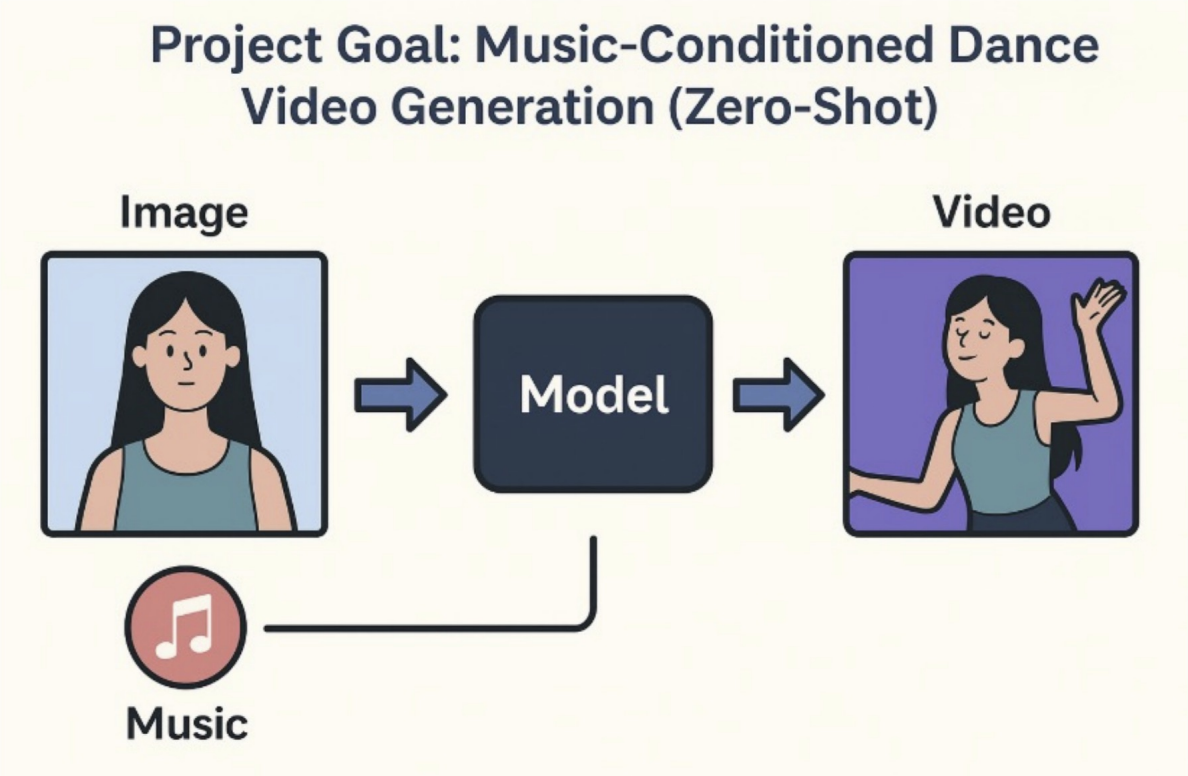


UNIVERSITY OF
CENTRAL FLORIDA

Audio-Image Conditioned Dance Video Generation using CTRL-Adapter

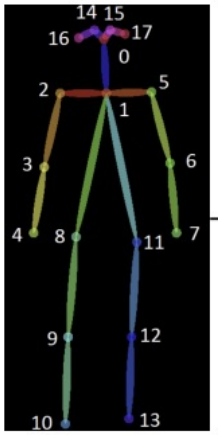
By Neel Shah, Mentors: Kevin Zhai, Soumik Ghosh, Dr. Mubarak Shah

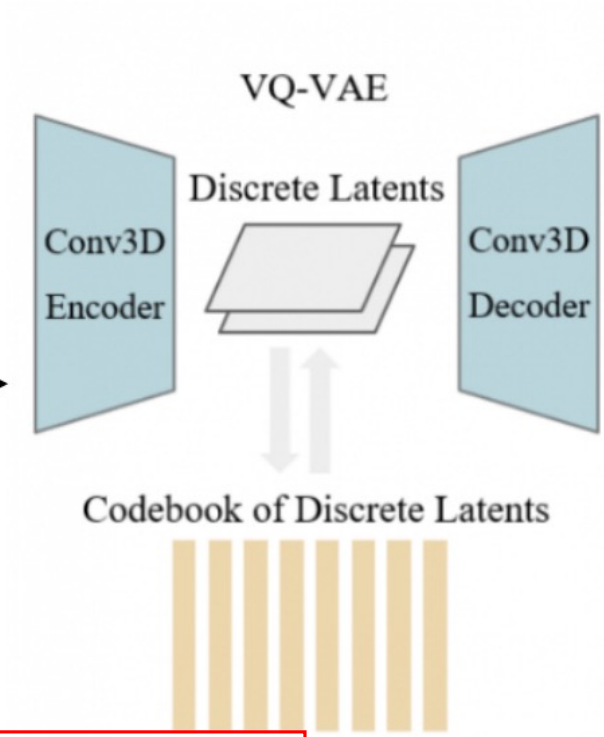
Goal


Project Goal: Music-Conditioned Dance Video Generation (Zero-Shot)
Image
Model
Video
Music

