Music-Aligned Holistic 3D Dance Generation via Hierarchical Motion Modeling

Supplementary Material

A. Additional Details of SoulDance Dataset

In Figure 5, we illustrate the relationships among dance genres, the number of dancers, and each dancer's proportion within the dataset. Figure 8 showcases various music-dance motions from different styles in the *SoulDance* dataset. The body and hand movements demonstrate remarkable diversity and precision, further enhanced by expressive facial motions, making the dances more dynamic and emotionally engaging.

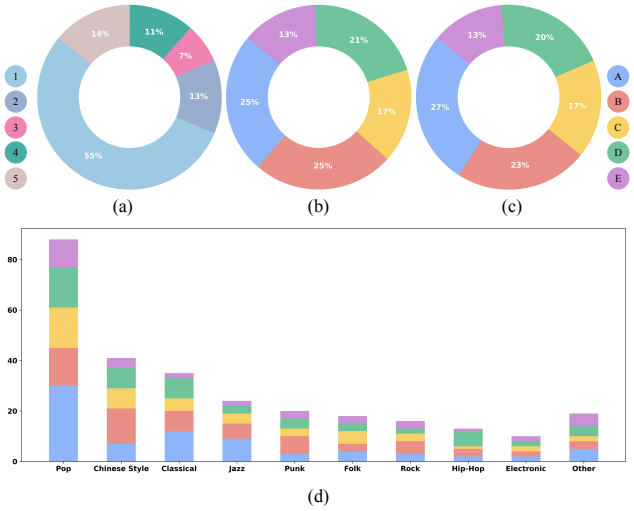

Figure 5. **Overview of the distribution of the *SoulDance* dataset.** (a) shows the distribution of dance sequences by dancer count (1–5), with most sequences featuring solo performances. (b) depicts the proportion of dance duration for each dancer (A-E). (c) illustrates the distribution of dance sequence counts per dancer. (d) displays the number of dance genres and the count of dance sequences per genre for each dancer.

B. Body-Hands Motion Refinement

The majority of existing datasets focus on body movement, typically utilizing 24 body joints from the SMPL model [37]. In contrast, our dataset captures detailed holistic dance motion, requiring the use of the SMPL-X model [39], which includes 22 body joints, 30 hand joints and is compatible with FLAME [35] parameters for facial expressions. As shown in Figure 1, once the body movement and hand gesture BVH data is acquired, we employ MotionBuilder [3] and Unreal Engine 5 [10] to retarget motions to the SMPL-X format. Throughout the retargeting process, we implement a workflow within Unreal Engine 5, including T-pose adjustments, bone length calibration, and joint name mapping to ensure precise alignment. When necessary, our team of engineers performs manual refinements to correct any non-physical joint behaviors, further enhancing the authenticity of the retargeted motions.

C. Transforming Face Blendshapes to FLAME

We follow the method introduced in EMAGE [36] to convert ARKit blendshape weights into FLAME parameters. Given the ARKit blendshape weights $\mathbf{b}_{\text{ARKit}} \in \mathbb{R}^{T\times 52}$, we aim to derive a transformation matrix $\mathbf{W} \in \mathbb{R}^{52\times 103}$ to map these into FLAME parameters $\mathbf{b}_{\text{FLAME}} \in \mathbb{R}^{T\times(100+3)}$, where the dimensionality of 100 corresponds to expression parameters, and 3 represents jaw movements. We leverage a set of handcrafted blendshape templates $\mathbf{v}_t \in \mathbb{R}^{52}$ on the FLAME model, structured according to ARKit's Facial Action Coding System (FACS) configuration. This setup enables direct control of the FLAME topology vertices $\mathbf{v}$ using the blendshape weights:

$$\mathbf{v} = \mathbf{v}_t^0 + \sum_{j=1}^{52} \mathbf{b}_{\text{ARKit},j} \cdot \mathbf{v}_t^j, \tag{10}$$

where $\mathbf{b}_{\text{ARKit},j}$ is the weight of the j-th ARKit blendshape, and $\mathbf{v}_t^j$ is the FLAME template vertex position. The term $\mathbf{v}_t^0$ denotes the initial template vertex positions in the FLAME model. We optimize $\mathbf{W}$ by minimizing the Euclidean distance $\|\tilde{\mathbf{v}}_j - \mathbf{v}_j\|_2$, where $\tilde{\mathbf{v}}$ represents vertices derived from FLAME's Linear Blend Skinning (LBS) function $\mathcal{V}(\mathbf{b}_{\text{FLAME}})$.

