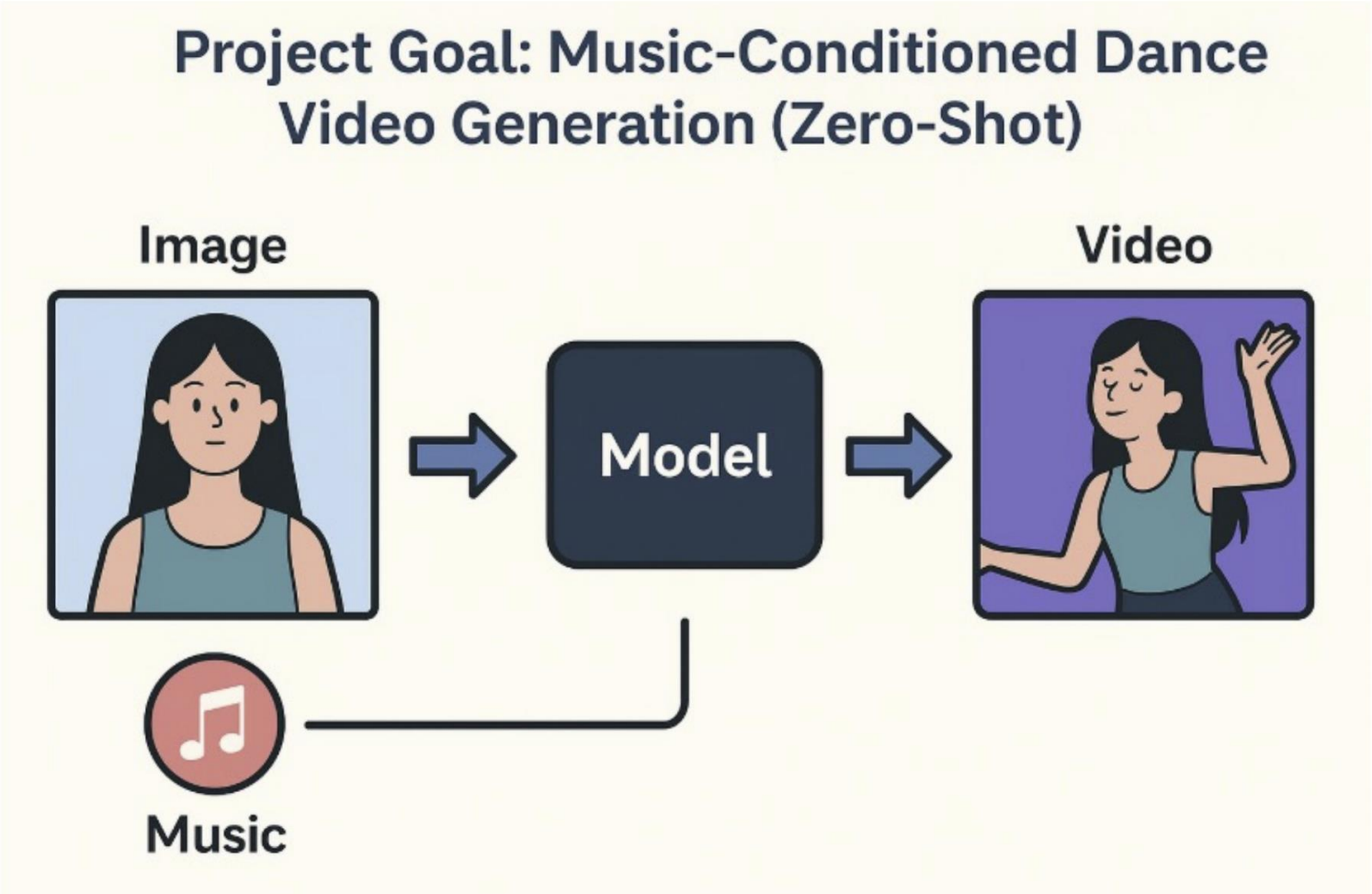


UNIVERSITY OF
CENTRAL FLORIDA

Audio-Image Conditioned Dance Video Generation using CTRL-Adapter

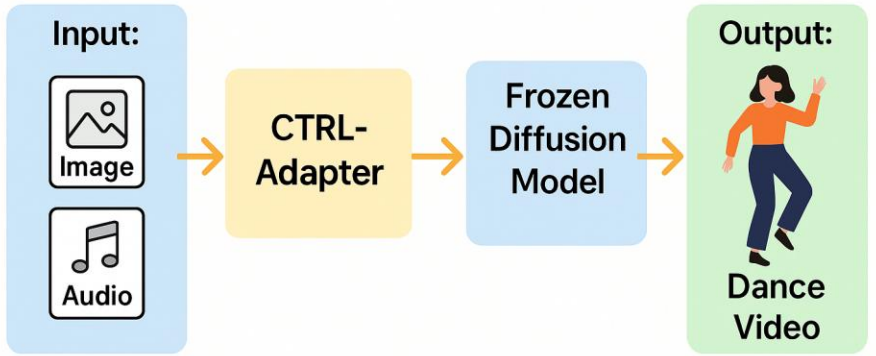
Mentors: Kevin Zhai, Soumik Ghosh, Dr. Mubarak Shah

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.



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.

Swapped Text Encoder CLIP with Audio Encoder CLAP

- **Replaced Text Prompts with Audio Embeddings**: Instead of using natural language text prompts (CLIP) , we extracted precomputed CLAP audio embeddings (.pt files) for each sample.

- **Modified Prompt Handling**: Replaced the section that loads text captions with logic to load audio embedding tensors from .pt files

