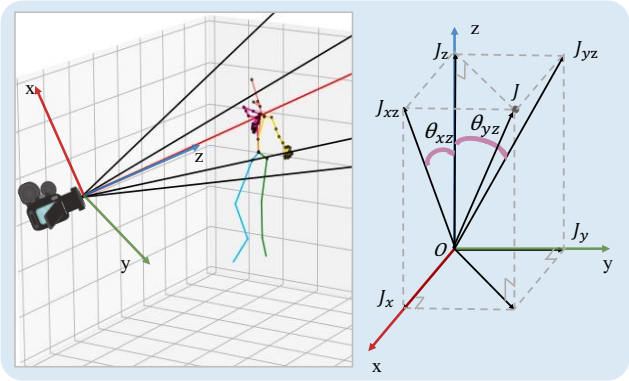
DanceCamera3D: 3D Camera Movement Synthesis with Music and Dance

Supplementary Material

1. Calculation of Joint Mask



(a) Global View (b) Camera Local View

Figure 1. **Details of joint mask calculation.**

As shown in Figure 1, we illustrate camera coordinates in (a) Global View and (b) Camera Local View. Given that O is the camera eye at frame i and J is an arbitrary joint at frame i, our target is to determine whether J is inside the camera view or not. Specifically, we first project J onto the xz-plane and the yz-plane to get J_{xz} and J_{yz}, as shown in (b) of Figure 1. Supposing that joint masks of joints inside camera view are 1 and others are 0, we can represent the joint mask of J as Jm:

$$Jm = \begin{cases} 1 & \theta_{xz} < Fov/2, \theta_{yz} < Fov/2, \\ 0 & \text{others}, \end{cases} \tag{1}$$

where θ_{xz} is the angle between $\overrightarrow{OJ}_{xz}$ and $\overrightarrow{OJ}_z$, θ_{yz} is the angle between $\overrightarrow{OJ}_{yz}$ and $\overrightarrow{OJ}_z$, and Fov is the field of camera view at frame i. In more detail, we use the cosine function to compare angles, because the cosine function is symmetrical about 0 and monotonically decreases on $[0, \pi)$. Thus, we can further represent Jm as:

$$Jm = \begin{cases} 1 & Cos(\theta_{xz}) > Cos(Fov/2), Cos(\theta_{yz}) > Cos(Fov/2), \\ 0 & \text{others}, \end{cases} \tag{2}$$

For computing $Cos(\theta_{xz})$ and $Cos(\theta_{yz})$, we take $Cos(\theta_{xz})$ as an example:

$$\begin{aligned} Cos(\theta_{xz}) &= \frac{\overrightarrow{OJ}_{xz} * \overrightarrow{z_0}}{||\overrightarrow{OJ}_{xz}|| * ||\overrightarrow{z_0}||}, \\ \overrightarrow{OJ}_{xz} &= \overrightarrow{OJ} - \overrightarrow{OJ}_y, \\ \overrightarrow{OJ}_y &= (\overrightarrow{OJ} * \overrightarrow{y_0})\overrightarrow{y_0}, \end{aligned} \tag{3}$$

where $\overrightarrow{x_0}, \overrightarrow{y_0}, \overrightarrow{z_0}$ are unit vectors of x, y, z axes. Here $\overrightarrow{x_0}, \overrightarrow{y_0}, \overrightarrow{z_0}$ and the position of O make up the camera-centric format representation xc, which is mentioned in Sec 3.2 and Sec 4.1 of the full paper.

2. Implementation of Body Attention Loss

In Sec 4.3, we represent our body attention loss $\mathcal{L}_{ba}$ as:

$$\mathcal{L}_{ba} = ||\boldsymbol{Jm} - \hat{\boldsymbol{Jm}} * \boldsymbol{Jm}||, \tag{4}$$

where $\hat{\boldsymbol{Jm}}$ means the generated joint mask and $\boldsymbol{Jm}$ means the ground-truth joint mask. This concise and clear representation denotes that we penalize the joints that are inside the camera view in ground truth but outside the camera view in synthesized results. However, in the actual implementation of $\mathcal{L}_{ba}$, we find that the calculation of joint mask is underivable. Thus, we implement $\mathcal{L}_{ba}$ as:

$$\begin{aligned} \mathcal{L}_{ba} = &\text{Relu}(\boldsymbol{Jm} * (\text{Cos}(\frac{Fov}{2}) - \text{Cos}(\boldsymbol{\theta_{xz}}))) \\ &+ \text{Relu}(\boldsymbol{Jm} * (\text{Cos}(\frac{Fov}{2}) - \text{Cos}(\boldsymbol{\theta_{yz}}))) \end{aligned} \tag{5}$$

where $\frac{Fov}{2}$ indicates the vector of camera field of view, $\boldsymbol{\theta_{xz}}$ and $\boldsymbol{\theta_{yz}}$ denote vectors of θ_{xz} and θ_{yz} respectively. In this way, for the joint outside the camera field of view in ground truth, the Jm is 0 and the corresponding impact to $\mathcal{L}_{ba}$ is 0. For the joint inside the camera field of view in ground truth, the Jm is 1, and the corresponding impact to $\mathcal{L}_{ba}$ is 0 only if this joint is inside the camera field of view in the generated result. Thus, this loss realizes a similar penalty as Equation 4.

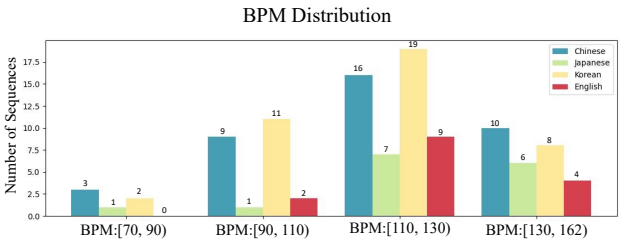


Figure 2. **BPM Distribution of DCM Dataset.**

3. Details of DCM Dataset

3.1. BPM

The BPMs (beat per minute) of music pieces in our DCM dataset range from 71 to 162. In detail, we illustrate BPM distribution with music categories in Figure 2.

Music Languages	Total Time	BPM		Camera Keyframes Interval		Number of Sequences		Frames of Sequences	
		Range	Average	Range	Average	Aligned	Split	Aligned	Split
Chinese	4298.8s	77.8∼143.6	119.2	1∼475	18.45	38	169	524∼8025	510∼1051
Japanese	2258.8s	86.1∼161.5	129.4	1∼220	6.67	15	88	1618∼7290	512∼1046
Korean	3996.5s	86.1∼161.5	119.8	1∼231	11.93	40	159	550∼6260	516∼1042
English	916.6s	99.4∼143.6	122.6	1∼228	18.57	15	38	829∼5737	510∼1046

Table 1. **Detailed statistics of the DCM dataset.** 'Aligned' means data after alignment among dance, camera, and music. 'Split' denotes data split into subsequences within 17∼35s which is more suitable for training.

3.2. Detailed Statistics

As shown in Table 1, we present more detailed statistics of our DCM dataset.