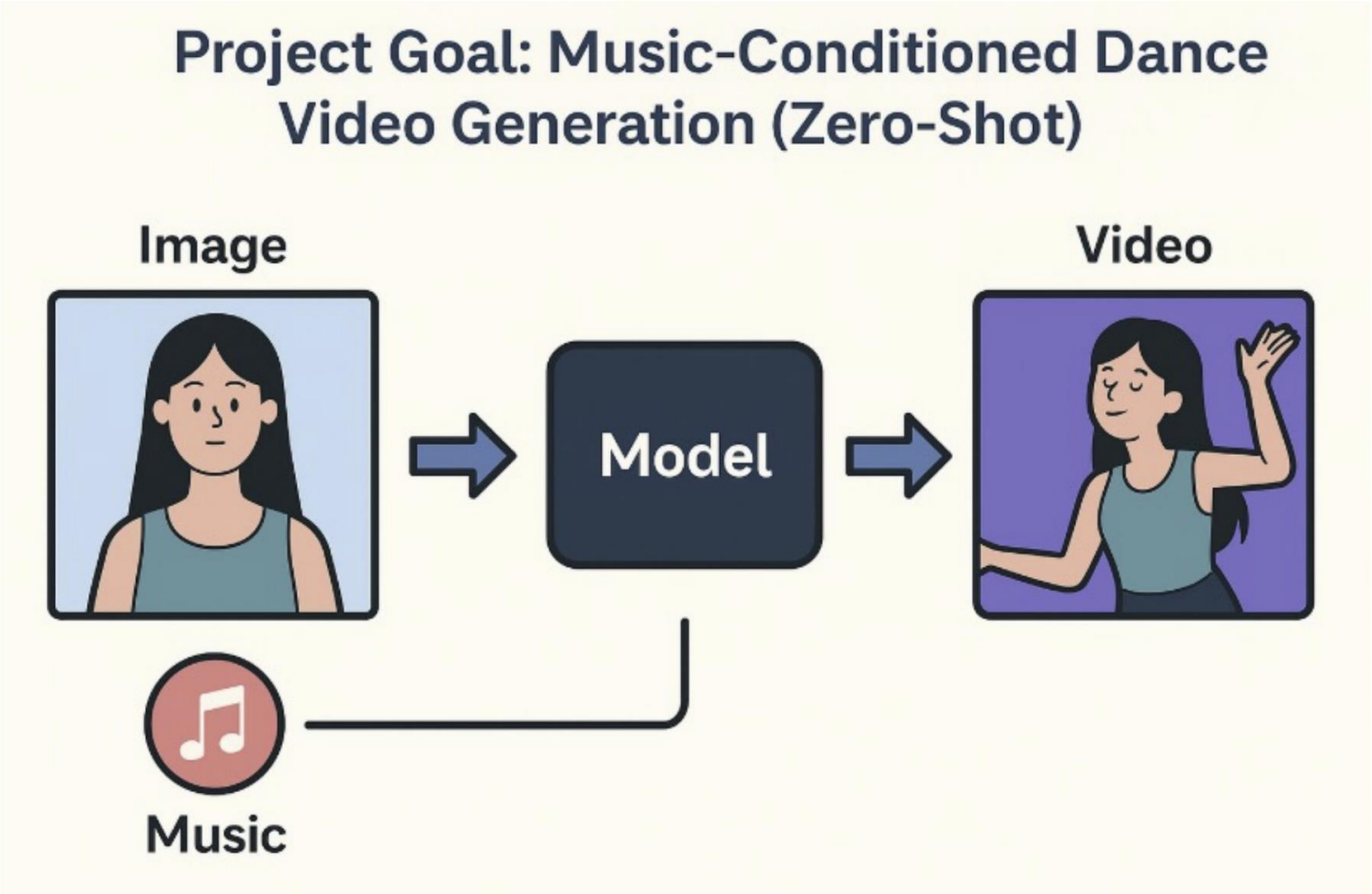


UNIVERSITY OF
CENTRAL FLORIDA

Audio-Image Conditioned Dance Video Generation using CTRL-Adapter

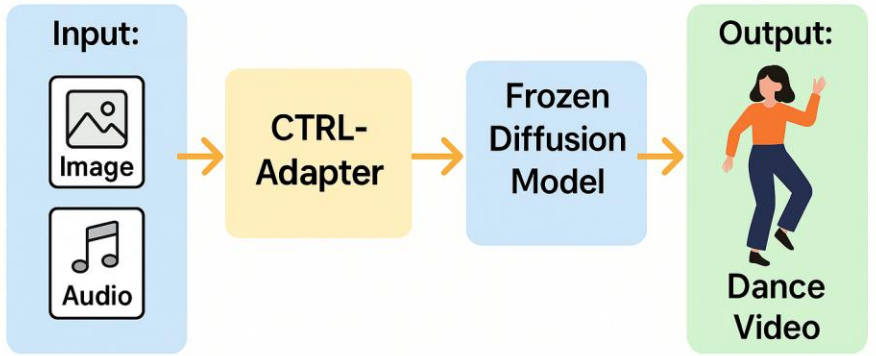
Mentors: Kevin Zhai, Soumik Ghosh, Dr. Mubarak Shah

