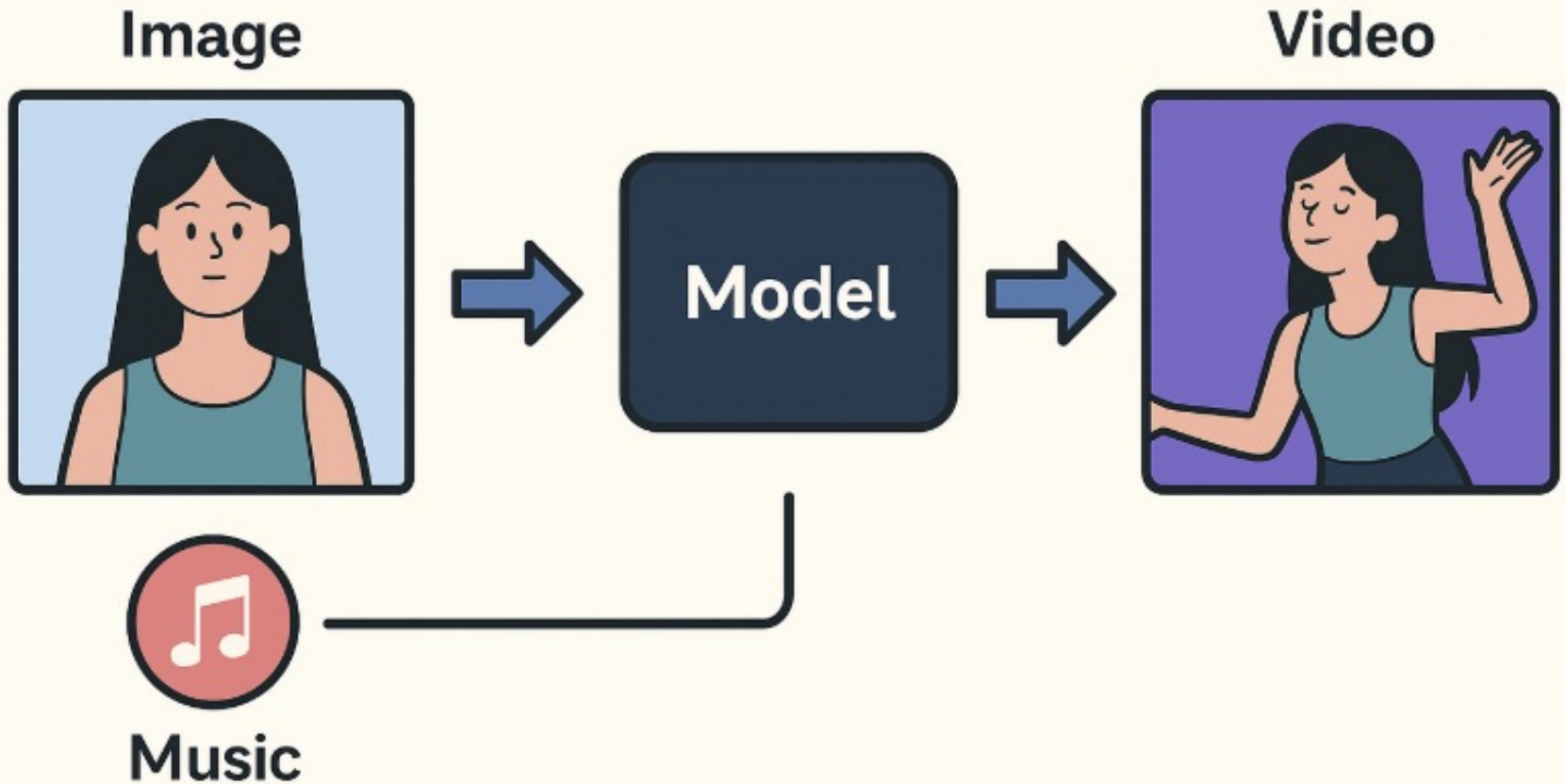
UCF

CENTER FOR RESEARCH IN COMPUTER VISION

Audio-Visual Conditioned Dance Video Generation with VQ-VAE and Diffusion Transformers