- **Updated Pipeline Input**: For models like svd , we passed the prompt_embeds (from CLAP) into the pipeline while setting the original prompt argument to "." or None, effectively bypassing the text encoder.

CLAP Audio Encoder: A Simple Overview

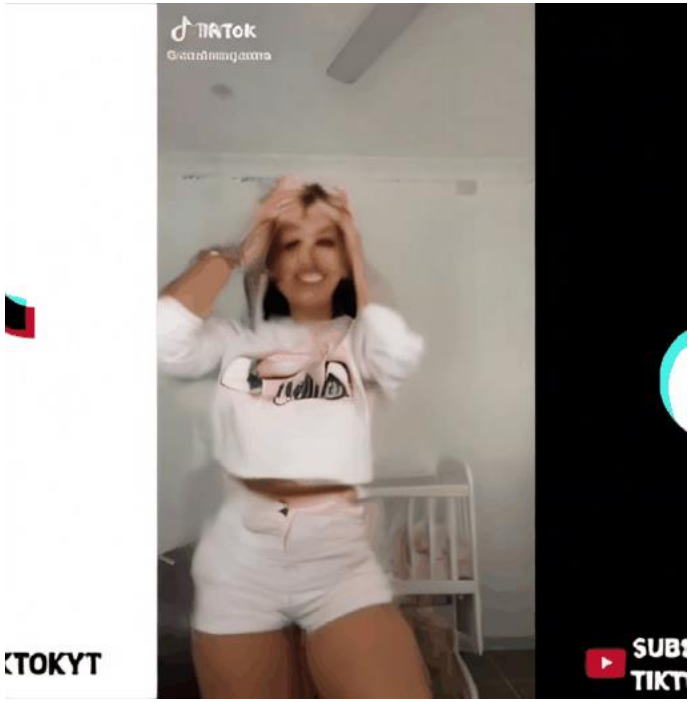
- **Audio to Vector Conversion**: CLAP takes an audio clip (like a music or speech waveform) and converts it into a fixed-size numerical vector that captures the important features of the sound.

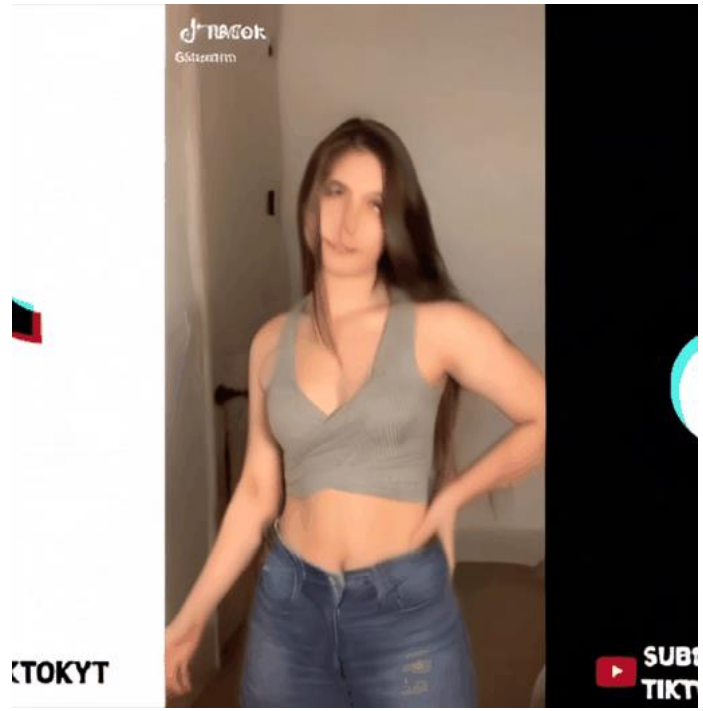
- **Trained with Text Alignment**: During training, CLAP learns to align audio with matching text (e.g., “a dog barking” ↔ sound of dog barking) using contrastive learning, so similar sounds and descriptions have similar embeddings.