D. Holistic Dance Motion Representation

Following the HumanML3D format [17] and HumanTomato format [38] for motion representation, we represent the holistic motion at each frame m_i as a tuple containing various motion attributes. Specifically, we define m_i by the root angular velocity $\dot{r}^a \in \mathbb{R}$ along the Y-axis, root linear velocities $\dot{r}^x, \dot{r}^z \in \mathbb{R}$ on the XZ-plane, root height $r^y \in \mathbb{R}$, local joint positions $\mathbf{j}^p \in \mathbb{R}^{3N-3}$, 6-DOF joint rotations [61] $\mathbf{j}^r \in \mathbb{R}^{6N-6}$, joint velocities $\mathbf{j}^v \in \mathbb{R}^{3N}$, and foot contact indicators $\dot{c} \in \mathbb{R}^4$. Here, $N = 52$ represents the total number of body-hand joints, utilizing 22 body joints and 30 hand joints as defined in the SMPL-X model [39]. For facial motion, we adopt the FLAME format [25], using $\mathbf{f} \in \mathbb{R}^{100}$ to represent facial expressions. Thus, each frame's whole-body motion is represented as $\mathbf{m}_i = \{\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y, \mathbf{j}^p, \mathbf{j}^r, \mathbf{j}^v, \dot{c}, \mathbf{f}\}$, with a total dimension of 723.

E. Dance Reconstruction Evaluation Metrics

During HRVQ training, it is essential to evaluate the reconstruction quality of dance. While body and hand movements can be assessed using the standard MPJPE [28], it fails to capture the accuracy of facial reconstruction. To address this, we introduce the Face Vertex Error (FVE), which quantifies the deviation of reconstructed facial sequences from the ground truth [11, 55]. FVE is computed by measuring the Euclidean distance between the ground truth and reconstructed facial vertices for each frame, then averaging these distances over the entire sequence:

$$FVE = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\sum_{j=1}^{V}(v_j - \tilde{v}_j)^2} \tag{11}$$

where v_j represents the ground truth positions of the facial vertices, and $\tilde{v}_j$ denotes the corresponding vertices reconstructed by HRVQ. The metric is averaged over N frames to evaluate facial reconstruction quality.

F. Additional Qualitative Results

Comparison with SOTA Methods. *SoulNet* demonstrates exceptional qualitative performance on both the AIST++ and *SoulDance* datasets. In Figure 10, FACT [31] generates dance sequences where, after the initial two seconds, body and hand movements become mostly static, and facial expressions are entirely absent. EDGE [52] produces convincing body movements but often fails to generate detailed hand motions and lacks expressive facial output. Bailando [49] captures body movements and facial expressions effectively but suffers from joint dislocations during turns and inadequate hand generation. FineNet [32] delivers satisfactory dance and hand motions but struggles with fine hand articulation and facial expressiveness. In contrast, *SoulNet* not only generates diverse and dynamic body movements but also excels in capturing intricate hand and facial details. As shown in Figure 9, on the AIST++ dataset, SoulNet achieves superior alignment with the musical beat and demonstrates greater diversity in generated dance sequences compared to other methods.

Generating Diverse Dances. SoulNet is used to generate three dance fragments from the same music clip. As illustrated in Figure 11, the generated dances exhibit significant diversity and richness in movement while maintaining alignment with the input music genre, showcasing the excellent multimodal capabilities of our method.

Comparison of Different VQ Methods. Figure 7 presents a comparison of dance motion reconstruction for a ground truth sequence using three quantization methods—HRVQ, RVQ, and VQ—all configured with a codebook size of 512. The results clearly demonstrate that HRVQ outperforms the other methods across all key aspects, including body reconstruction (rows 1 and 3), hand gestures (row 2), and facial expressions (row 4). Furthermore, Figure 6 shows that HRVQ produces the most stable and consistent dance sequences, followed by RVQ, with VQ performing the worst. These findings underscore HRVQ's superior ability to capture fine-grained and expressive dance motions, significantly surpassing RVQ and VQ in reconstruction quality.

