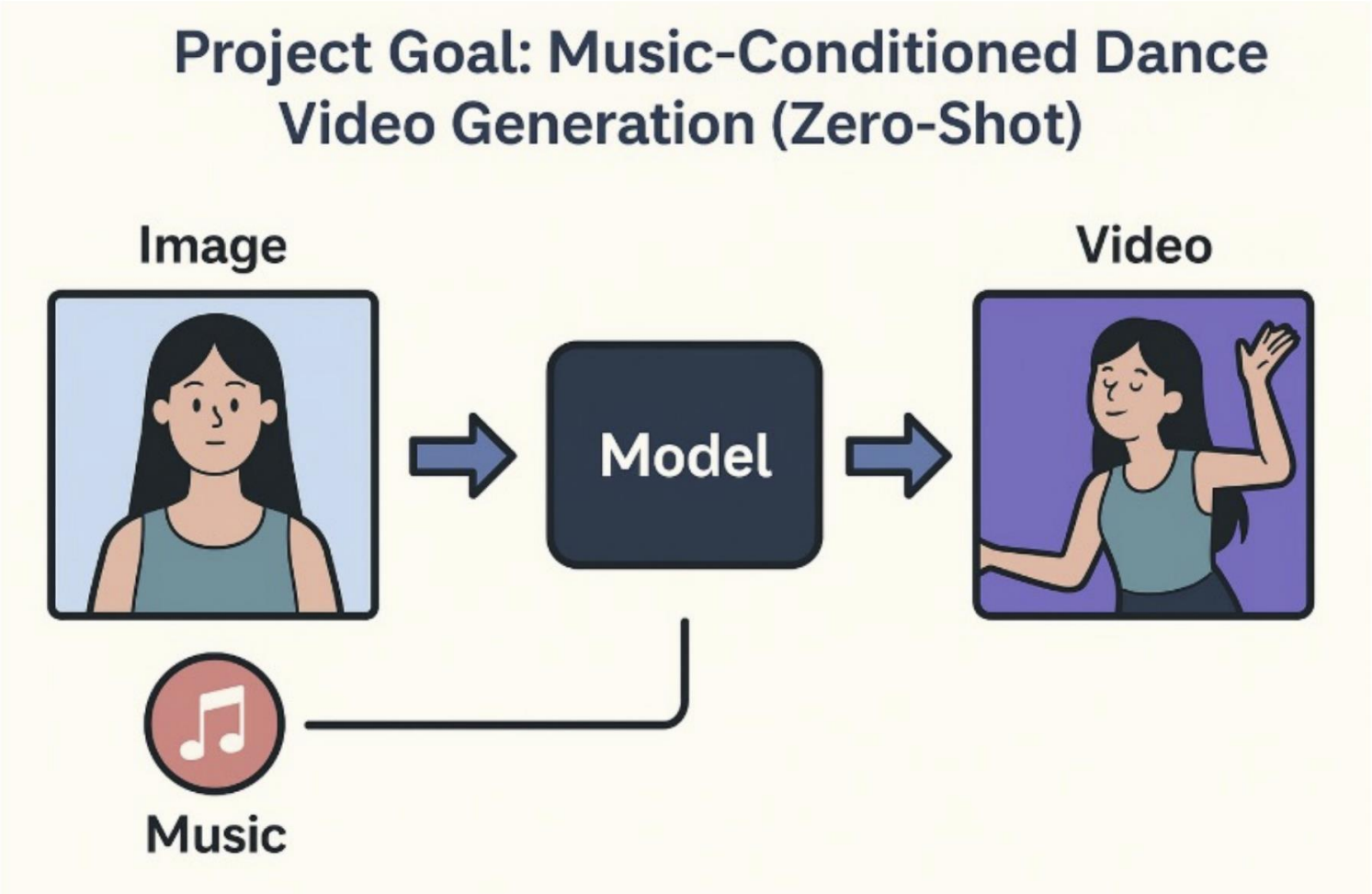


UNIVERSITY OF
CENTRAL FLORIDA

Audio-Image Conditioned Dance Video Generation using CTRL-Adapter

Mentors: Kevin Zhai, Soumik Ghosh, Dr. Mubarak Shah

Problem Statement



BACKGROUND

- **Challenge**: Training VQ-VAE and DiT models required significant compute and large datasets, which were not feasible.
- **Approach**: Use CTRL-Adapter to condition a frozen video diffusion model with image and audio via lightweight adapters.