Problem Statement



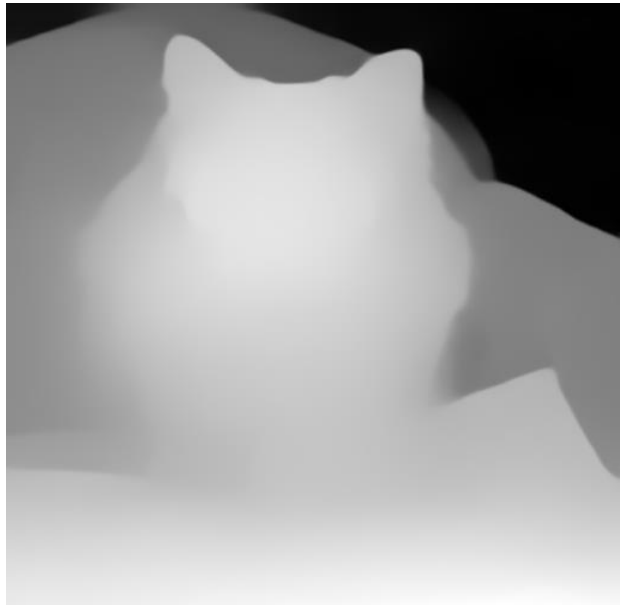
BACKGROUND

- **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.



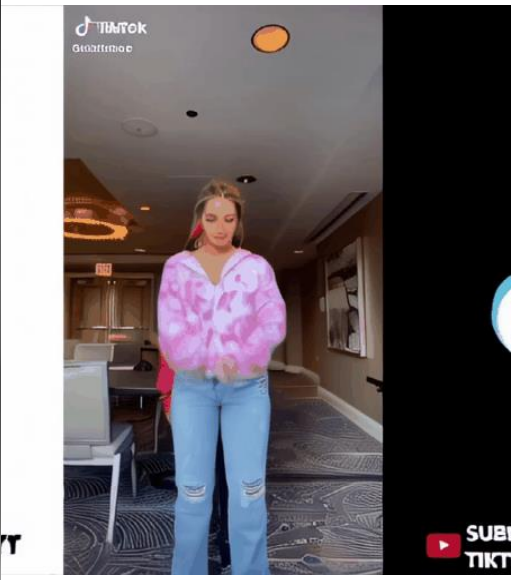
Results

Input Frames = 14, Ouptut Frames = 14 Control types "depth", Control guidance = 0
Input Captions: Newspaper_cat = "", Person = ""



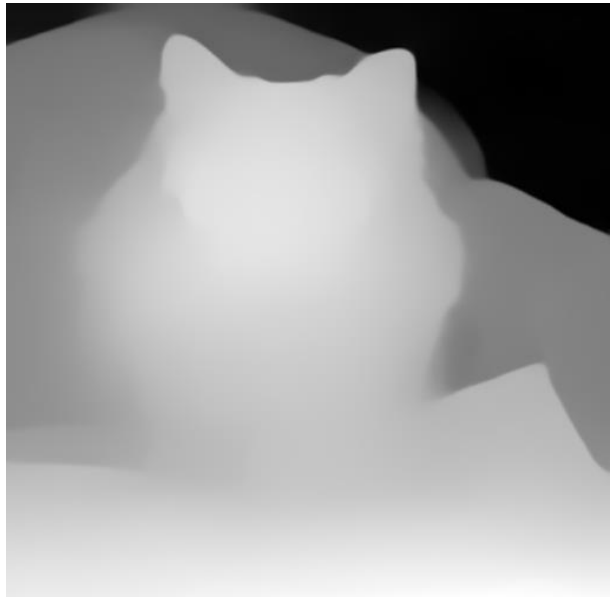






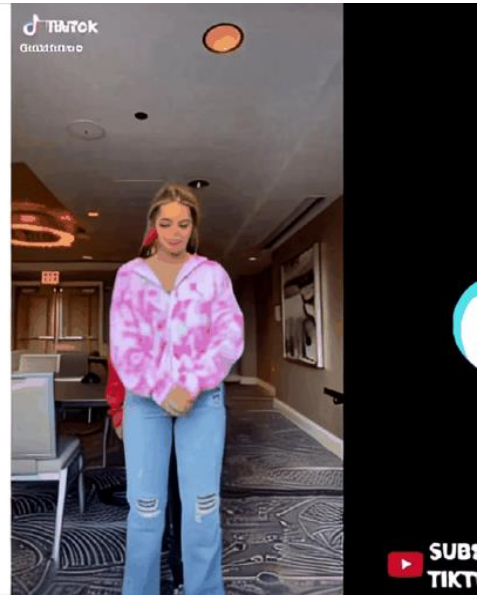
Results

Input Frames = 14, Output Frames = 14, Control types "depth", Control guidance "0.8"
Input Captions: Newspaper_cat = "", Person = ""









Plan for Next Week

- Try Removing the Control Condition i.e. Depth and Control guidance and run experiments using the basic Vanilla model.
- Replace the text encoder in Ctrl Adapter with CLAP (Contrastive Language Audio Pretraining)
- Implement and test a simple video+audio data loader using the AudioX clips and Images from the tiktok dataset
- Run small-scale baseline experiments to verify end-to-end data and pipeline consistency

THANK YOU