A SUPPLEMENTARY MATERIALS

A.1 Subjective Evaluation

We conducted a subjective evaluation involving 20 participants. During the evaluation, the participants watched a video with the combined input motion and the corresponding dance music. For the music, we prepared a total of 15 tracks, which were evenly divided into three groups: our generated music, ground truth music, and a randomly selected ground truth. We asked the participants two sets of questions for each video. The first set focused on the audio quality, while the second set evaluated the correlation between the music and the dance motion. For the specific questions and format used in this evaluation, please refer to Figure 1.

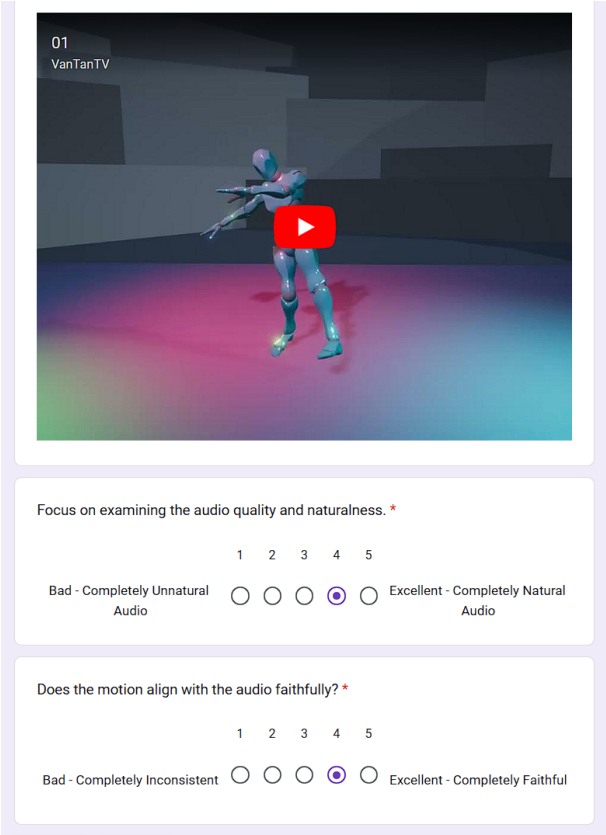

Figure 1: Subjective Evaluation.

A.2 Ablation Study

We investigated the motion features utilized in our method, which comprised p, q, $\dot{p}$, and $\dot{q}$. These quantities represent the position, orientation, linear velocity, and angular velocity of all joints, respectively. As shown in Table 1, augmenting the linear velocity $\dot{p}$ and angular velocity $\dot{q}$ resulted in an improvement in the quality of the generated music in terms of beats coverage score, beats hit score, and Frechet Audio Distance (FAD) score. However, it also led to a decrease in the beat align score and genre KLD score. On the other hand, following the approach of previous music-to-dance research [1] and used only the position p and orientation q as our motion feature, it proved beneficial for aligning the beats between the motion and the music. Overall, our findings suggest that the inclusion of $\dot{p}$, and $\dot{q}$ in the motion feature offers advantages in certain aspects of music generation, while using only p and q improves the beat align score, indicating a better correlation between motion and music.

A.3 Performance Evaluation

We conducted performance evaluations to demonstrate the efficiency of our method compared to the CDCD model [2]. As shown in Table 2, our model demonstrates several advantages over the CDCD model. It excels in terms of GPU memory usage, training time efficiency, and model size. However, due to the iterative denoising process in the continuous latent space, our model takes longer for inferences when compared to the CDCD model.

Model	Beats Coverage Score ↑	Beats Hit Score ↑	Frechet Audio Distance ↓	Beat Align Score ↑	Genre KLD ↓
Ours	**93.5**	**86.0**	**4.96**	0.212	0.604
Ours (no p)	85.9	78.4	5.78	0.211	0.434
Ours (no q)	79.8	72.3	5.99	0.198	0.365
Ours (no $\dot{p}$)	87.7	80.2	5.42	0.214	**0.339**
Ours (no $\dot{q}$)	87.7	81.3	5.17	**0.215**	0.433

Table 1: Ablation Study for AIST++ Dataset.

Model	No. of GPUs	Training Time	Model Size	Inference Time
CDCD [2]	4 RTX A5000	2 days	6.4 GB	**2.95 s**
Ours	**1 RTX A5000**	**1 day**	**0.85 GB**	17.7 s

Table 2: Performance Evaluation.

REFERENCES

[1] Jo-Han Tseng, Rodrigo Castellon, and C. Karen Liu. 2022. EDGE: Editable Dance Generation From Music. *ArXiv* abs/2211.10658 (2022).
[2] Ye Zhu, Yuehua Wu, Kyle Olszewski, Jian Ren, S. Tulyakov, and Yan Yan. 2022. Discrete Contrastive Diffusion for Cross-Modal Music and Image Generation.