Current SOTA Method

2D Pose Estimation



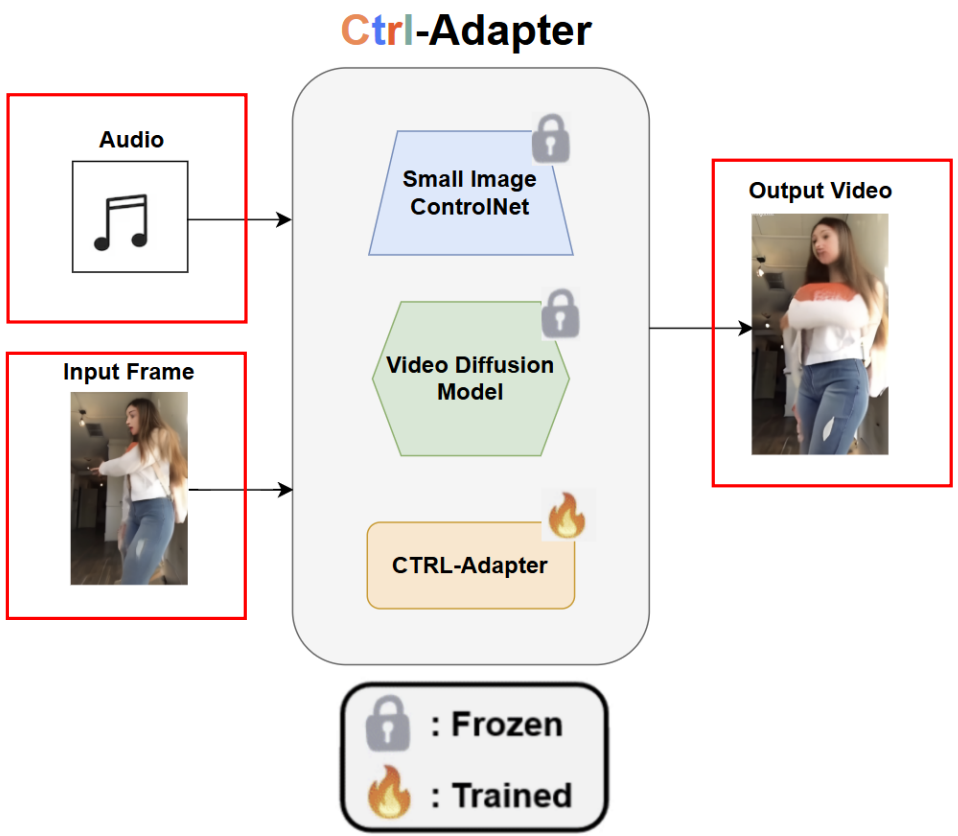


Diffusion Transformer

Requires Large Computing

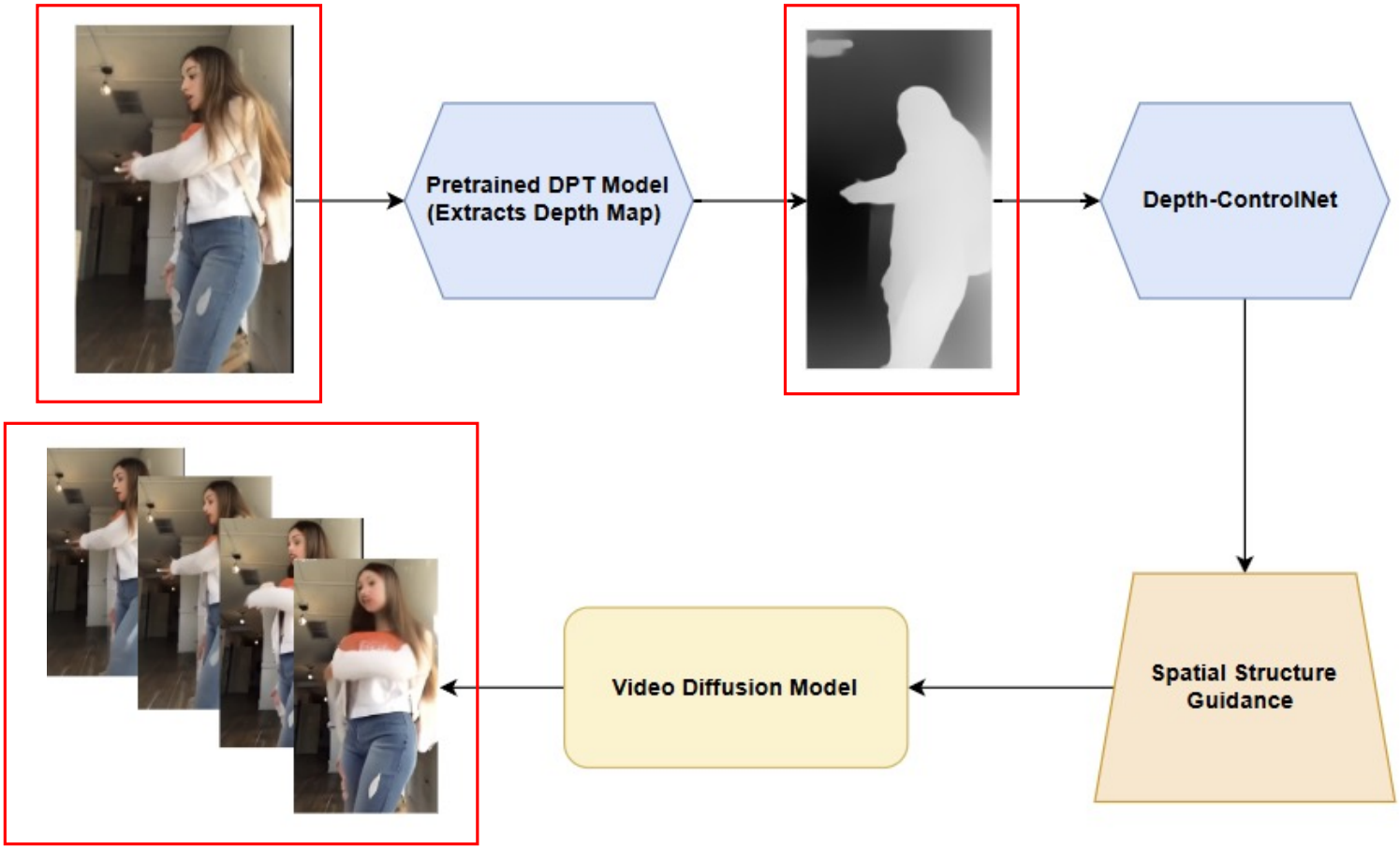
Complex Training

Method

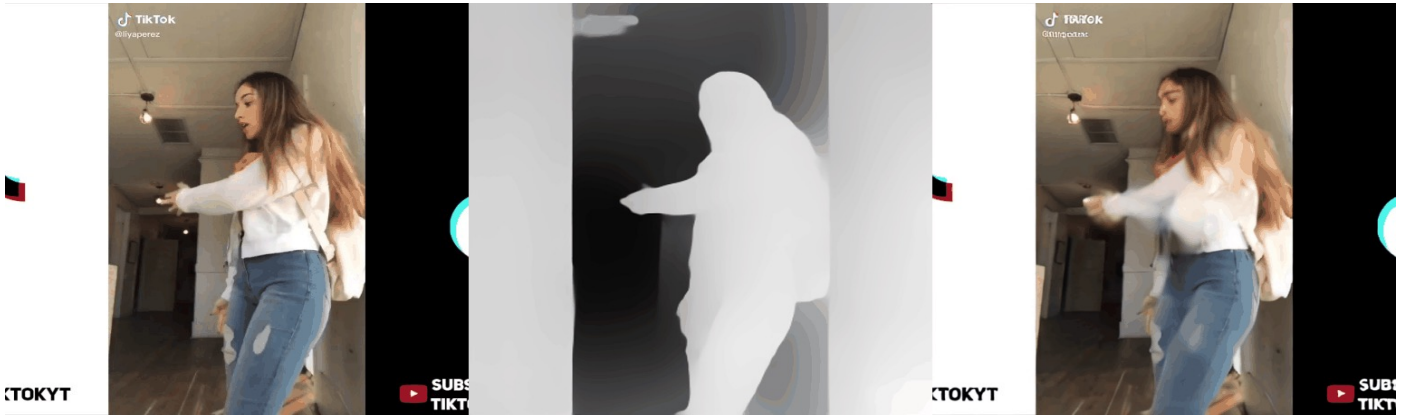


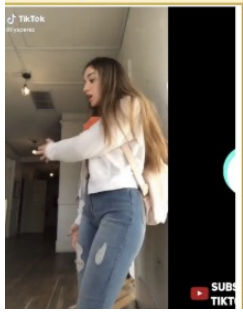
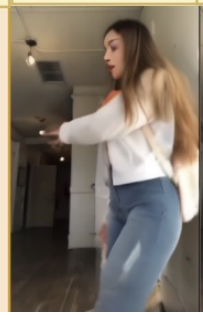
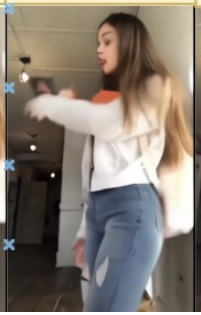
- Small Image ControlNet and Video Diffusion Model (SVD) weights are frozen
- No training is done on the base diffusion model
- Text encoder is replaced with a pretrained audio encoder
- Only the Ctrl-Adapter is trained to enable audio-conditioned video generation

Control Condition – Depth Map



Results



1 Input Frame	Frame Number	2	4	6	8	10	12	14
	Model Output: 14 Generated Frames							

Results



THANK YOU