O. Vinyals, and K. Kavukcuoglu. Neural discrete representation learning. In* NeurIPS, *2017.* 1

[17] *S. Oore, I. Simon, S. Dieleman, D. Eck, and K. Simonyan. This time with feeling: Learning expressive musical performance.* Neural Computing and Applications, *pages 955–967, 2020.* 1

[18] *A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks.* arXiv preprint arXiv:1511.06434, *2015.* 2

[19] *A. Razavi, A. van den Oord, and O. Vinyals. Generating diverse high-fidelity images with vq-vae-2. In* NeurIPS*, 2019.* *1*

[20] *X. Ren, H. Li, Z. Huang, and Q. Chen. Self-supervised dance video synthesis conditioned on music. In* ACM MM*, 2020.* *1*

[21] *T. Salimans and D. P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In* NeurIPS*, 2016.* *2*

[22] *E. Shlizerman, L. Dery, H. Schoen, and I. Kemelmacher-Shlizerman. Audio to body dynamics. In* CVPR*, 2018.* *1*

[23] *K. Su, X. Liu, and E. Shlizerman. Audeo: Audio generation for a silent performance video. In* NeurIPS*, 2020.* *1*

[24] *S. Tsuchida, S. Fukayama, M. Hamasaki, and M. Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In* Proceedings of the 20th International Society for Music Information Retrieval Conference, (ISMIR)*, 2019.* *2*

[25] *T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In* CVPR*, 2018.* *2*

[26] *B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network.* arXiv preprint arXiv:1505.00853*, 2015.* *2*