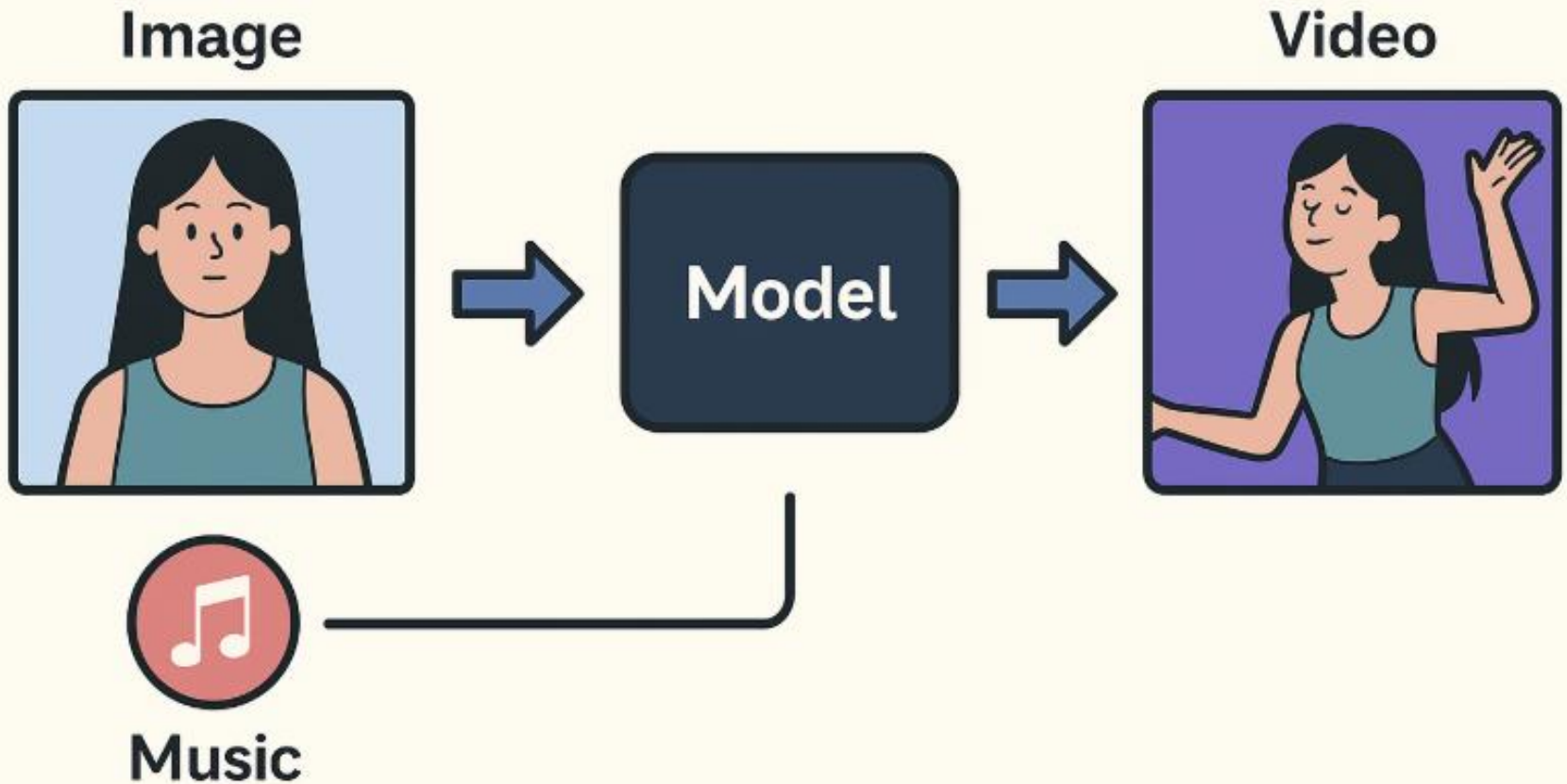
UCF

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
- Looked into different Datasets like AIST++