- **Useful for Machine Learning Models**: The resulting audio embeddings can be used directly in models (like diffusion or classification) as a rich, compressed representation of the sound—just like how CLIP is used for images and text.

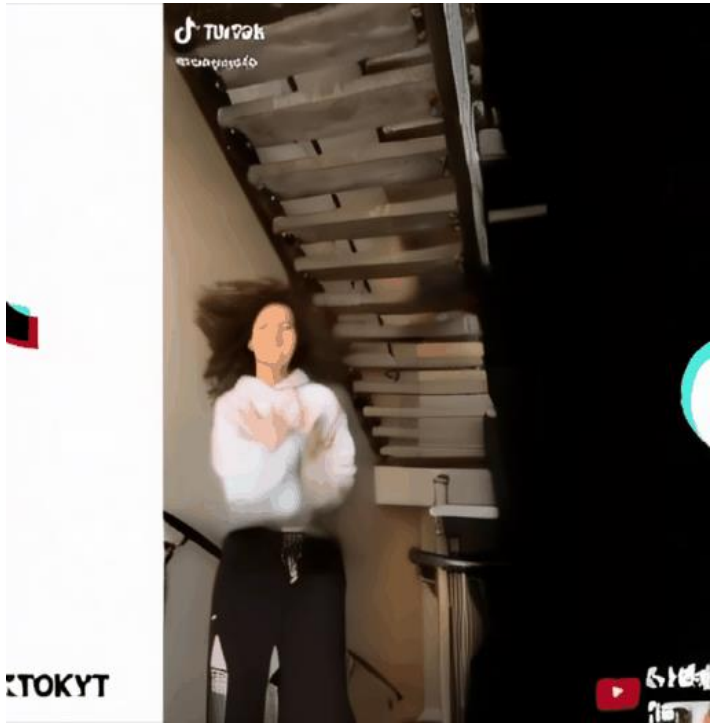
Input Frame = 1 , control guidance = 0, Output Frames = 8

