Scalable Group Choreography via Variational Phase Manifold Learning Supplementary Material

Nhat Le[1], Khoa Do[2,3], Xuan Bui[2,3], Tuong Do[1,4], Erman Tjiputra[1], Quang D.Tran[1], and Anh Nguyen[4]

[1] AIOZ, Singapore
[2] University of Science, Ho Chi Minh City, Vietnam
[3] Vietnam National University, Ho Chi Minh City, Vietnam
[4] University of Liverpool, Liverpool L69 3BX, United Kingdom

1 Architecture Details

The detailed network architecture of our proposed Phase-conditioned Dance VAE (PDVAE) is provided in Figure 1. Particularly, all Transformer modules consist of 5 blocks with 512 dimensions for keys, queries, and values, using 8 heads for multi-head attention [8] and feedforward layers with hidden dimensions set to 512, along with a dropout rate of 0.1 to mitigate overfitting. 1D convolutional layers with a kernel size of 31 and a padding size of 15 are employed to encode the local relationship between motion frames, which corresponds to a 1-second receptive field of the motion data. Additionally, the MLP phase predictor comprises two hidden layers (512, 256) with Siren activations [7] to effectively capture the periodicity of the movements. We elaborately design our dance generator architecture to capture the intricate long-range temporal dependencies and local phase timing information inherent in dance movements, facilitating the generation of high-fidelity and temporally consistent dance sequences.

1.1 Encoder

The Encoder takes both motion and music sequences as input, and produces a distribution over possible latent phase variables capturing the cross-modal relationship between them, which is parameterized by a Gaussian distribution. We implement the Encoder following the Transformer Decoder architecture where the Cross-Attention mechanism [8] is utilized to learn the relationship between the motion and the music. In particular, the input motion $\mathbf{x}$ is treated as the query while the audio $\mathbf{a}$ is projected into the corresponding key and value. We then calculate their cross-attention as follows:

$$\mathbf{Q}_e = \mathbf{W}_e^Q \mathbf{x}, \quad \mathbf{K}_e = \mathbf{W}_e^K \mathbf{a}, \quad \mathbf{V}_e = \mathbf{W}_e^V \mathbf{a}, \tag{1}$$

$$\text{CrossAtt}(\mathbf{x}, \mathbf{a}, \mathbf{a}) = \text{softmax}\left(\frac{\mathbf{Q}_e \mathbf{K}_e^\top}{\sqrt{d_k}}\right)\mathbf{V}_e, \tag{2}$$

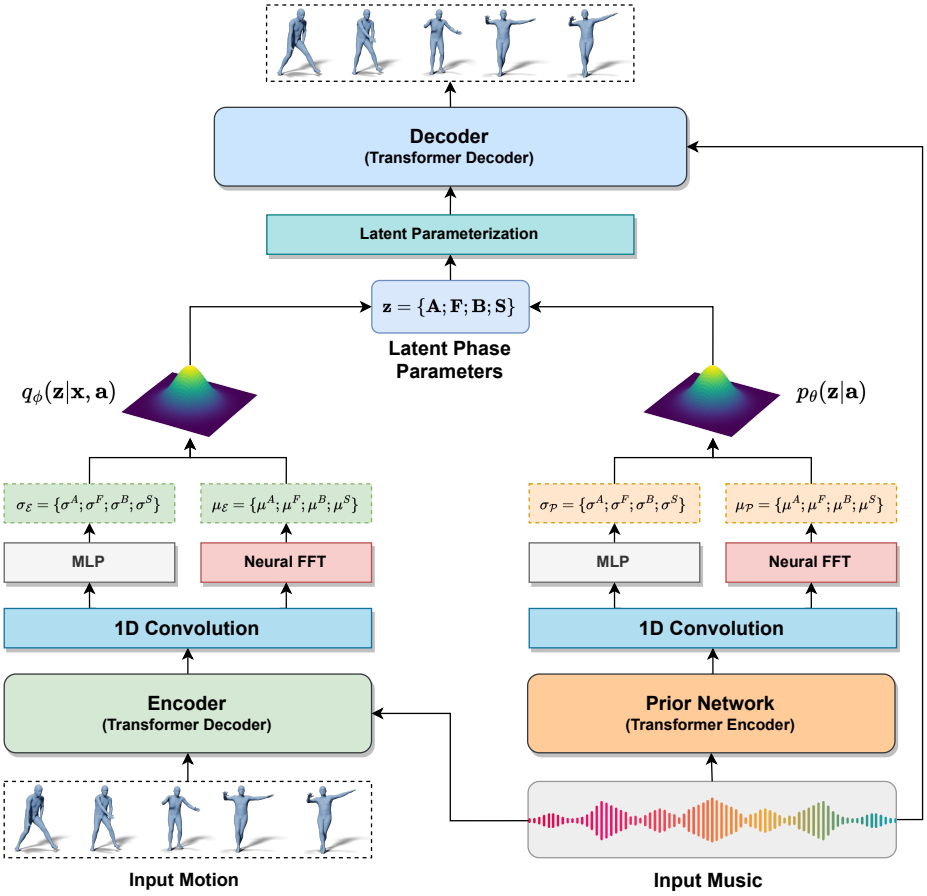

Fig. 1: Detailed network architecture of Phase-conditioned Dance VAE (PDVAE).

