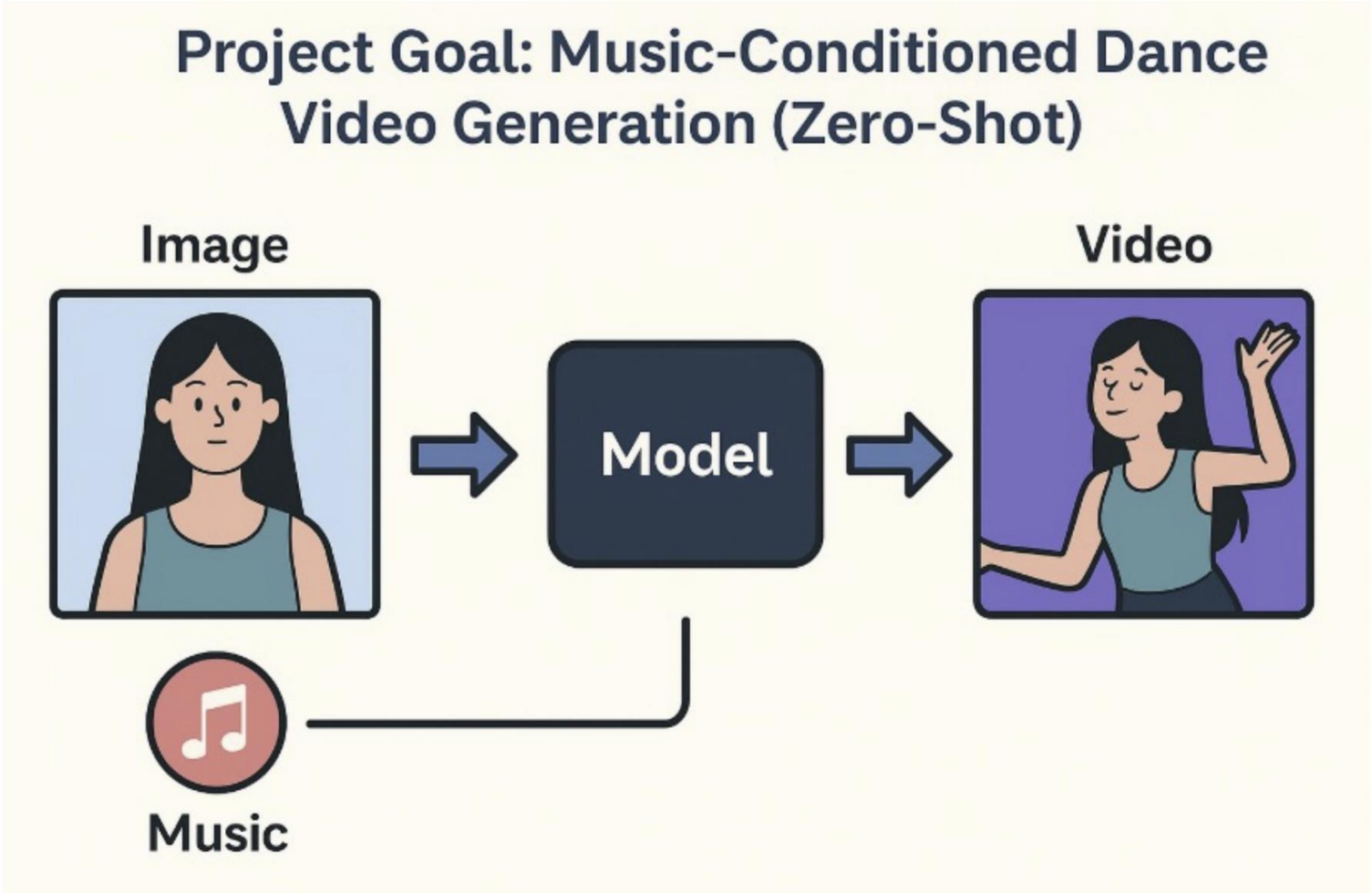


UNIVERSITY OF
CENTRAL FLORIDA

Audio-Image Conditioned Dance Video Generation using CTRL-Adapter

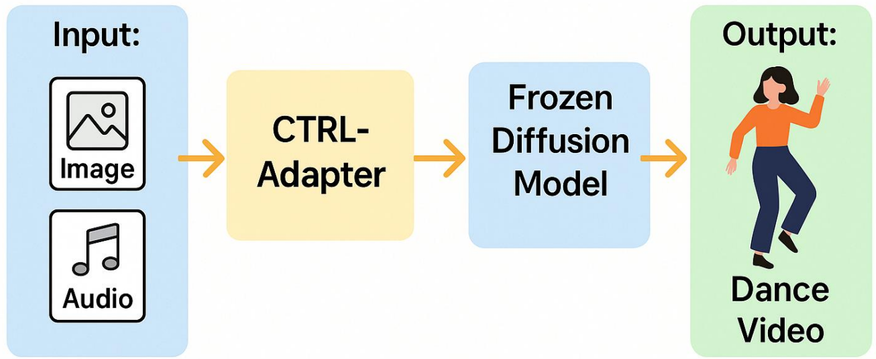