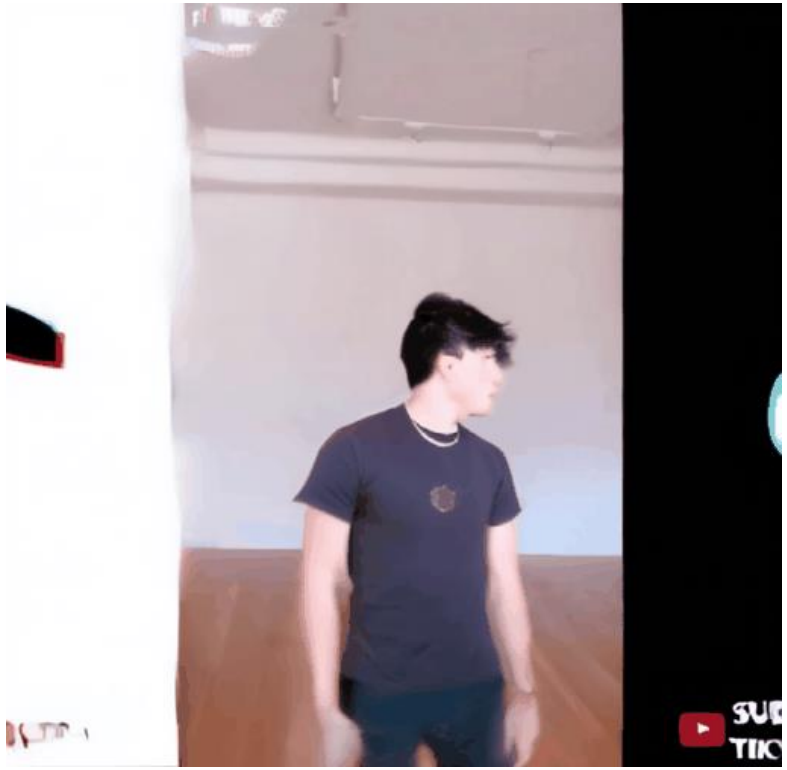


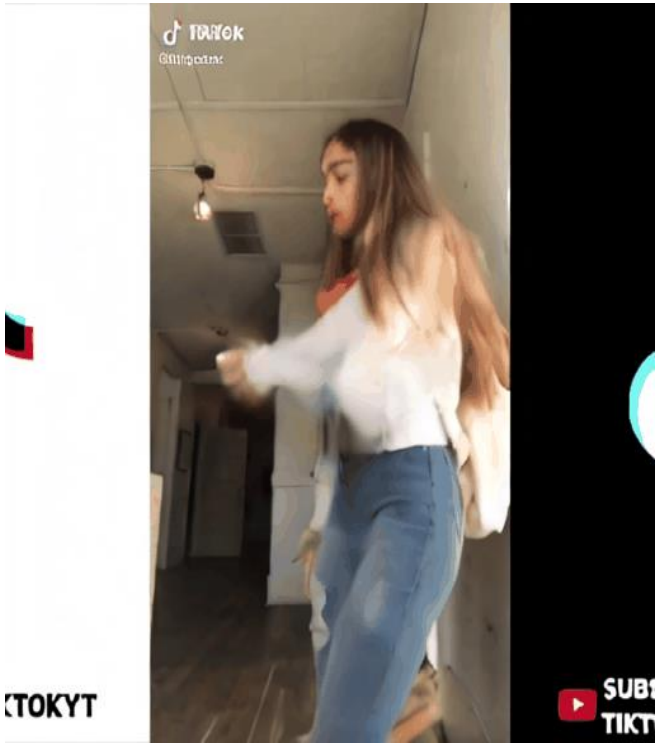




Input Frames = 1, Control Guidance = 0, Output Frames = 8

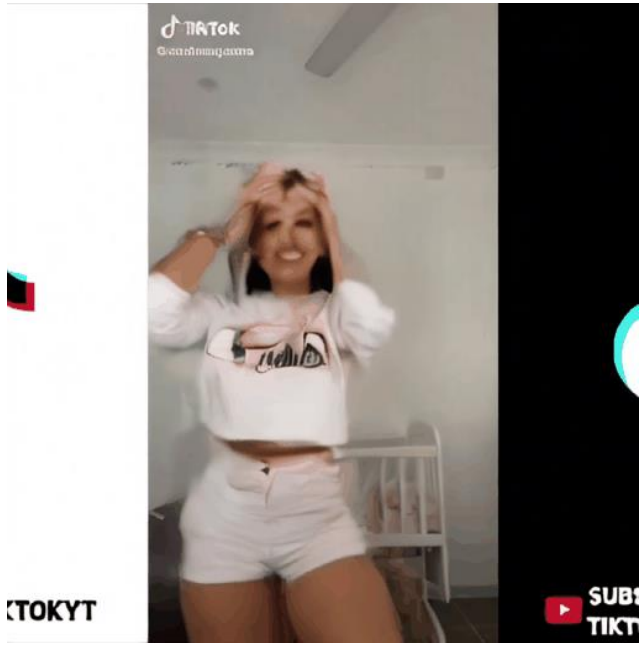


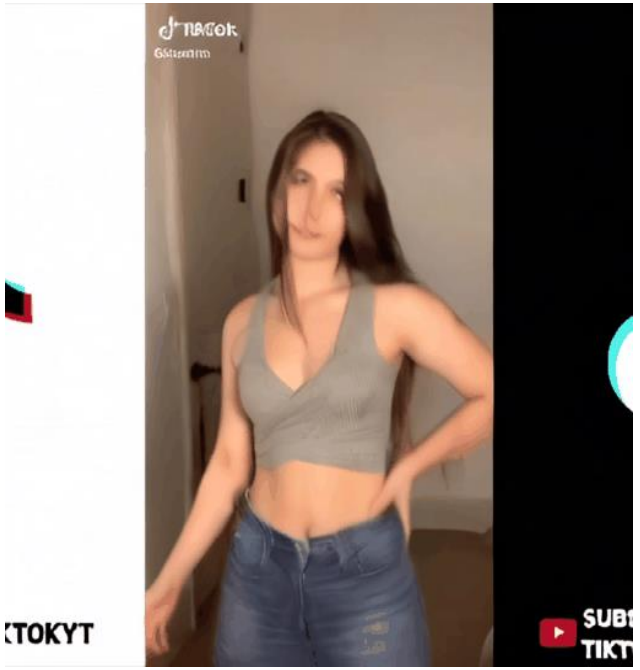




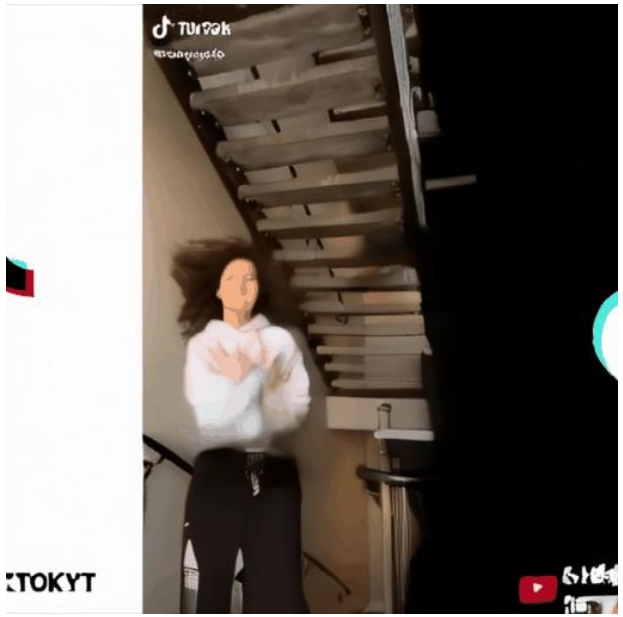
Input Frames = 14, control guidance = 0, Output Frames = 8