where $\mathbf{W}_e^Q, \mathbf{W}_e^K \in \mathbb{R}^{d \times d_k}$ and $\mathbf{W}_e^V \in \mathbb{R}^{d \times d_v}$ are learnable projection matrices to map the inputs into query, key, and value, respectively. Additionally, since the transformer attention mechanism mainly models the long-range dependencies with global receptive field, we further adopt 1D temporal convolutional layers stacking on top of the transformer to capture the local relationship and periodicity between motion frames. Subsequently, the neural FFT (Fast Fourier Transform) layer is applied to extract the mean embeddings in the frequency domain given the encoded latent curves, whereas an MLP with two layers is utilized to predict the variance of the phase parameters. Those predictions form a latent Gaussian distribution over the possible phase parameters that can well describe the essential information of a specific group dance. We then sample a set of phase parameters $\mathbf{z} = \{\mathbf{A}; \mathbf{F}; \mathbf{B}; \mathbf{S}\}$ from this distribution and compute the periodic parameterizations for decoding the original motion signals.

1.2 Prior Network

Since only the music is available for inference, we need to learn a prior to match the posterior distribution of the motion to draw new samples. Instead of imposing the prior distribution to be standard normal as in conventional VAE [3], we relate it to the music condition. Learning the conditional prior significantly improves the ability of the CVAE to generalize to diverse types of music and motion. Therefore, our conditional prior is defined as:

$$p_\theta(\mathbf{z}|\mathbf{a}) = \mathcal{N}(\mathbf{z}; \mu_\mathcal{P}, \sigma_\mathcal{P}) \tag{3}$$

where a Transformer model is used to encode the input conditioning music sequence and predict the corresponding μ_θ and Σ_θ. We implement the Prior network akin to the Encoder network, however, here we use self-attention mechanism (Transformer Encoder) [8] to capture the multi-scale rhythmic patterns and global music context.

$$\mathbf{Q}_p = \mathbf{W}_p^Q \mathbf{a}, \quad \mathbf{K}_p = \mathbf{W}_p^K \mathbf{a}, \quad \mathbf{V}_p = \mathbf{W}_p^V \mathbf{a}, \tag{4}$$

$$\text{SelfAtt}(\mathbf{a}, \mathbf{a}, \mathbf{a}) = \text{softmax}\left(\frac{\mathbf{Q}_p \mathbf{K}_p^\top}{\sqrt{d_k}}\right)\mathbf{V}_p, \tag{5}$$

where $\mathbf{W}_p^Q, \mathbf{W}_p^K \in \mathbb{R}^{d \times d_k}$ and $\mathbf{W}_p^V \in \mathbb{R}^{d \times d_v}$ are learnable projection matrices. Similarly, we also adopt temporal convolution, neural FFT, and MLP layers to predict the variational phase manifold from the music sequence that matches the posterior latent.

1.3 Decoder

Conditioned on the periodic latent paramerizations, the Decoder aims to produce human motion for a given music duration in one go (i.e., non-autoregressively). This architecture design is more suitable for our spatial-temporal phase representation as it does not suffer from the error accumulation problem (i.e., the prediction error accumulates over time since the current-frame outputs are used as inputs to the next frame), thus can generate natural dance sequences with high overall temporal alignments. Our Decoder is also based on Transformer Decoder architecture that takes a single latent vector z, the music features a as inputs, as well as the periodic parameterization sequence $\hat{\mathbf{L}}$ to start decoding:

$$\begin{aligned} p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{a}) &= \mathcal{D}(\mathbf{z}, \mathbf{a}), \\ \mathcal{D}(\mathbf{z}, \mathbf{a}) &= \text{CrossAtt}(\hat{\mathbf{L}}, \mathbf{a}, \mathbf{a}) \\ \hat{\mathbf{L}} &= \mathbf{A} \cdot \sin(2\pi \cdot (\mathbf{F} \cdot \mathcal{T} - \mathbf{S})) + \mathbf{B} \end{aligned} \tag{6}$$

where CrossAtt is calculated similarly to Equation 1. Here, we also utilize cross-attention model where we consider the music features as key and value, and the latent periodic sequence as the query. Additionally, we introduce skip connections from the phase features to each block of the network, enhancing the influence

of the periodic signals. By utilizing these techniques, we can more effectively model the periodicity inherent in the motion data, which can improve our dance synthesis performance.