Mentors: Kevin Zhai, Soumik Ghosh, Dr. Mubarak Shah

Problem Statement



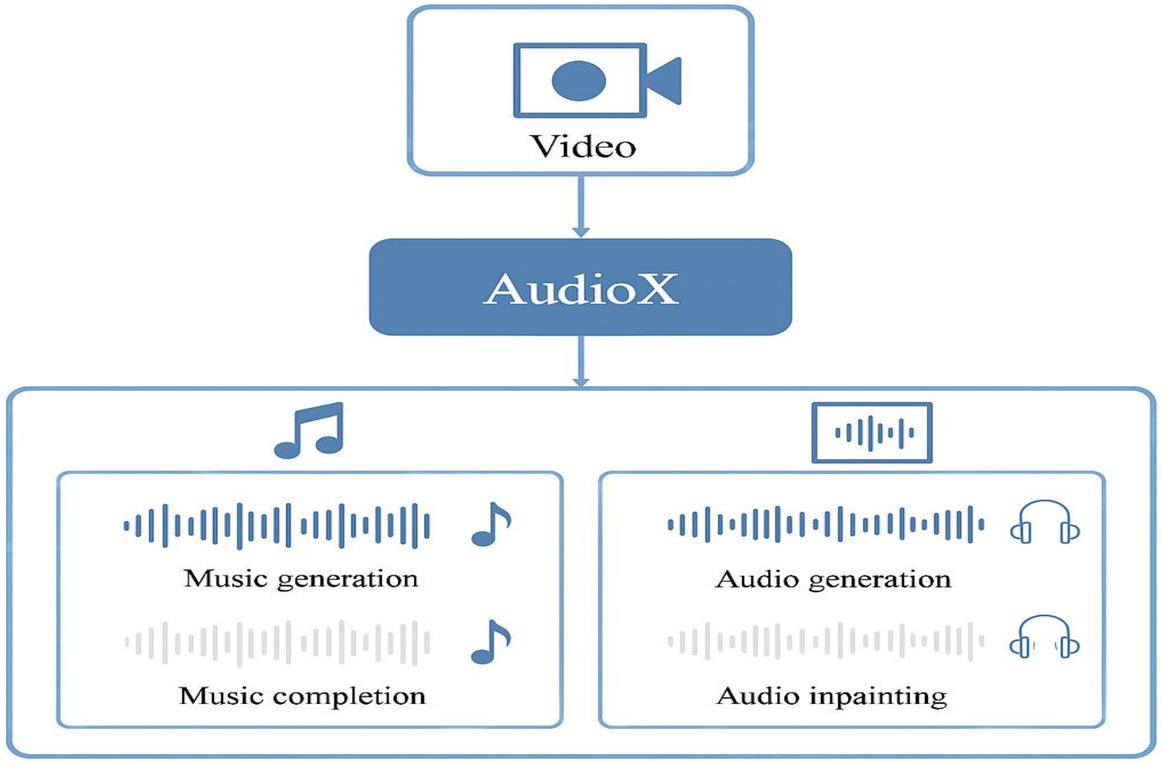
BACKGROUND

- **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.



