X-Dancer: Expressive Music to Human Dance Video Generation

Supplementary Material

In the supplementary material, we provide additional details on our curated in-house dataset (Section 1). In Section 2, we provide a user study to further compare X-Dancer to all baselines. We discuss more use cases of X-Dancer in Section 3, and present some failure cases and discuss some limitations of our method (Section 4). For more dynamic visual results, please refer to our offline webpage.

1. Dataset

Our in-house video dataset is sourced from a third-party curation service, featuring monocular recordings of everyday dance performances by diverse individuals worldwide. To ensure quality, we filter out videos shorter than 5 seconds, those with low resolution or poor quality (HyperIQA score [7] below 40), videos containing multiple people or with a human bounding box smaller than 27.5% of the frame, and footage with non-static cameras (corner pixel standard deviation exceeding 20). After filtering, the dataset comprises 76,818 dance videos, totaling around 360 hours. All filtered videos are center-cropped to 896×512 resolution and resampled to 30 fps.

To analyze the dataset, we employ the state-of-the-art vision-language model (VLM) Qwen2.5-VL [1], yielding the following statistics:

- Gender distribution: 78% female, 22% male.
- Age distribution: 68% of dancers are classified as young, with the rest as middle-aged or older.
- Ethnicity distribution: 49% White, 26% Asian, 11.2% Latino, 9% Black, with the rest others.
- Video duration distribution: 10.04% of videos are between 5-10 seconds, 33.06% between 10-15 seconds, 49.72% between 15-20 seconds, 6.25% between 20-25 seconds, and 0.94% exceed 25 seconds.
- Recording environment: 88% of videos are recorded indoors.
- Dance styles: 89% are categorized as freestyle/hip-hop, while the remaining consist of popping and locking.
- Motion characteristics: 52% feature medium-level movements (e.g., hip swaying, arm waving), while 47.4% exhibit strong motion variations (e.g., leg lifts, body rotations, and translations). 96% of videos do not contain large body turning.

Different from the AIST [9] and FineDance [3] dataset which capture professional dancers in a multiview setup, our dataset consists of monocular recordings of everyday individuals, offering wider accessibility and greater diversity in both dance motions and identity features. However, compared to professional dancers, our curated performances may exhibit weaker beat alignment, mostly frontal movements and less distinct genre characteristics, often leaning towards freestyle movement.

Figure 1. Human image animation after finetuning motion transformer with 30 dance videos of Subject Three.

2. User Study

In addition to our quantitative and qualitative comparisons to various baseline methods provided in Section 4.2 of our main paper, including Hallo [10], Bailando [6] + PoseGudier, and EDGE [8] + PoseGuider. We conduct a user study to further assess perceptual quality, motion fidelity and consistency with the music across all methods.

We generated 25 dance videos for each method conditioned on randomly selected reference images with music tracks. Each video is 10 seconds long, recorded at 30 FPS, totaling 300 frames per video. Evaluation was conducted using a questionnaire distributed via Google Sheets. Participants were asked to choose the best-performing video among candidates based on five specific metrics: (1) Human Identity Consistency, (2) Music Beat Alignment, (3) Music Style Alignment, (4) Motion Consistency and Naturalness, and (5) Overall Quality. In total, 375 responses were gathered from 15 independent participants with no prior knowledge of video generation techniques. For each question sample, the videos from different methods were randomly permuted. The reference images were also provided to assist participants in assessing identity preservation.

Our method substantially surpasses all the baselines across different metrics, as evidenced in Table 1.

3. More Visual Results.

Please refer to our webpage for all the dynamic visual results.

Single Reference, Multiple Music. Given a single reference image, X-Dancer demonstrates the ability to generate diverse and expressive dance motions, maintaining consistent movement styles that adapt to different music genres and beat flows. This underscores X-Dancer's capability to

Table 1. User study. The values indicate user preference ratios (%).

Method	ID Consistency	Beat Alignment	Style Alignment	Motion Consistency	Overall
Hallo [10]	18.67	2.13	1.60	1.87	1.87
Bailando [6]+PG	3.73	10.93	6.40	3.73	3.47
EDGE [8] + PG	4.80	24.80	17.87	11.47	9.87
X-Dancer	**72.80**	**52.13**	**74.13**	**82.93**	**84.80**

effectively interpret the global music context while synchronizing seamlessly with local rhythmic beats.

Single Music, Multiple References. We present diverse dance videos generated by X-Dancer for various reference images, all driven by the same music track. While maintaining a consistent dance style synchronized to the shared music, each generated video also reflects the personalized attributes derived from its corresponding reference image, showcasing X-Dancer's adaptability and attention to individual identity details.

Single Music, Single Reference. We demonstrate a variety of dance movements generated from a single reference image, all driven by the same music track. While all dance movements are well-aligned with the music beats, our model exhibits the ability to produce diverse and dynamic dance motions, highlighting its versatility and creativity.

Finetuning for Characterized Choreography. While our method operates as a zero-shot pipeline, generalizing seamlessly to new reference images and music inputs, it can also be fine-tuned for characterized choreography using only a few sample dance videos. This adaptability is challenging for 3D motion generation models like EDGE [8] and Bailando [6], which require intricate multiview captures or extensive effort in creating 3D dance movements. As shown in Fig. 1 and our supplementary video, our method successfully captures and mimics the specific choreography after fine-tuning with only 30 dance videos from diverse performers, showcasing its efficiency and versatility in adapting to specific dance styles.

Additional Results. We provide additional results including baseline comparisons and results of X-Dancer-AIST (discussed in line 449 - 455 of the main paper).

4. Limitations and Failure Cases.

In addition to the limitations and future work discussed in Section 5 of our main paper, we would like to discuss some additional limitations and failure cases. Specifically, noticeable rendering artifacts remain observable, particularly in the face and hands, and color flickering or over-saturation can occasionally occur. However, as indicated in Table 4 of the main paper (GT Pose + PG), these artifacts persist even when using ground-truth poses, and the overall video quality improvement remains marginal. Such issues are challenges across human image animation models and are orthogonal to the core contributions of our work . Future work could explore stronger video diffusion base models such as [2, 5, 11], and incorporate specialized embeddings or priors for these fine-grained regions [4] to mitigate these problems. Furthermore, while our method effectively generates motions informed by the identity and shape of the reference image, there are cases where the generated motions become inconsistent with the background scene or reference identity. This misalignment can lead to unnatural motions and, in severe cases, pronounced rendering artifacts (Figure 2).

Figure 2. Failure cases. Please note the rendering artifacts on the face and hands, as well as the unnatural renderings under challenging poses.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[2] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2

[3] Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li.

Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *ICCV*, 2023. 1

[4] Supreeth Narasimhaswamy, Uttaran Bhattacharya, Xiang Chen, Ishita Dasgupta, Saayan Mitra, and Minh Hoai. Handiffuser: Text-to-image generation with realistic hand appearances. In *CVPR*, 2024. 2

[5] OpenAI. Sora: Creating video from text. *https://openai.com/sora/*, 2024. 2

[6] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *CVPR*, 2022. 1, 2

[7] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *CVPR*, 2020. 1

[8] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *CVPR*, 2023. 1, 2

[9] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *ISMIR*, 2019. 1

[10] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Luc Van Gool, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. 1, 2

[11] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2