🎯 Goals for Next Week

- Begin **training the DiT model** with cross-modal conditioning

Pose Tokenization with VQ-VAE

Pose Estimation

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

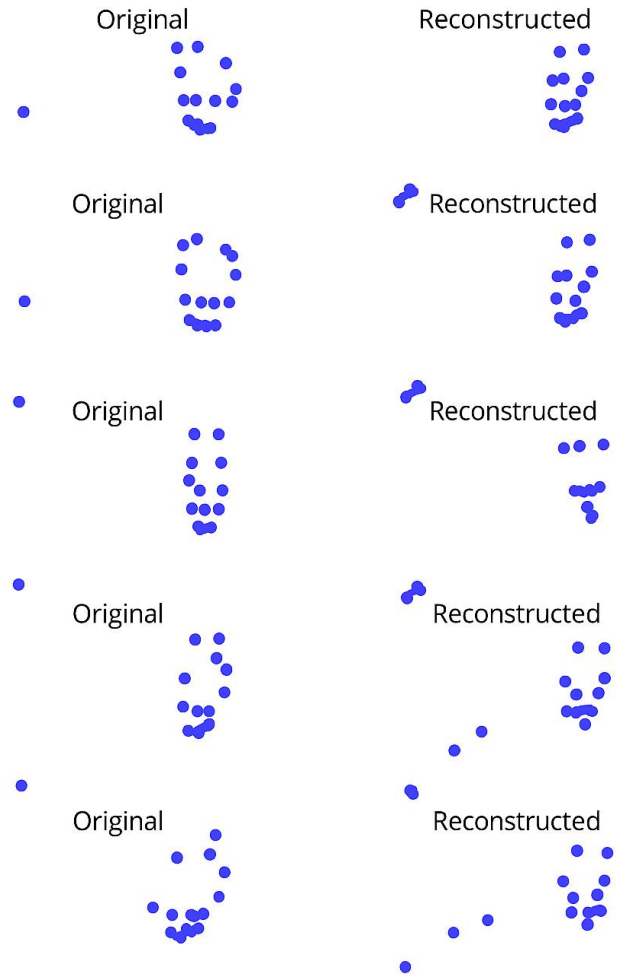
2D Keypoints → VQ-VAE → Pose tokens

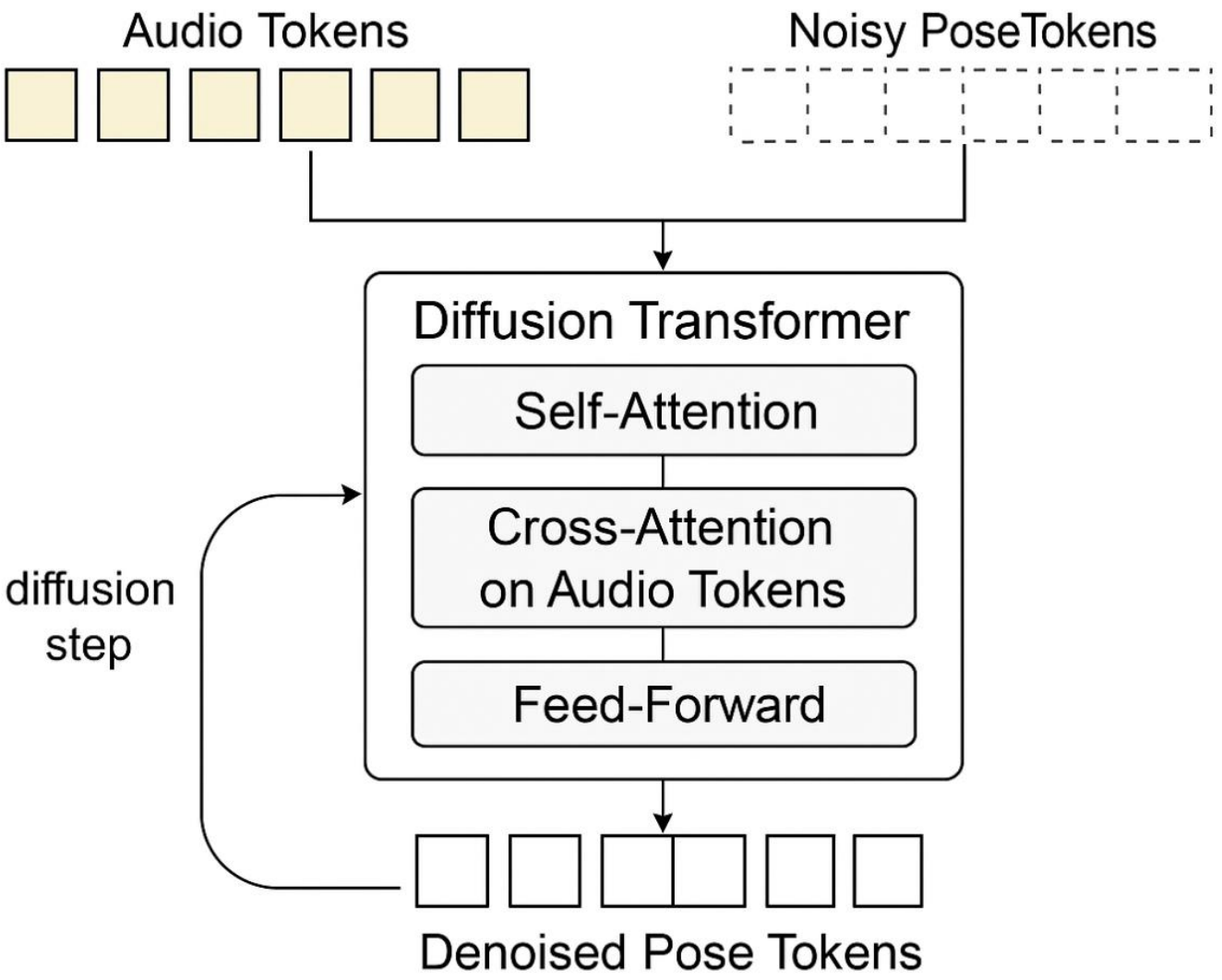
Why Part-wise Tokenization?

- 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.




Audio Tokens
Noisy PoseTokens
Diffusion Transformer
Self-Attention
Cross-Attention
on Audio Tokens
Feed-Forward
diffusion
step
Denoised Pose Tokens

Challenges & Our Contribution

- Challenge 1: Shared Latent Space (X-Dancer)
  Previous work forced audio and pose into a single latent space, limiting modality-specific details.
- Challenge 2: U-Net Backbone
  Used image-based U-Nets, not designed for long sequential token modeling.
- Our Approach: Diffusion Transformer + Cross-Attention
  We use a DiT backbone with separate audio and pose streams, linked by cross-attention for better temporal alignment.

Thank you