By Neel Shah & John Benjamin

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



Progress So Far

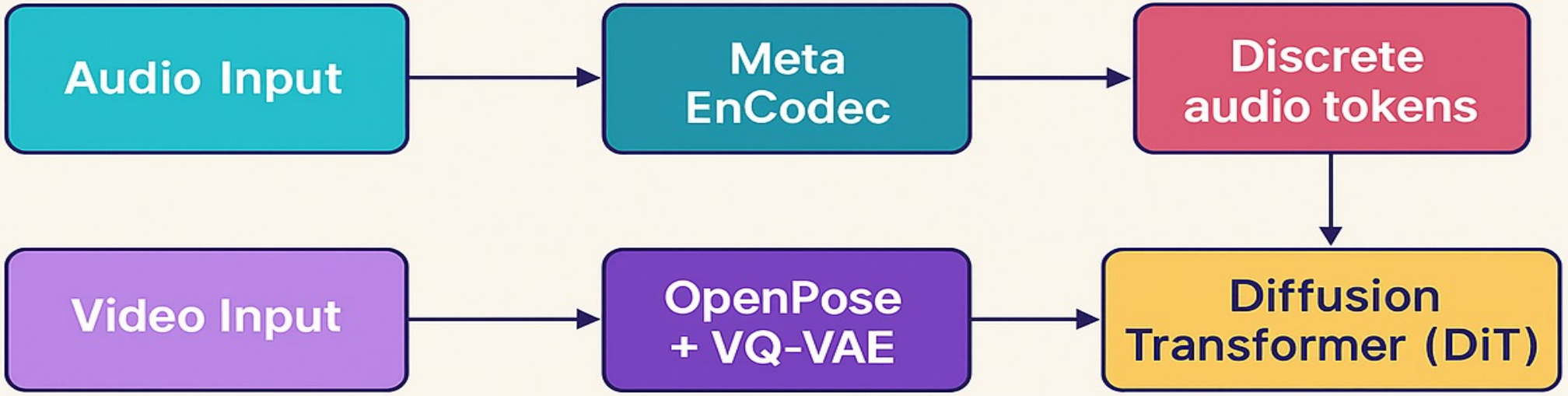
- Completed **literature review** and designed the **full architecture** for audio-visual conditioned generation
- Collected and preprocessed dataset:
  - Trimmed video clips to **10 seconds**
  - Ensured **single-person dancing** per clip
  - Verified **continuous music** in each segment
- Extracted **2D poses** using **DWPOSE**
- Trained **VQ-VAE** to convert poses into discrete tokens
- Used **Meta's EnCodec** to generate audio token embeddings
- Successfully loaded and tested **DiT-XL/2** model from the *Scalable Diffusion Transformers* paper

🎯 Goals for Next Week

- Implement **contrastive learning** for pose and audio token alignment
- Begin **training the DiT model** with cross-modal conditioning

Project Architecture Overview

System Pipeline



🧠 **Key Idea:** Rather than using a shared encoder, we use modality-specific tokenizers and allow the DIT to learn laten structure.

Audio Tokenization with Meta EnCodec

🧠 Tool: Meta EnCodec4kHz

- Pretrained audio codec model from Meta AI
- Operates at 50 kHz (required for EnCodec)

Preprocessing Pipeline

Convert stereo → mono

Resample to 24 kHz (required for EnCodec)