Figure 6. **Comparison of Visualization Results.** We visualize the dance generation results on the SoulDance dataset using different methods. The dashed line represents the motion trajectory along the direction of gravity, where smaller fluctuations indicate more stable generated dance motions.

G. Implementation Details.

For HRVQ, we employ residual blocks for both the encoder and decoder, with a downscale factor of 4. Each vector quantizer consists of 6 layers, with each layer's codebook containing 512-dimensional codes. The transformation process uses a MLP and a 1D convolution. The quantization dropout ratio, q, is set to 0.2. For MAGM, we use 6 transformer layers and 6 residual transformer layers, with 8 attention heads and a latent dimension of 512. The learning rate reaches 2×10^{-4} after 2000 iterations, following a linear warm-up schedule for training all models. The batch size is set to 256 for training HRVQ and 64 for training MAGM. During inference, we apply a classifier-free guidance (CFG) scale of 4 and 5. For training MMR module, we use the AdamW optimizer with a learning rate of 1×10^{-4} and a batch size of 32. The latent dimensionality of the embeddings is set to $d = 256$. We set the temperature τ to 0.1, and the weight for the InfoNCE loss to 0.1. The threshold for filtering negatives is set to 0.8. All experiments are conducted on 4 NVIDIA V100 GPUs, and the whole process is completed within three days.

Figure 7. **Visualizes dance motion reconstruction on the *SoulDance* dataset.** From left to right, the columns represent the ground truth (GT), HRVQ, RVQ, and VQ results, respectively.

H. Training: Music-Motion Retrieval Module

Dataset. To establish robust dance-music alignment priors, we gathered high-quality open-source music-dance datasets: AIST++ [31], Finedance [32], PhantomDance [30], and *SoulDance* (totaling 25 hours). Following the preprocessing protocol of EDGE [52] with default temporal segmentation parameters, we derive two specialized motion representations for training distinct Music-Motion Retrieval modules. For Body-Alignment MMR, all sequences are processed using the HumanML3D [16] motion representation ($D_m = 263$). For Whole-Alignment MMR, SoulDance dataset is reformatted via our Holistic Dance Representation (Appendix D, $D_m = 723$), encoding holistic motion movements. Finally, all datasets are partitioned into training/validation/test splits (8:1:1 ratio) using

a stratified strategy that preserves music genre and dance style distributions.

Technical Details. Crucially, we enforce temporal alignment constraints between *SoulNet* and MMR training subsets within AIST++ and SoulDance to prevent data leakage—ensuring no overlapping music clips or motion segments exist across models. Two specialized MMR modules, MMR_{body} and $\text{MMR}_{\text{whole}}$, are pre-trained to provide supervisory signals for the respective losses. MMR_{body} is trained on body-only motion data from public datasets (AIST++ [31], FineDance [32]], PhantomDance [30]) using a 263-dimensional motion representation ($D_m = 263$). In contrast, $\text{MMR}_{\text{whole}}$ is trained on a subset of SoulDance holistic motion data (including body, hands and face) with a 723-dimensional representation ($D_m = 723$).

Training Details. We adopt the same encoder and decoder architectures as TEMOS [41] for training our MMR module, with modifications applied only to the encoder dimensions, while keeping the decoder parameters unchanged. Implementation details are consistent with TMR [42]. For optimization, we use the AdamW optimizer with a learning rate of 10^{-4} and a batch size of 128, as batch size is a critical hyperparameter for the InfoNCE loss. The latent embedding dimensionality is set to $d = 256$, with the temperature τ set to 0.1 and the weight of the contrastive loss term λ_{NCE} set to 0.1. The threshold for filtering negative samples is configured at 0.8.

Experiments. We conducted both qualitative and quantitative experiments to evaluate the performance of the MMR module. As shown in Table 8, MMR demonstrates exceptional retrieval capabilities. Visualized retrieval results further validate this observation. For comparison, we provide two music samples, each paired with two high-similarity dance sequences and two low-similarity dance sequences. Figure 12 and the demo examples illustrate that our retrieval results align better with the beat and rhythm. In these examples, higher similarity scores indicate a stronger correlation between the music and the retrieved dance motions.

