Audio-Image Conditioned Dance Video Generation using CTRL-Adapter

Neel Shah[1] Kevin Zhai[2] Soumik Ghosh[2] Mubarak Shah[2]

[1]Rutgers University [2]University of Central Florida

July 31, 2025

Abstract

Recent diffusion-based models can generate high-quality videos, but multimodal conditioning—particularly with audio and images—remains limited in control and efficiency. This project focuses on generating dance videos conditioned on a single image and audio clip. Instead of training a VQ-VAE and Diffusion Transformer from scratch, we use the CTRL-Adapter framework to inject external conditions into a frozen pre-trained video diffusion model. The image encodes the appearance of the subject, while the audio guides the style and rhythm of the motion. CTRL-Adapter integrates these inputs through lightweight adapters without altering the base model. Our objective is to produce a temporally coherent motion that matches the reference identity and aligns with musical features. This design improves controllability while significantly reducing training complexity.

1 Introduction

Dance is a powerful medium of self-expression and communication, intricately linked to music and cultural identity. With the proliferation of short-form video platforms such as TikTok and YouTube Shorts, user-generated dance videos have become a dominant form of social interaction and content creation. However, choreographing expressive and musically synchronized dance movements typically requires time, skill, and practice. From a computational perspective, synthesizing realistic dance videos is particularly challenging due to the complex, free-form nature of dance and its dependence on rhythm, tempo, and human appearance.

Recent advances in diffusion-based generative models have shown remarkable success in video synthesis. However, these models often lack precise multimodal conditioning, especially with audio and image inputs, limiting their ability to generate temporally coherent and expressive motion that aligns with musical cues. Furthermore, training large-scale generative models such as VQ-VAE and Diffusion Transformers from scratch remains computationally expensive and data intensive.

To address these limitations, we propose a novel approach for generating dance videos conditioned on a single static image and an accompanying audio clip using the CTRL-Adapter framework. Rather than training an end-to-end system, we leverage a frozen pre-trained video diffusion model and inject control signals via lightweight adapter layers. The image input preserves the appearance and identity of the subject, while the audio embedding—extracted using a pretrained CLAP encoder—modulates the generated motion to align with rhythm and style. Our framework enables efficient, controllable generation of dance videos without modifying the base diffusion model, offering a scalable solution to multimodal video synthesis.

2 Related Work

Prior work on music-driven dance generation has primarily focused on synthesizing 3D motion sequences from audio inputs. However, these approaches often suffer from limited expressiveness due to small datasets and do not scale well to diverse human appearances. X-Dancer [1] addresses these limitations by proposing a zero-shot, 2D pose-based image animation pipeline that generates lifelike dance videos from a single static image and a music clip. Their framework consists of a compositional VQ-VAE for encoding 2D whole-body pose sequences, a GPT-style transformer for generating music-synchronized motion tokens, and a diffusion model to render high-resolution video frames that preserve identity and background context. Unlike prior 3D methods such as Bailando [38] and EDGE [44], which often omit head and hand movements, X-Dancer models fine-grained motions across all body parts, improving the realism and expressiveness of synthesized videos.

The X-Dancer pipeline also introduces innovations in pose tokenization and cross-modal conditioning. By partitioning the body into five regions (upper/lower body, left/right hands, head), it enables fine-grained control over motion representation. The transformer autoregressively generates future pose tokens using both global music embeddings and local beat-aligned features, extracted from pretrained Jukebox and Librosa models. These motion tokens are then decoded into spatial guidance using an AdaIN-based motion decoder and passed to a pretrained video diffusion model for synthesis. Evaluations show that X-Dancer outperforms baselines across metrics such as Fréchet Video Distance (FVD), identity similarity, and beat alignment score (BAS), establishing a new benchmark for music-conditioned dance video generation [Chen et al., 2025].