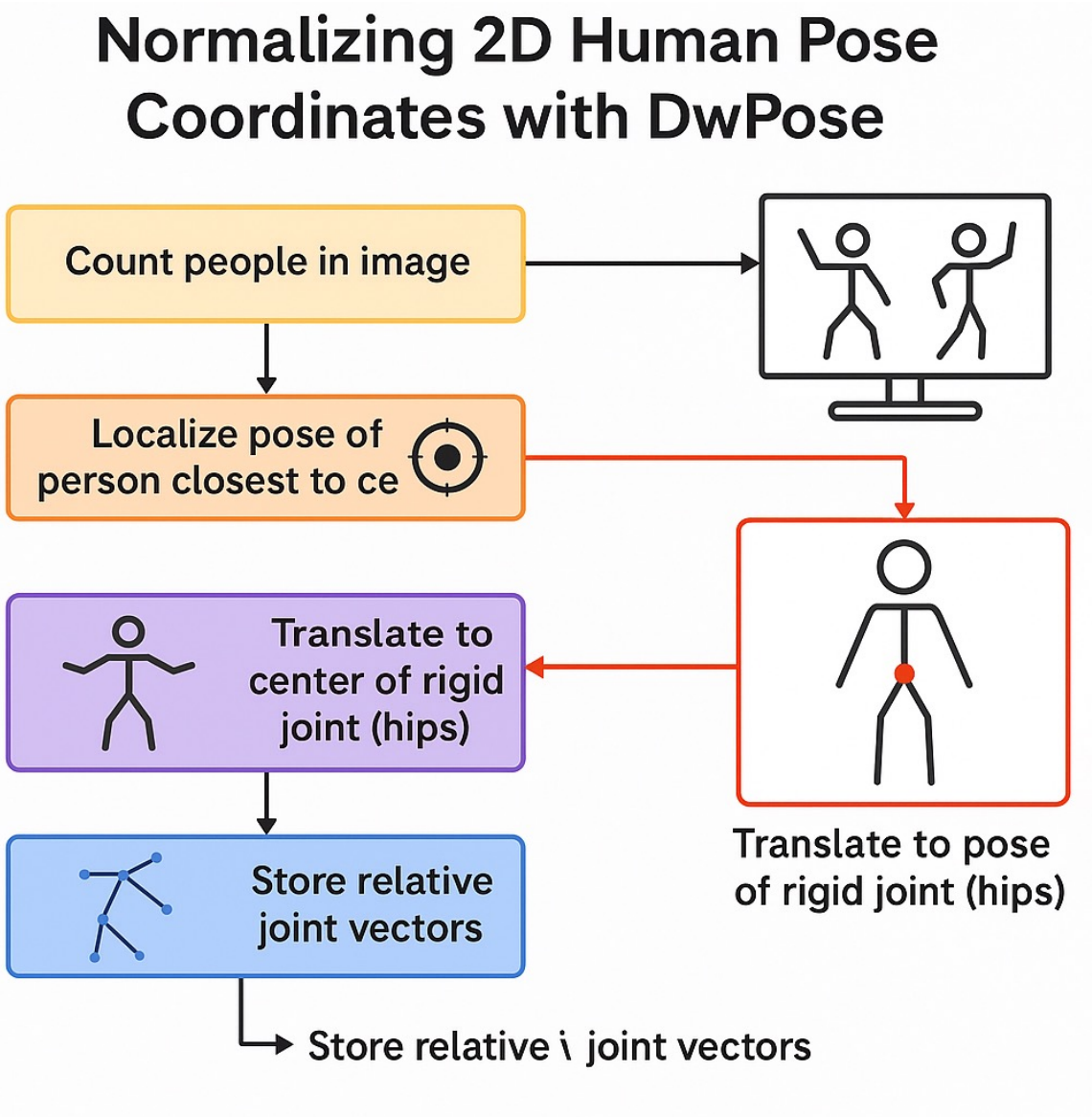
Encode with EnCodec

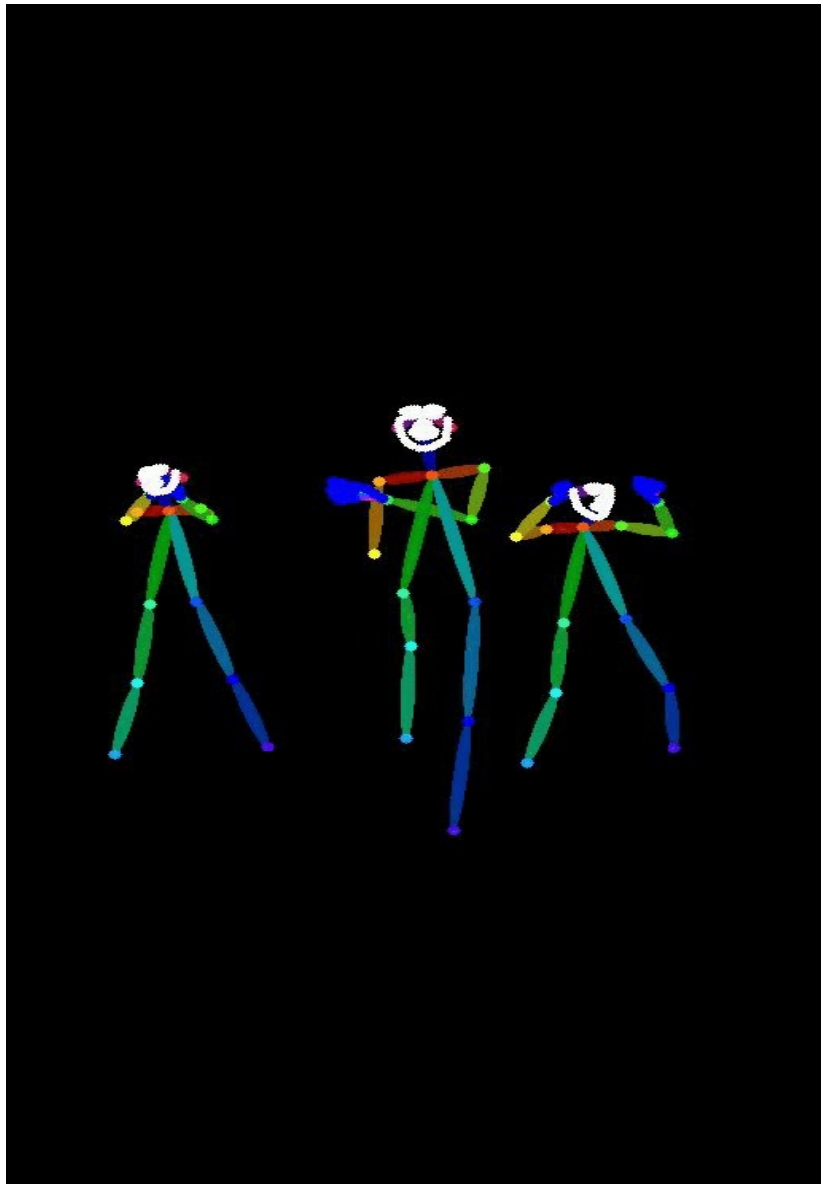
Output stored as .npz with shape (T, 32)

✅ Benefits

- Highly compressed yet semantically rich audio representation
- Retains musical structure (beat, rhythm, energy

Normalizing 2D Human Pose Coordinates with DwPose





Pose Tokenization with VQ-VAE

Pose Estimation

- Extract 2D keypoints using OpenPose or MediaPipe @ 25 FPS
- Normalize joint coord-inates (zero-centered, scale-invariant)