Retrieval Task	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓
Music-Motion Retrieval	42.04	56.95	64.93	73.94	84.62	2.00
Motion-Music Retrieval	42.00	57.52	65.62	74.48	84.22	2.00

Table 8. **Retrieval results on the dance dataset.** Both Music-Motion Retrieval and Motion-Music Retrieval tasks maintain Recall@1 performance above 40%.

Supervised MAGM. The pretrained MMR module provides music-motion alignment supervision for the MAGM dance generation pipeline. A critical challenge arises from the discrete token inputs to MAGM, while the MMR's alignment losses $\mathcal{L}_{\text{Align-body}}$ and $\mathcal{L}_{\text{Align-whole}}$ require continuous motion representations for contrastive learning. How can we bridge this gap and enable gradient propagation through the non-differentiable process? As illustrated in Fig. 3, we address this issue in three steps. First, the discrete discrete tokens are decoded into continuous motion sequences $M \in \mathbb{R}^{T \times D_m}$ via the hierarchical residual vector quantization decoder $\mathcal{D}_{\text{whole}}$; second, the reconstructed motion M is encoded through the MMR's motion encoder $\mathcal{E}_{\text{motion}}$ to obtain latent code $\mathbf{z}$, enabling the computation of InfoNCE loss with music features $\mathbf{c}$; and third, to enable end-to-end training despite discrete token sampling, we employ Gumbel-Softmax relaxation [21] during token generation, which provides a continuous approximation of the discrete sampling process and allows gradient flow through the otherwise non-differentiable quantization step, with the temperature parameter τ annealed during training to progressively sharpen the distribution.

I. User Study Details

A/B videos are randomly sampled clips from different datasets or generated by different methods, presented to users for comparison and evaluation. As shown in Figure 14, after watching dance videos A and B, participants were asked to answer the following questions:

- Please rate A/B based on your level of preference.
- Considering only the body movements of A/B, please rate based on your level of preference.
- Considering only the hand movements of A/B, please rate based on your level of preference.
- Considering only facial expressions, how well does A/B convey the emotional tone of the music? Please rate.
- How well does A/B align with the rhythm of the music? Please rate.

We conducted a user study on the dance datasets, selecting four different music genres from the *SoulDance* dataset. For each genre, a random music-dance sequence was chosen and compared with sequences of the same genre from the AIST++ [31] and FineDance [32] datasets. Participants were then asked to rate various performance aspects for each comparison.

In addition, we performed a user study on different dance generation methods. Under identical music conditions, we conducted pairwise comparisons between results generated by *SoulNet* and those produced by FACT [31], Bailando [49], EDGE [52], FineNet [32] and ground truth. Participants rated different aspects of each generated dance sequence. Training and generation were carried out separately on both the AIST++ [31] and SoulDance datasets.

Figure 8. **Showcase of various dance styles in the *SoulDance* dataset.** The *SoulDance* dataset demonstrates high motion quality and diversity across multiple dance styles.

Figure 9. **Qualitative generation result comparisons** for a *Rock* song in the AIST++ dataset.

Figure 10. **Qualitative generation result comparisons** for a *Pop* song in the *SoulDance* dataset.

Figure 11. **Diversity of generated dances.** The *SoulNet* method demonstrates rich diversity under identical input music of the *Chinese Style* genre, encompassing variations in body movements, hand gestures, and facial expressions.