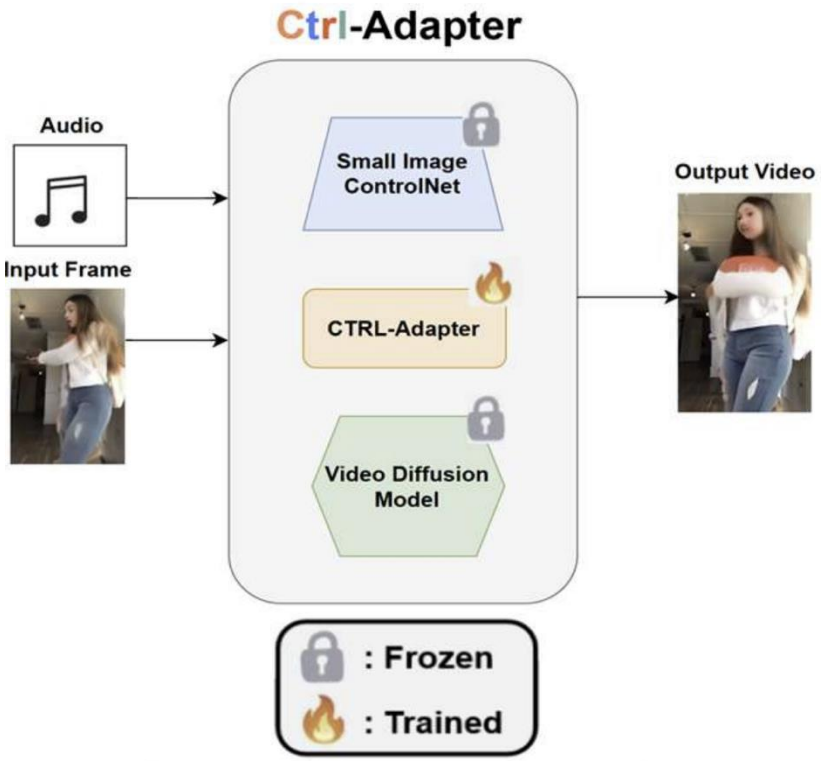


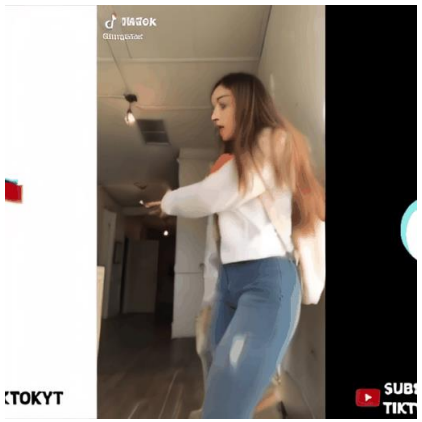
Figure 1: Video Generation Method

Control Condition – Depth Map

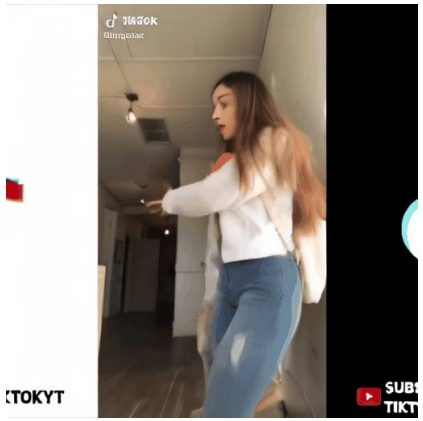


- Grayscale images showing object distance from the camera.
- Used in Ctrl-Adapter as control signals to guide video generation.
- Ctrl-Adapter extracts depth maps using a pretrained DPT model during inference.

