Scaling short-form internet chaos.

Motion-Conditioned Video Generation

AnimateAnything · StableAnimator · MimicMotion

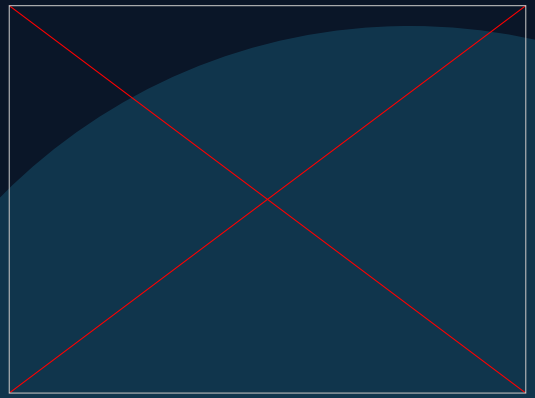
Soham Dasgupta | Zizheng Zhou

Motion-Conditioned Video Generation

Generating more TikTok videos!!

Motion-Conditioned Video Generation

Motion-Conditioned Video Generation



Background

Diffusion Models · Latent Diffusion

Diffusion Models: The Core Idea

Forward Process (Add Noise)

q(x□ | x□₋₁) — Gradually adds Gaussian noise over T steps

Reverse Process (Denoise)

pθ(x□₋₁ | x□) — A neural network learns to reverse noise step by step

Key Takeways

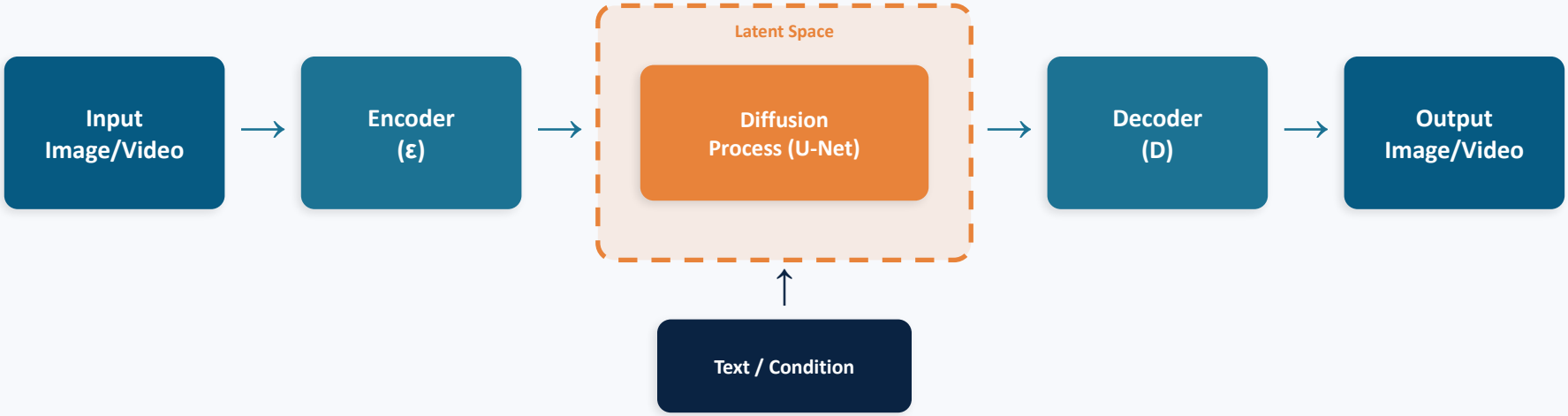
Training:
Learn to predict noise ε from noisy input x□

Loss Function:
$L = ||\varepsilon - \varepsilon\theta(x□, t, C)||^2$

where C = guidance condition (text, image...)

Latent Diffusion Models (LDMs)

Key idea: Run diffusion in a compressed latent space, not pixel space — much more efficient!



