POPDG: Popular 3D Dance Generation with PopDanceSet

Supplementary Material

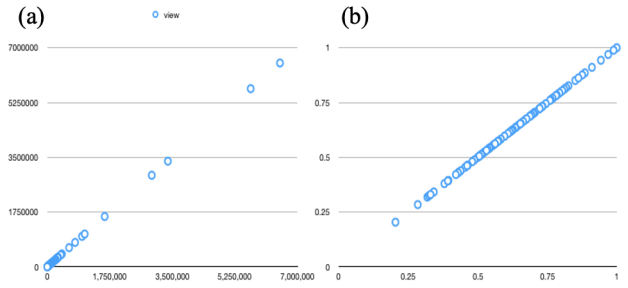

Figure 6. **Data Preprocessing.** (a) The graph represents the original distribution of the view counts for dance videos, showing significant variation in the data distribution; (b) This is the view counts after log normalization, which now exhibits a much more even distribution.

7. PopDanceSet Construction Details

The POPDataset was established in September 2022, with the dance videos primarily spanning from September 15, 2021, to September 15, 2022. In the data preprocessing phase, we initially randomly selected 100 dance videos from BiliBili's dance section and collected data on the number of coins, favorites, danmus, comments, views, likes, and shares. We visualized the view count data as in Fig. 6 (a).

From Fig. 6 (a), it is evident that top popular videos have view counts several orders of magnitude higher than average popular videos, and the same happens to other factors. Therefore, direct linear normalization is not suitable in this case. Instead, we employ non-linear normalization (log normalization) for preprocessing the data of the videos, as shown in Fig. 6 (b).

The core of this experiment in selecting popular dance videos lies in constructing a popularity function. BiliBili's recommendation algorithm for dance videos is given by Eq. (14)[1].

$$Recommendation = \frac{W * N}{n_{views}} \tag{14}$$

where $W = [1.2, 0.9, 1.2, 1.2, 0.75, 1.2, 1.8]$ and $N = [n_{coins}, n_{favorites}, n_{danmucounts}, n_{comments}, n_{views}, n_{likes}, n_{shares}]^T$. This formula indicates that the recommendation function considers multiple factors of a video, not just its view count. The function calculates the growth of these facotrs within a specific time frame, with a $Recommendation$ value greater than 1 significantly increasing the likelihood of the video being recommended on the homepage. Our popularity function was built upon this basis. By omitting the denominator in the formula, we obtained the total values of the video up to the time of data collection. We can then select relevant variables through multiple linear regression and t-tests, with results as in Tab. 6:

Table 6. Estimated Value Ranges of Variables from Multiple Linear Regression and T-Test

Variable	Lower Bound	Upper Bound
bias	0.042	0.046
n_{coins}	-0.002	0.008
$n_{\text{favorites}}$	0.019	0.030
$n_{\text{danmucounts}}$	0.004	0.011
n_{comments}	-0.002	0.008
n_{views}	0.798	0.814
n_{likes}	0.086	0.098
n_{shares}	0.019	0.027

Note: This table presents the lower and upper bounds of variable estimates resulting from a multiple linear regression analysis followed by a T-test. The bounds signify the expected range of values for each variable.

Thus, we eliminated the number of coins and comments from the model and, after another round of multiple linear regression and t-test, obtain the formula presented as Eq. (1).

Following the selection measures described, we ultimately filtered out 263 (around 10% of all collected data) dance videos with a POP greater than 0.85 from a year's span of videos. We also edited these videos into 760 clips featuring relatively high-quality dance content, distributed as Tab. 7:

Table 7. Statistical Distribution of Dance Duration

Duration (s)			Total (s)
Short	Medium	Long	
352 (46.3%)	394 (51.8%)	14 (1.8%)	12818.924

Note: Duration categories are defined as Short (<12s), Medium (12-29.5s), and Long (>29.5s). Percentages represent the proportion of total dances falling within each category.

8. Loss Function

The whole loss is shown as Eq. (10). And in Sec. 4.4 we have already shown velocity and acceleration loss and body