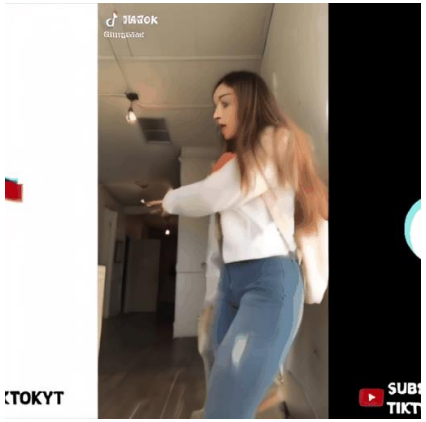
Input Frames = 1, control guidance = 0, Output Frames = 14



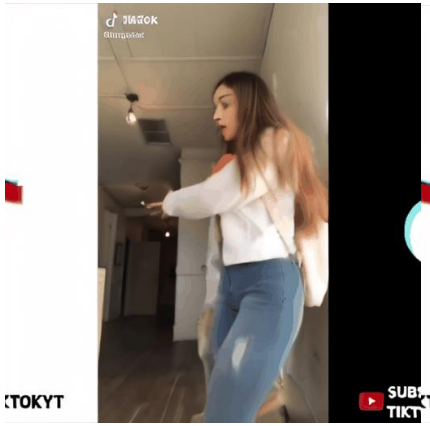
Audio file 1



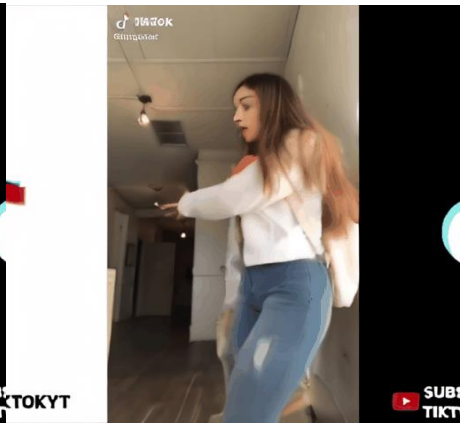
Audio file 2



Audio file 3

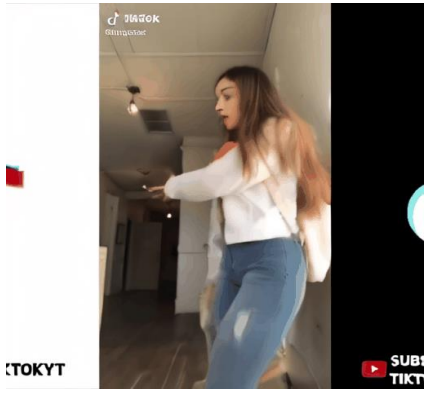


Audio file 4

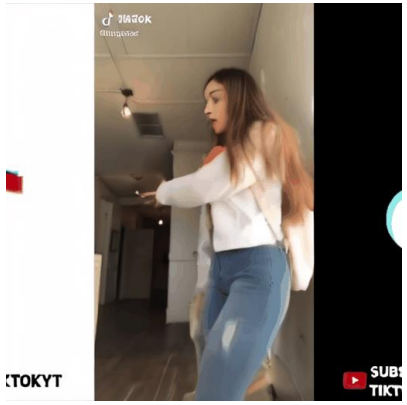


Audio file 5

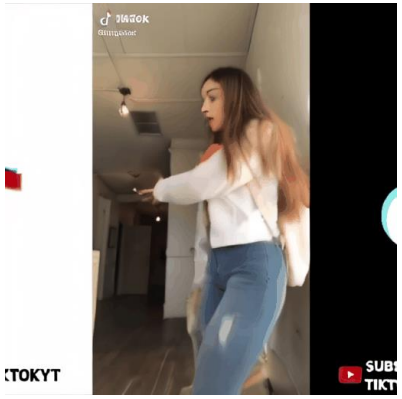
Input Frames = 14, control guidance = 0, Output Frames = 14