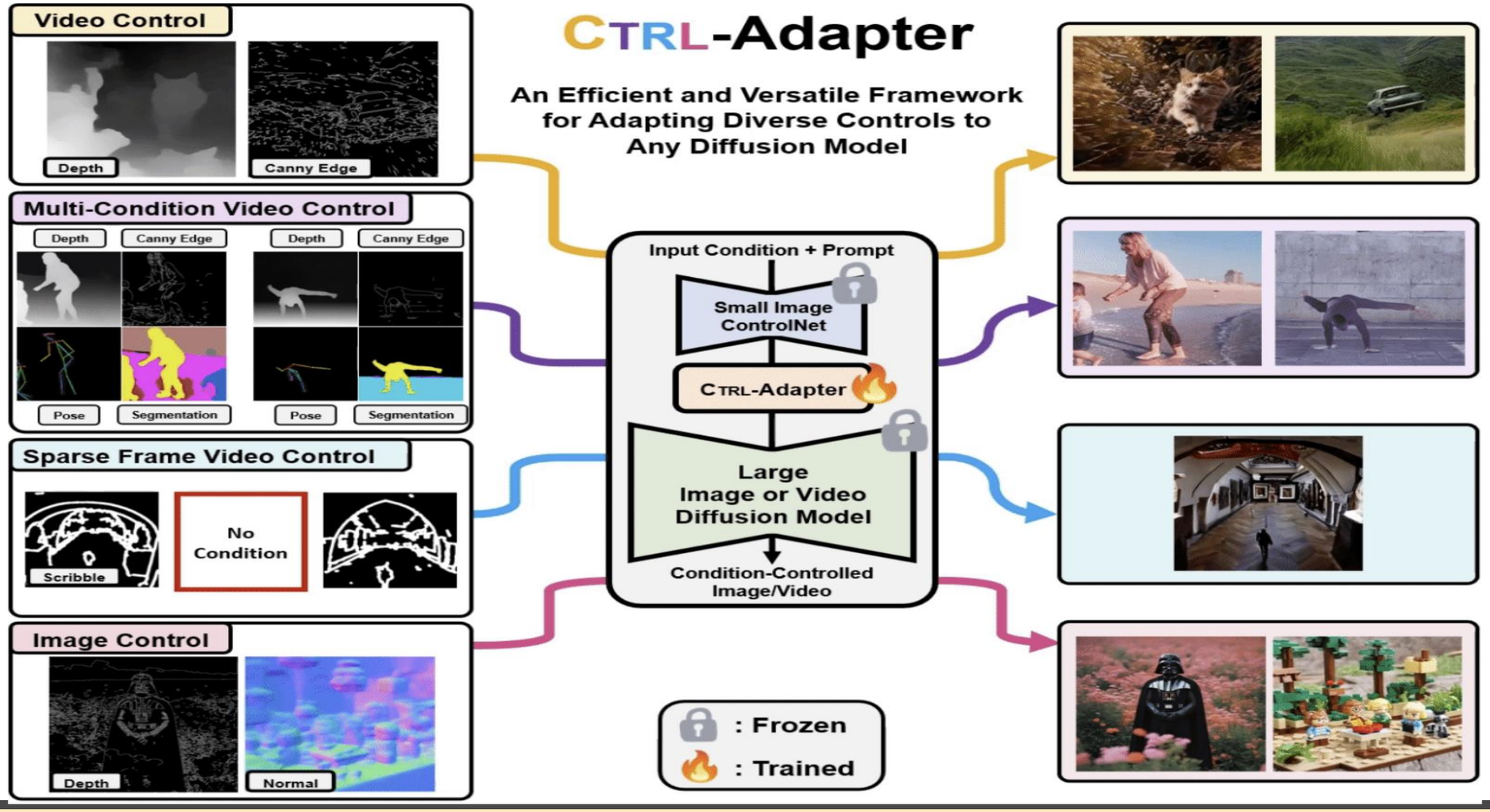
Dataset

Tiktok Dataset is published in CVPR 2021 introduced in the paper: Learning High Fidelity Depths of Dressed Humans by Watching Social Media Dance Videos





Ctrl Adapter



Plan for Next Week

- Read ControlNet and Ctrl-Adapter papers to clarify key differences and challenges
- Set up the coding environment with pretrained weights
- 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