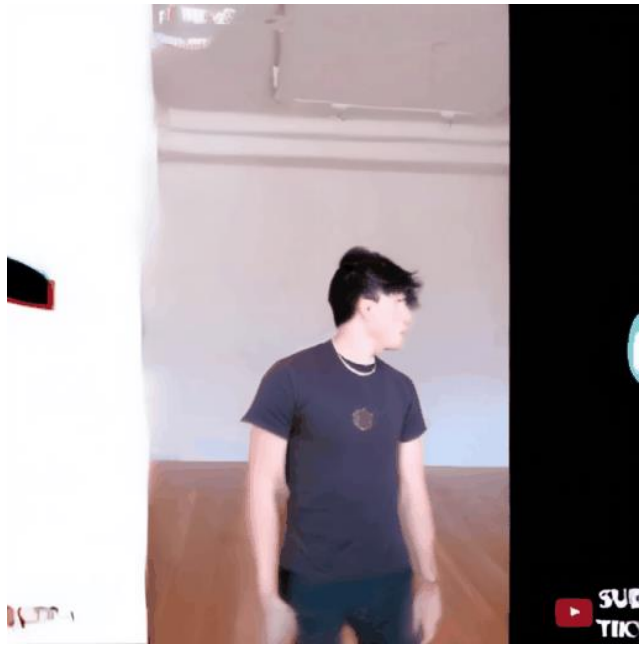


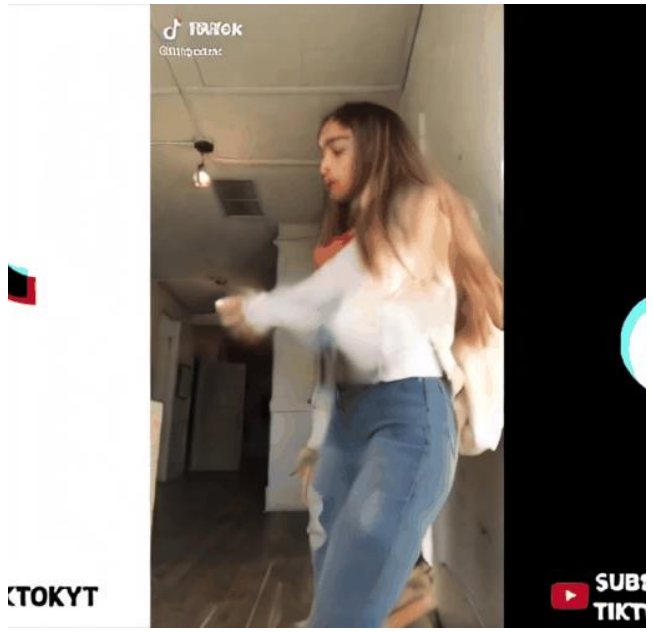




Input Frames = 14 , control guidance 0, Output Frames = 8







Plan for Next Week

- Run baseline experiments on the entire dataset to verify end-to-end data and pipeline consistency
- Start fine-tuning the model
- Finish writing the report and the poster

THANK YOU