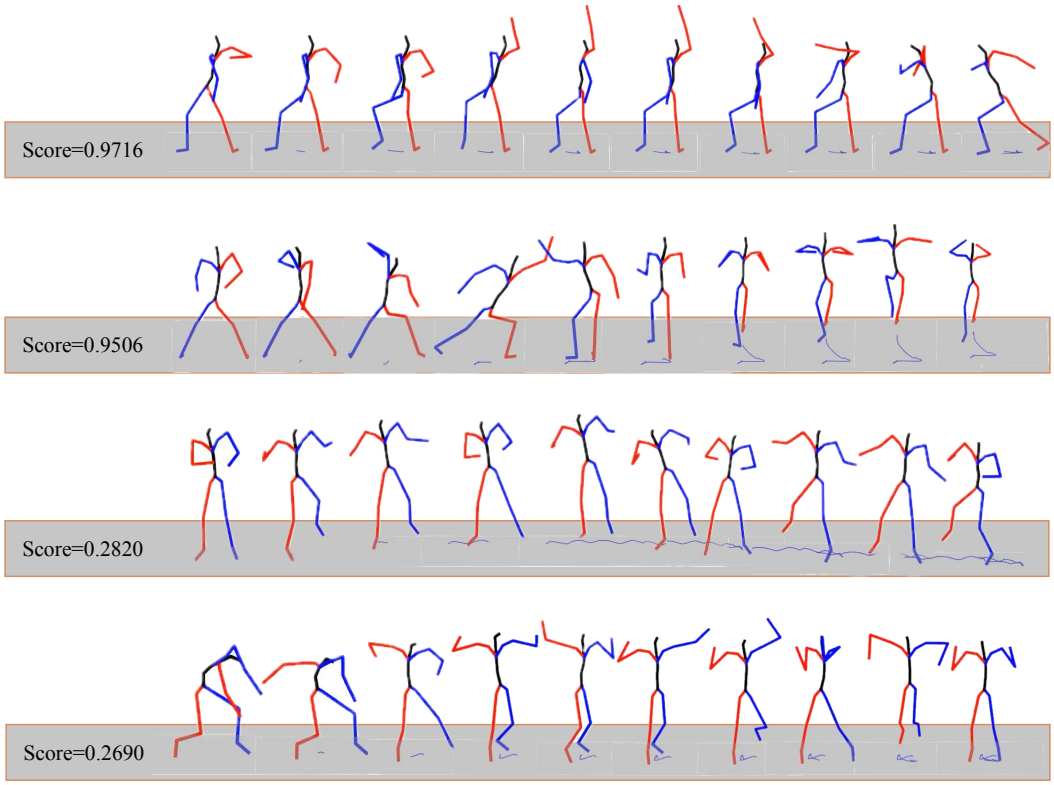

Figure 12. **Qualitative results of Music-Motion Retrieval.** For the *Pop* music genre, higher similarity scores indicate greater correspondence between the retrieved dance motions and the input music.

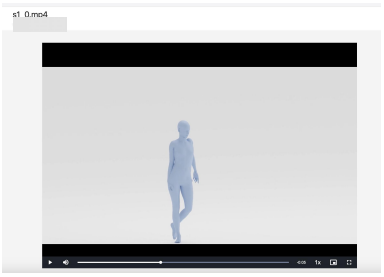

Figure 13. **Screenshot of video page in the user study.** The interface provides independent A/B video links, allowing users to view each corresponding video separately.

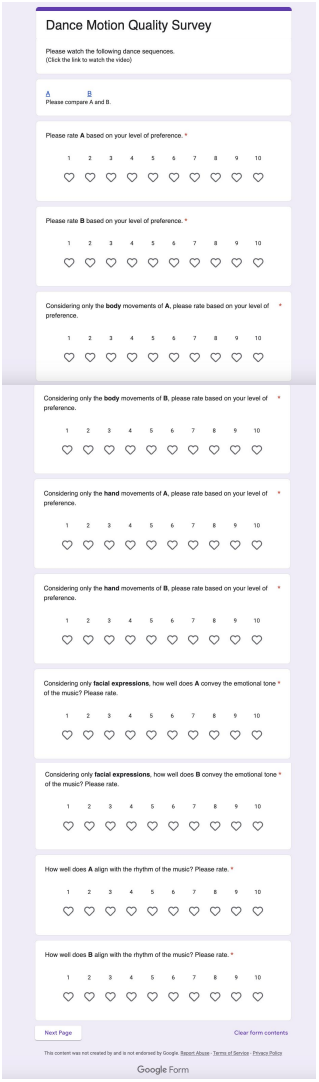

Figure 14. **User interface of our surveys.** The interface presents a set of questions alongside two videos, A and B. Screenshots of the videos linked in the survey are shown in Figure 13.