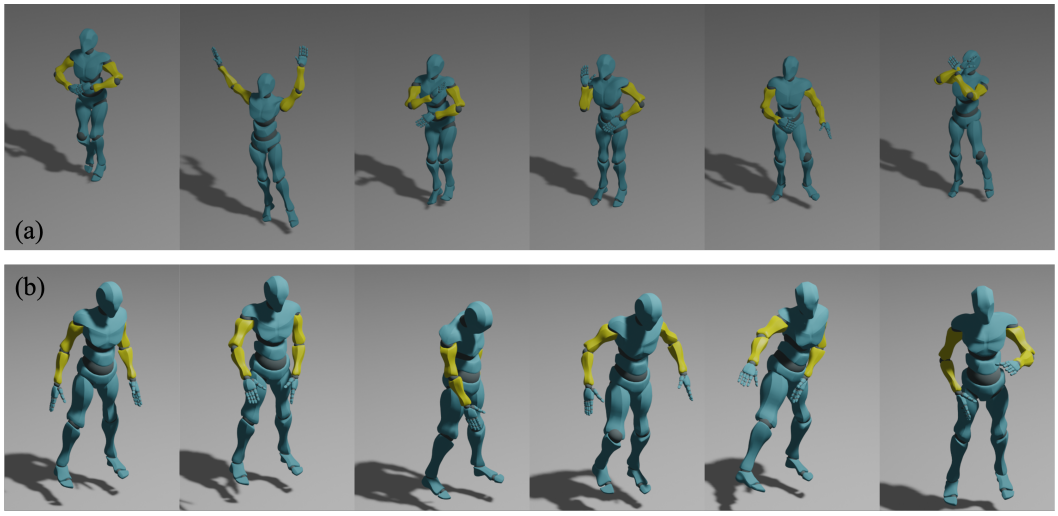

Figure 7. **Comparison of visual effects between PopDanceSet and AIST++.** (a) shows the dance generation results from PopDanceSet, and (b) shows those from AIST++. Both are comparisons of dance postures at the same frame every second under the same background music. Compared to dances generated based on AIST++, PopDanceSet undoubtedly exhibits richer and more captivating movements.

loss, here is the FK loss, as Eq. (15):

$$\mathcal{L}_{\text{FK}} = \frac{1}{N}\sum_{i=1}^{N} \|FK(\boldsymbol{x}^{(i)}) - FK(\hat{\boldsymbol{x}}^{(i)})\|_2^2 \qquad (15)$$

As mentioned in Sec. 4.4, FK(·) denotes the forward kinematic function that converts joint angles into joint positions. Therefore, FK loss is the positional comparison between the generated dance and ground truth in 3D space.

9. PBC and PFC

The EDGE [41] constructs the PFC (Physical Foot Contact) evaluation metric based on the following two assumptions:

- On the horizontal (xy) plane, any center of mass (COM) acceleration must be due to static contact between the feet and the ground. Therefore, either at least one foot is stationary on the ground or the COM is not accelerating.
- On the vertical (z) axis, any positive COM acceleration must be due to static foot contact.

The PFC derived from these two assumptions is Eq. (17):

$$s^i = ||\overline{\boldsymbol{a}}^i_{\text{COM}}|| \cdot ||\mathbf{v}^i_{\text{Left Foot}}|| \cdot ||\mathbf{v}^i_{\text{Right Foot}}||, \qquad (16)$$

$$PFC = \frac{1}{N \cdot \max_{1\le j\le N} ||\overline{\boldsymbol{a}}^j_{\text{COM}}||} \sum_{i=1}^{N} s^i, \qquad (17)$$

where $\overline{\boldsymbol{a}}^i_{\text{COM}} = \begin{pmatrix} a^i_{\text{COM},x} \\ a^i_{\text{COM},y} \\ \max(a^i_{\text{COM},z}, 0) \end{pmatrix}$.

In the SMPL human body model, the COM (Center of Mass) is represented by the 0th joint at the hip, which is also the root joint in Eq. (12). The essence of these two assumptions is that if the body's root joint has acceleration in any direction on the XYZ plane, it means at least one foot must be firmly planted on the ground, as it requires force to initiate movement. Since at least one foot is on the ground, the velocity of that foot should be zero. Thus, the core of PFC is to measure the extent of implausible movements where the body's root joint is accelerating while both feet are moving (i.e., both have velocity). However, this calculation only considers the plausibility of lower body dance movements and overlooks the analysis of upper body movements' plausibility, such as the arms, head and neck. For instance, if the generated dance involves minimal lower body movement but excessive upper body swaying, it would be deemed highly plausible under the PFC metric. Therefore, there's a significant need to also take the upper body into consideration.

In dance, although the upper body movements are relatively independent, we can still observe constraints similar to those between the root joint and the feet within the upper body joints. As illustrated in Eq. (12), whether the left and right chest (referred to as the left and right inshoulder in the SMPL model) and neck joints (i.e., joints 12, 13, and 14 in Fig. 3(a)) accelerate during a dance largely depends on whether the hands and head are moving, that is, whether they have velocity. Specifically, the movements of the hands and head do not necessarily cause movements in the left and right chest and neck joints, but if the latter do move, it generally indicates that the hands and head have also changed position, thus possessing velocity. Unlike Eq. (17), which calculates the irrationality of movements, Eq. (12) adds a calculation for the rationality of movements. Therefore, in PBC, the value of the original PFC needs to be negated,

enabling PBC to reasonably calculate the rationality of full-body dance movements.

10. Visual Effects Comparison

Fig. 7 showcases a comparison of typical dance clips from PopDanceSet and AIST++. As outlined in Sec. 5.5, comparing dances generated from the same model trained on different datasets under the background of the same wild music allows for a clearer distinction of which dataset's dances are more appealing. From Fig. 7, it's evident that dances generated from PopDanceSet are noticeably more engaging. In contrast, dances from AIST++ tend to be more rigid, with several seconds of movement being merely slight adjustments of a single pose. Clearly, the diversity of movements from PopDanceSet, especially in the arm parts, makes these dances more captivating. The only drawback is that the AIST++, with its collection of human keypoints data from nine camera angles, offers somewhat greater stability in the dancer's center of mass compared to PopDanceSet.