3 Method

3.1 Architecture Overview

Our goal is to generate a temporally coherent sequence of dance video frames conditioned on a single input image and an audio clip. We build upon the *CTRL-Adapter* framework, which enables conditioning of a frozen diffusion model using external control signals. In the original implementation, CTRL-Adapter leverages a CLIP text encoder; in our approach, we replace the text encoder with a CLAP-based audio encoder, allowing us to condition the video generation on music instead of textual prompts.

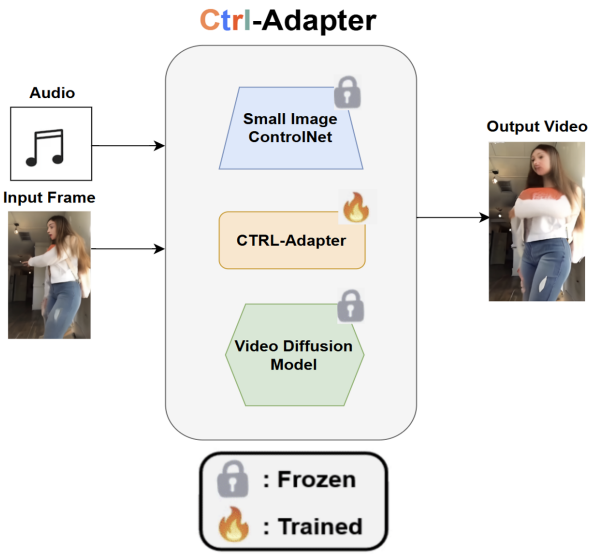


Figure 1: Model architecture

The CTRL-Adapter framework extends ControlNet by injecting conditioning signals—such as depth maps, poses, or other modalities—into a frozen video diffusion model via lightweight adapter layers. The base diffusion model remains unchanged during training. In our setup, we provide two inputs: a single reference image representing the appearance of the subject and a precomputed audio embedding representing the musical context. The audio embedding is extracted using a pretrained CLAP audio encoder and replaces the text embedding used in the original ControlNet pipeline.

As illustrated in Figure 2, we use depth maps as the control condition to guide spatial structure during the denoising process. If a depth map is not precomputed, the CTRL-Adapter automatically ex-

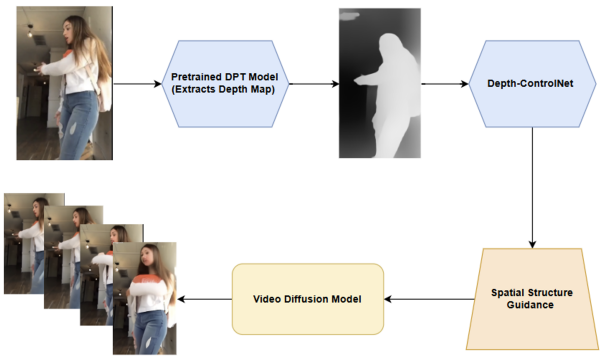


Figure 2: Control Condition:Depth Map

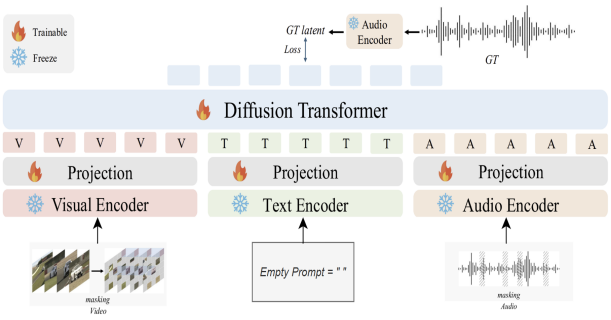


Figure 3: AudioX Diffusion Transformer

tracts it using a pretrained DPT model. The extracted depth map is passed into a specialized Depth-ControlNet module, which processes the spatial information and produces a guidance signal referred to as *spatial structure guidance*.