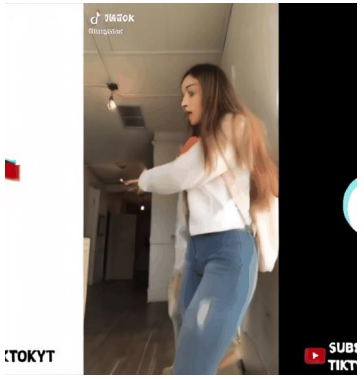
Audio file - 1



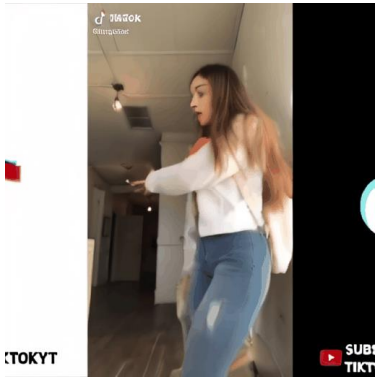
Audio file - 2



Audio file - 3



Audio file - 4



Audio file - 5

Challenges

- Training failed when I swapped the Text Encoder with the Audio Encoder
- Couldn't debug due to the time constraints
- Couldn't access the Newton cluster over the weekend, Monday and Tuesday

Poster Completed

CENTER FOR RESEARCH IN COMPUTER VISION

Audio-Image Conditioned Dance Video Generation using Ctrl-Adapter

Neel Shah[1], Kevin Zhai[2], Soumik Ghosh[2] , Mubarak Shah[3]

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

RUTGERS

Introduction

Goal:
- Generate dance videos conditioned on a single Image and an audio clip
- Ensure temporal coherence matching appearance & rhythm
- Improve multimodal control in diffusion models

Motivation:
- Current video diffusion lacks precise audio & Image conditioning
- Training VQ-VAE & DiT from scratch is not feasible
- Audio & Image multimodal control remains underexplored
- CTRL-Adapter injects conditions without altering the base model

Materials & Methods

Dataset:
- CVPR 2021 TikTok dance dataset
- Original clips are silent
- For each video, we synthesized a unique audio track using the same text prompt
- Audio generation is motion-driven, producing a different soundtrack per Video

Model:
- Frozen pre-trained video diffusion Unet
- CLIP Image encoder
- CLAP audio encoder (replaces the original text encoder)
- CTRL-Adapter lightweight adapter layers

Forward Pass:
- Encode Image & audio into embeddings
- Inject embeddings via CTRL-Adapter into Unet denoising
- Decode denoised latents into final video frames

Results



Model Architecture

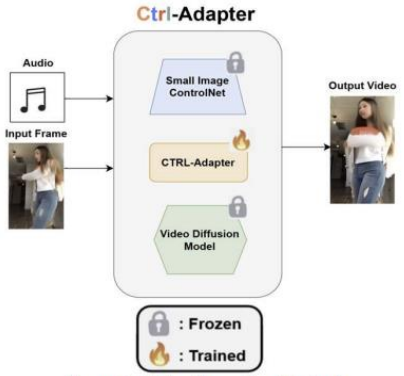


Figure 1: Video Generation Method

Audio Generation Architecture

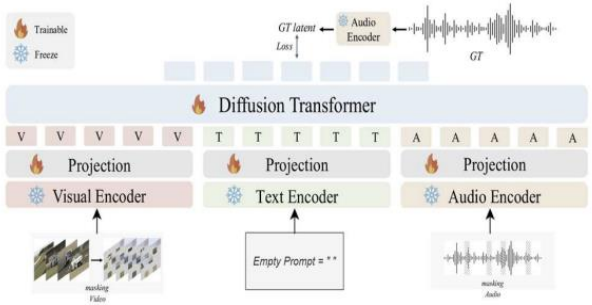


Figure 2: AudioX Diffusion Transformer for Anything-to-Audio Generation

Conclusion

Future Work:
- Test the model with longer videos
- Increase the number of processed frames

References

[1] Lin, H., Cho, J., Zala, A., and Bansal, M., "Ctrl-Adapter: An Efficient and Versatile Framework for Adapting Diverse Controls to Any Diffusion Model", arXiv e-prints, Art. no. arXiv:2404.09967, 2024. doi:10.48550/arXiv.2404.09967.

[2]Tian, Z., "AudioX: Diffusion Transformer for Anything-to-Audio Generation", arXiv e-prints, Art. no. arXiv:2503.10522, 2025. doi:10.48550/arXiv.2503.10522.

Acknowledgement

Dr. Mubarak Shah, Dr. Lobo, Kevin Zhai, Soumik Ghosh, Melody Halbert

Plan for Next Week

- Train the model on the entire dataset to verify end-to-end data and pipeline consistency
- Start fine-tuning the model
- Finish writing the report

THANK YOU