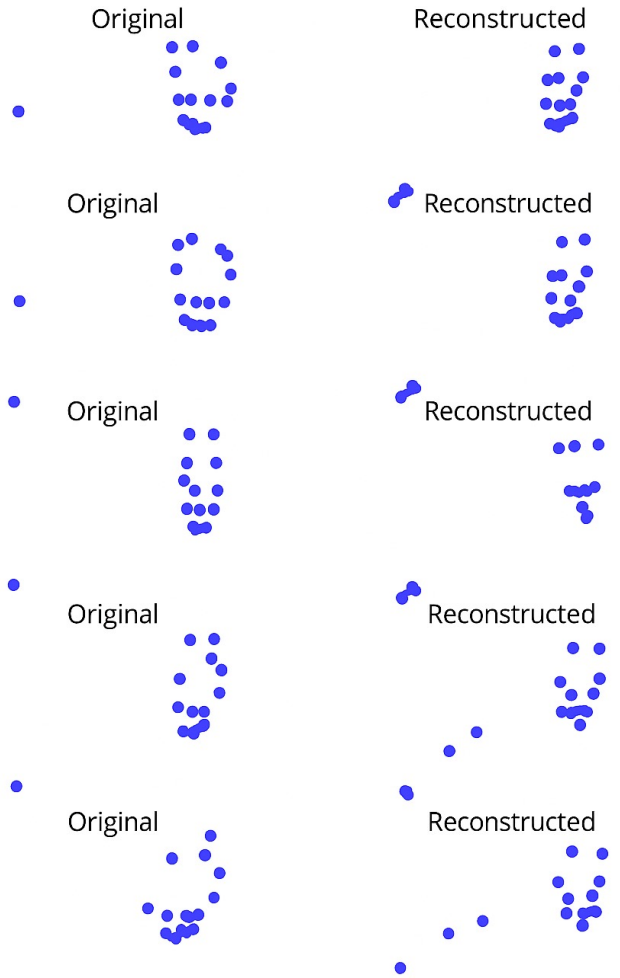
2D Keypoints → VQ-VAE → Pose tokens

Why Part-wise Tokenization?

- Captures high-frequency motion details (e.g., hands)
- Improves expressivity and decoding fidelity

Pose Point Reconstruction

Evaluating how well the VQ-VAE model reconstructs 2D human poses from its discrete tokens.