3.2 Video Generation Pipeline

Our pipeline takes a single input frame and an audio clip. The depth map is either provided or extracted from the image. This depth map is processed by Depth-ControlNet and converted into control features. These control features, along with the audio embedding and the frozen image appearance, are then injected into the backbone video diffusion model through the adapter layers. The video diffusion model synthesizes a sequence of 14 output frames that align with the appearance, depth structure, and musical features of the input. This design enables motion synthesis with fine-grained spatial control and temporal coherence, without training diffusion model.

4 Dataset

We conduct our experiments using the TikTok dance video dataset introduced in the CVPR 2021 paper by Zheng et al., *Learning High Fidelity Depths of Dressed Humans by Watching Social Media Dance Videos*. This dataset contains a large collection of monocular dance videos sourced from TikTok, showcasing diverse clothing styles, human body types, and expressive dance movements. The original videos are silent and do not include any background audio.

For our work, we trimmed and finalized a dataset of 546 video clips, each exactly 10 seconds long, ensuring uniform duration across all samples. To enable audio-conditioned training, we paired each clip with a unique synthesized audio track generated using the *AudioX* framework. AudioX is a diffusion-based transformer model for anything-to-audio generation. We used an empty prompt during generation, allowing the model to produce motion-driven audio that reflects the visual dynamics of each video. This process resulted in a curated dataset of 546 synchronized audio-video pairs, providing a consistent and diverse set of samples for training and evaluation.

5 Results

Figure 4 presents qualitative outputs from our model, where 14 video frames are generated based on a single reference image and an audio prompt. The model effectively preserves subject identity and appearance while generating temporally coherent motion sequences aligned with the rhythm and style of the audio. These results highlight the model's ability to generalize across various subjects and motion dynamics while maintaining consistent visual quality.

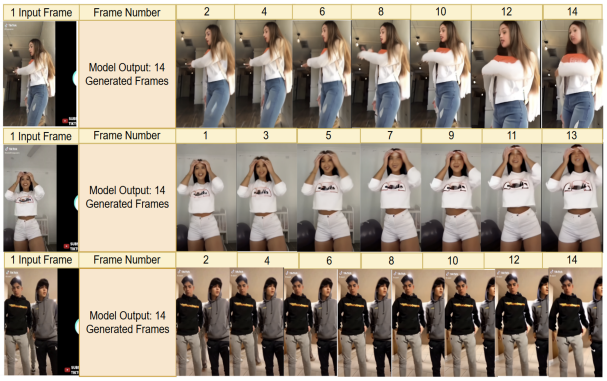


Figure 4: Qualitative results showing 14 generated frames for three different subjects conditioned on a single input image and audio clip.

6 Conclusion

Moving forward, we plan to experiment with generating longer video sequences to better capture extended dance motions and transitions. We also aim to increase the number of output frames per sample, which could lead to smoother and more detailed motion. These improvements will help us explore the full potential of audio-conditioned dance video generation and make the results more expressive and realistic.

References

[1] Zeyuan Chen, Hongyi Xu, Guoxian Song, You Xie, Chenxu Zhang, Xin Chen, Chao Wang, Di Chang, and Linjie Luo. X-dancer: Expressive music to human dance video generation. *arXiv preprint arXiv:2502.17414*, 2025.

[2] Jade Copet et al. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, 2023.

[3] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. *arXiv preprint arXiv:2101.08779*, 2021.

[4] Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. *arXiv preprint arXiv:2404.99697*, 2024.

[5] Zeyue Tian, Yizhu Jin, Zhaoyang Liu, Ruibin Yuan, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Audiox: Diffusion transformer for anything-to-audio generation. *arXiv preprint arXiv:2503.10522*, 2025.

[6] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[7] Li Zhao and Zhengmin Lu. Dancefusion: A spatio-temporal skeleton diffusion transformer for audio-driven dance motion reconstruction. *arXiv preprint arXiv:2411.04646*, 2024.