1.4 Global Motion Predictor

Following previous works [2,6,9], we presume that the character's global translation is conditioned on its local poses. We utilize a Transformer network to generate the 3D global translation τ of the root joint from the local joint rotations, positions, and velocities as inputs. To alleviate ambiguity in the output (e.g, same trajectories with different starting points are equally weighted), rather than directly outputting the root position, we opt to predict its velocity υ. The global translations τ can be obtained by integrating the trajectory over time as follows:

$$\tau_{t+1} = \tau_t + \upsilon_t \Delta_t \tag{7}$$

However, the integrating process may likely suffer from error accumulation, which leads to the trajectory easily getting off track. This problem is even more severe in the upward movement (vertical direction), where the character may gradually float into the air or under the floor. To avoid this phenomenon, we directly predict the height τ^h of the root joint, which is reasonable since it lies in a region bounded by the height of the character. Then, our global trajectory predictor is optimized with the following objective:

$$\begin{gathered} \mathcal{L}_{\text{global}} = \mathcal{L}_{\text{trans}} + \mathcal{L}_{\text{vel}} \\ \mathcal{L}_{\text{trans}} = \sum_{t=1}^{T} \|\tau_t - \hat{\tau}_t\|, \quad \mathcal{L}_{\text{vel}} = \sum_{t=1}^{T} \|\upsilon_t - \hat{\upsilon}_t\| \end{gathered} \tag{8}$$

Our Global Motion Predictor is based on a similar architecture to our Decoder, consisting of 3 Transformer Decoder blocks. However, here we use the generated local motion from Decoder as the input sequences, along with the corresponding phase features for cross-attention to be applied.

To avoid the intersection problem, the predictor is also conditioned on the neighboring trajectories from other previously generated dancers in the scene. Specifically for crowded scene, we identify the trajectories of neighbor dancers by using K-d tree clustering of the 3D positions [1]. The trajectories are then flattened and processed by a 5-layer MLP each neighbor, yielding feature vectors of size 512 that summarize the trajectory context. Subsequently, those feature vectors are concatenated with the phase features and utilized as the conditions for cross-attention. In addition, the starting position τ_0 can either be controlled by manual user inputs or randomly sampled by fitting a Gaussian Mixture Model (GMM) over possible starting points in the training data.

2 Phase Manifold Visualization

We visualize the 2D PCA embedding of the phase manifold as depicted in Figure 2. In the visualization, we can see that each point is positioned along a smooth

cyclical curve, indicating that motions generated by our method can mostly preserve a unidirectional property. By ensuring this attribute within our system, we can produce realistic and faithful movements that are consistent through time. Furthermore, we also observe that dances belonging to the same group (same color curves) are located in close proximity to each other. This indicates that each group of dancers can be distinctively represented and distinguishable based on their phase features, showing that our method can successfully capture different key characteristics of dances with different styles or groups, and effectively encode them into unique generative continuous phase manifolds. The phase manifold can be utilized to effectively generate consistent and natural movements for an infinite number of dancers via a scalable decoding process.

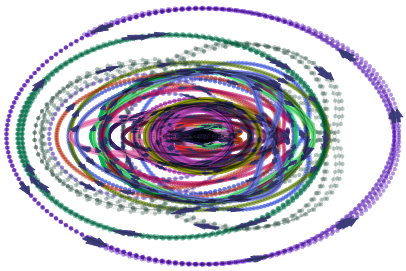

Fig. 2: Visualization of the phase manifold by 2D PCA embedding for different dances. Curves with the same color represent the temporally connected phase features of dance sequences from the same group.

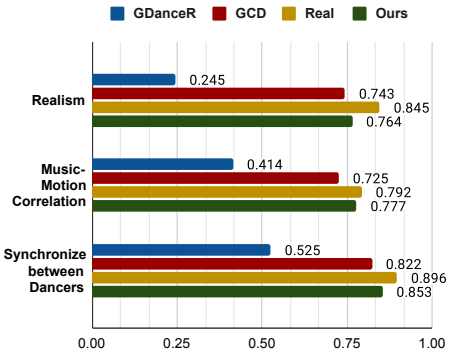

Fig. 3: User study results in three criteria: Realism; Music-Motion Correspondence; and Synchronization between Dancers.

3 Extended Scalability Analysis

Figure 4 demonstrates that our method consistently delivers stable and competitive outcomes irrespective of the number of dancers. This suggests that our proposed approach effectively mitigates the scalability concerns inherent in group dance generation while maintaining the overall performance quality of each dance motion. This underscores the robustness and efficacy of our method across various scenarios, ensuring reliable results regardless of the scale of the dance ensemble.