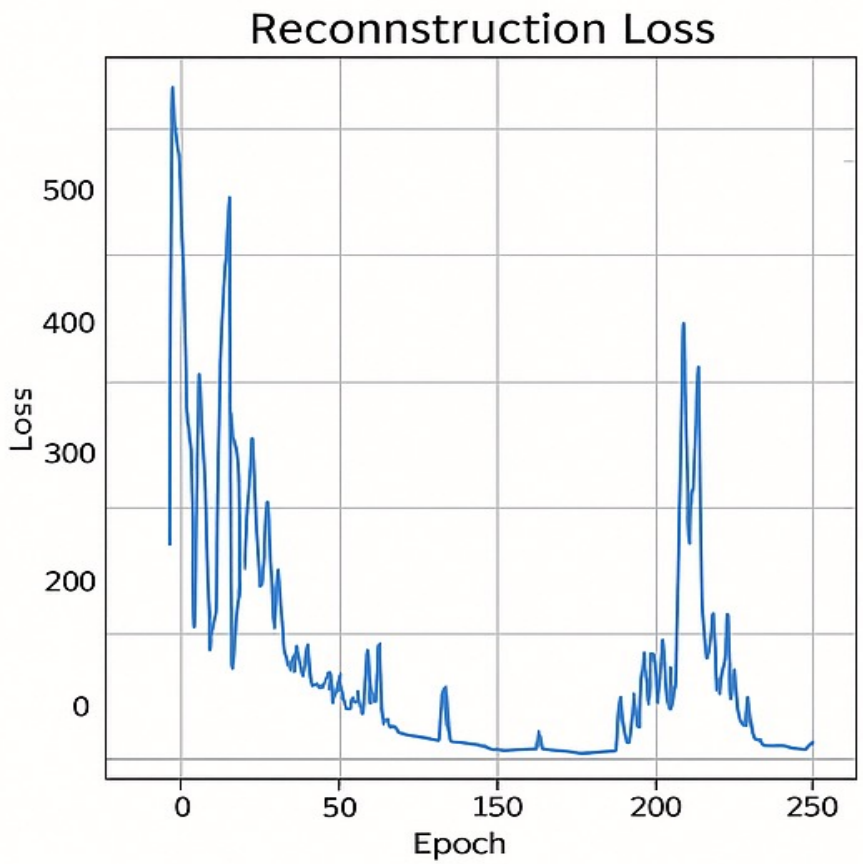
Results and Observations

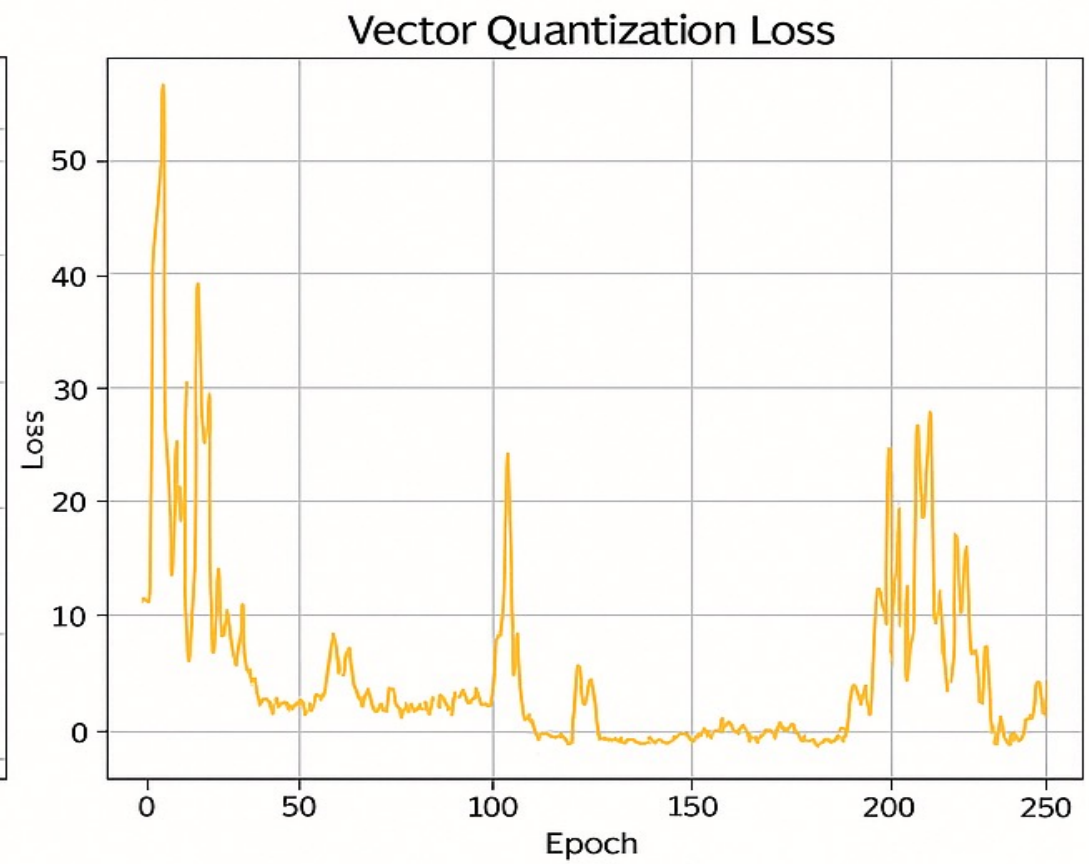
Training Setup

- Epochs: 250
- Optimizer. Adam, LR = 16 - 3
- Loss = MSE (reconstruction) + VQ loss

Results Summary:

- Steady drnLP): [36] ⟶ [128] -[3]
- L\erned PyTorch





DIT Model and Contrastive Learning Plan

- **Model Used**: DiT-XL/2 (from *Scalable Diffusion Models with Transformers*, 512×512 pretrained checkpoint)
- **Purpose**: Leverage powerful transformer-based diffusion model for generating realistic human motion sequences from pose tokens
- **Input Modalities**:
- Pose Tokens (from VQ-VAE)
- Audio Tokens (from Meta's EnCodec)
- **Goal**: Use **contrastive learning** to project both audio and pose tokens into a **shared latent space** before feeding into DiT
- **Why Contrastive?** Helps align pose and audio representations — enabling better temporal and semantic correlation for generation

Thank you