4 User Study

User studies are vital for evaluating generative models, as user perception is pivotal for downstream applications; thus, we conducted a study with 70 partic-

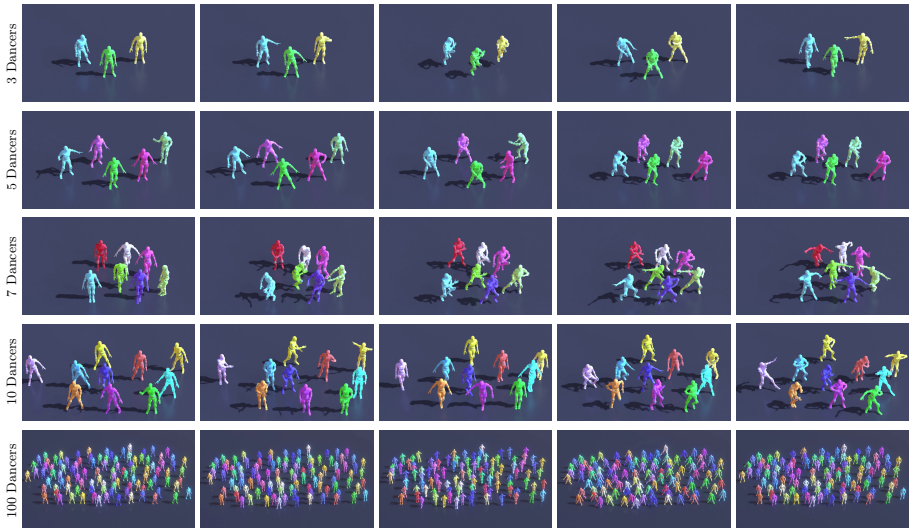

Fig. 4: Extended Visualization of scalable dancers.

ipants, diverse in background, with experience in music and dance, aged between 20 to 40, consisting of approximately 47% females and 53% males, to assess the effectiveness of our approach in group choreography generation.

Participants evaluated dancing animations based on three criteria: realism, music-motion correspondence, and synchronization between dancers, rating them on a scale of 0 to 10 with descriptors such as "very poor," "acceptable," and "very good," which were subsequently normalized to a range of $[0, 1]$.

The user study involved a total of $100 * 4$ samples, which included songs not present in the training set, encompassing samples from GDanceR [5], GCD [4], real dance clips, and results generated by our proposed method. Figure 3 illustrates the average scores for the mentioned criteria across the three experiments. Notably, our method received significantly higher ratings than GDanceR and GCD across all three criteria. These findings underscore that our method can achieve comparable scores with other state-of-the-art approaches while addressing the scalability issue during dance motion generation.

5 Social Impacts

Our research aims to produce realistic dance movements. These lifelike motions have significant potential in various domains, including animation and entertainment industries. However, it's crucial to acknowledge the ethical implications associated with generating such content. Since the dance generation can be implemented into fake dance moves of idols without their consent or knowledge, there arises a concern regarding the misuse of generated content for malicious purposes like deep-fakes, which can be used to encourage downloading videos

that contain harmful malwares. Therefore, we advocate for further exploration into systems capable of detecting augmented or fake media and verifying the authenticity of real media, thus mitigating potential risks and safeguarding against misuse.

6 Human Subjects Data

The AIOZ-GDance and AIST-M datasets are publicly available but no clear statement about the consent of the human subjects is provided. However, our work does not use or exploit any personally sensitive information such as subject identity or appearance.

References

1. Bentley, J.L.: Multidimensional binary search trees used for associative searching. Communications of the ACM (1975)
2. He, C., Saito, J., Zachary, J., Rushmeier, H., Zhou, Y.: Nemf: Neural motion fields for kinematic animation. NeurIPS **35**, 4244–4256 (2022)
3. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
4. Le, N., Do, T., Do, K., Nguyen, H., Tjiputra, E., Tran, Q.D., Nguyen, A.: Controllable group choreography using contrastive diffusion. ACM Transactions on Graphics (TOG) (2023)
5. Le, N., Pham, T., Do, T., Tjiputra, E., Tran, Q.D., Nguyen, A.: Music-driven group choreography. In: CVPR (2023)
6. Li, J., Villegas, R., Ceylan, D., Yang, J., Kuang, Z., Li, H., Zhao, Y.: Task-generic hierarchical human motion prior using vaes. In: International Conference on 3D Vision (3DV) (2021)
7. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. In: NeurIPS (2020)
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
9. Yuan, Y., Iqbal, U., Molchanov, P., Kitani, K., Kautz, J.: Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In: CVPR (2022)