Efficiency

4–8× spatial compression reduces compute dramatically

Easier Manifold To Learn

It assume the entire image distribution lies on much smaller manifold

AnimateAnyone

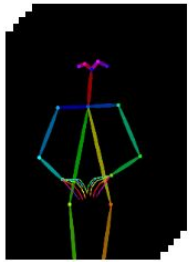
AnimateAnyone

Reference Image

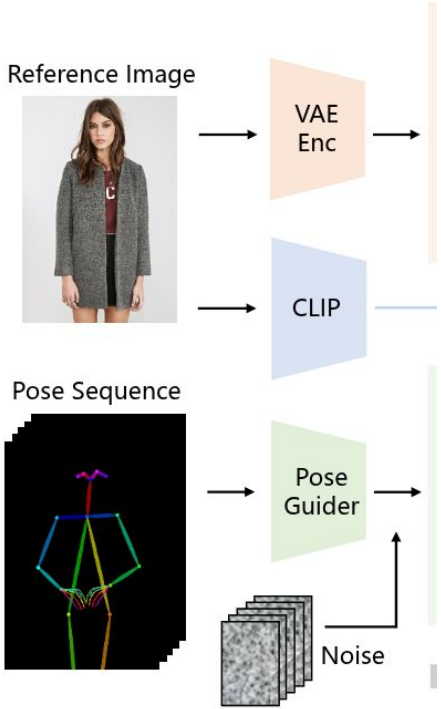
-

-

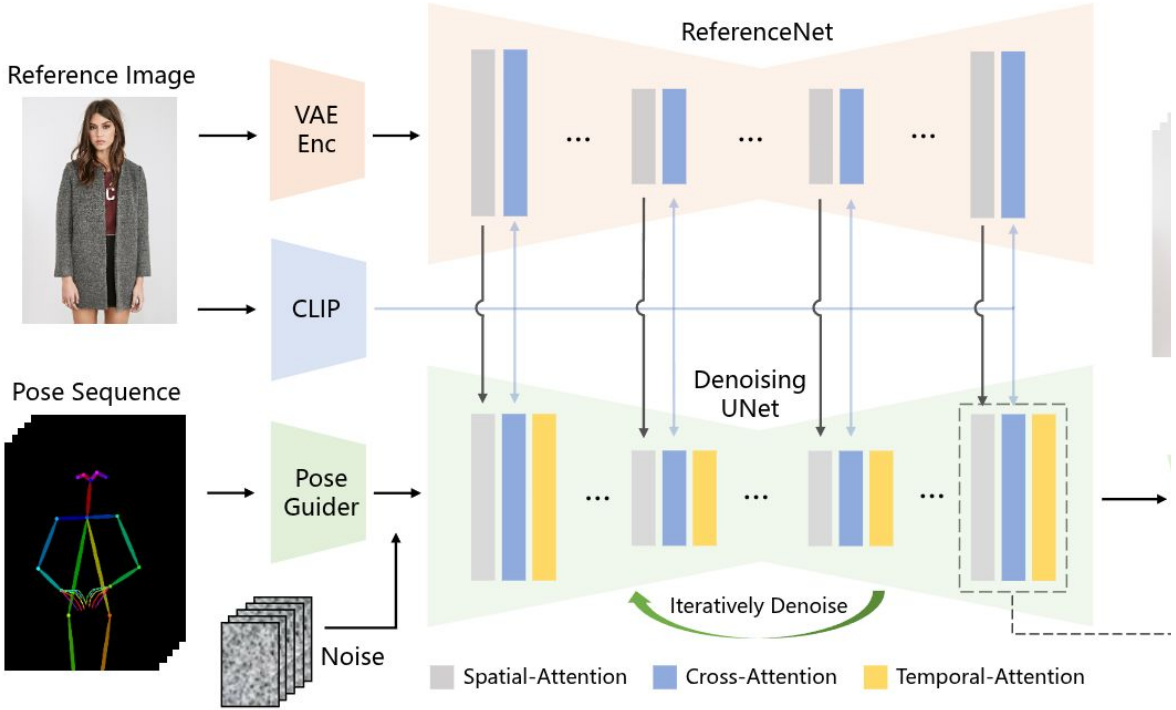
Pose Sequence



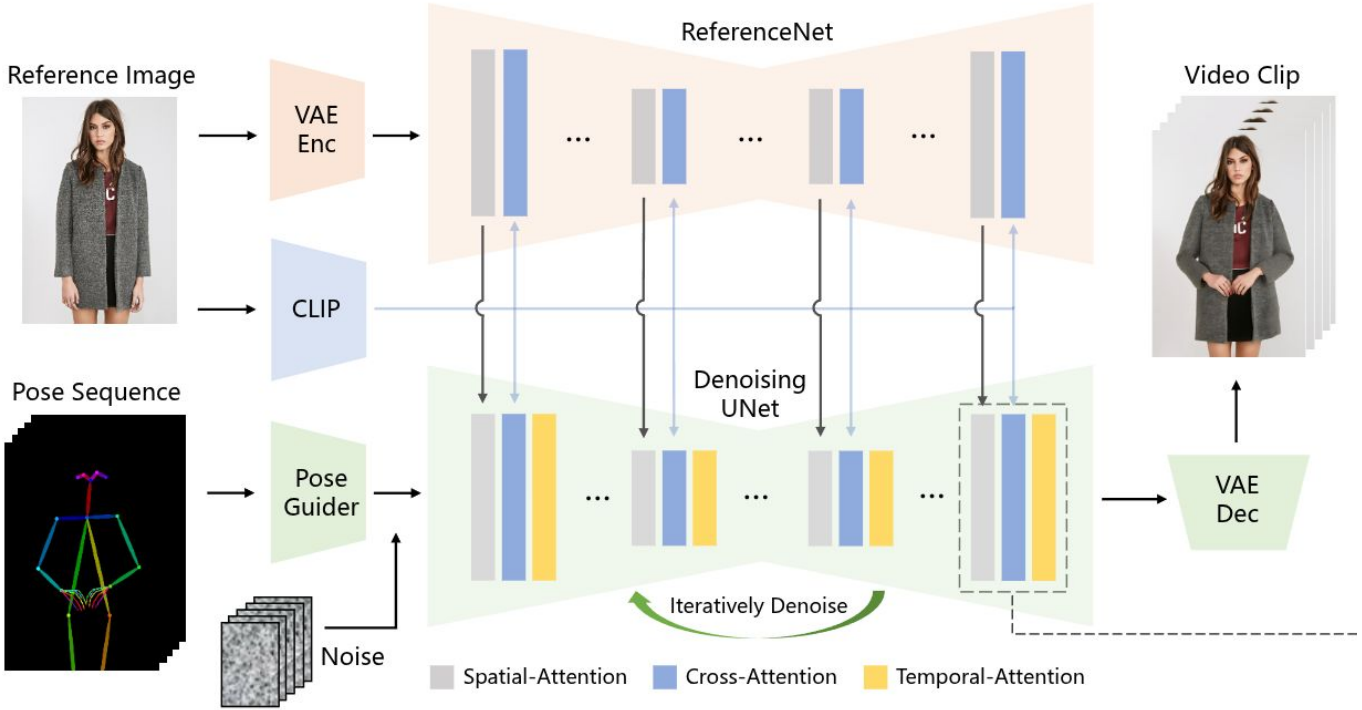
AnimateAnyone



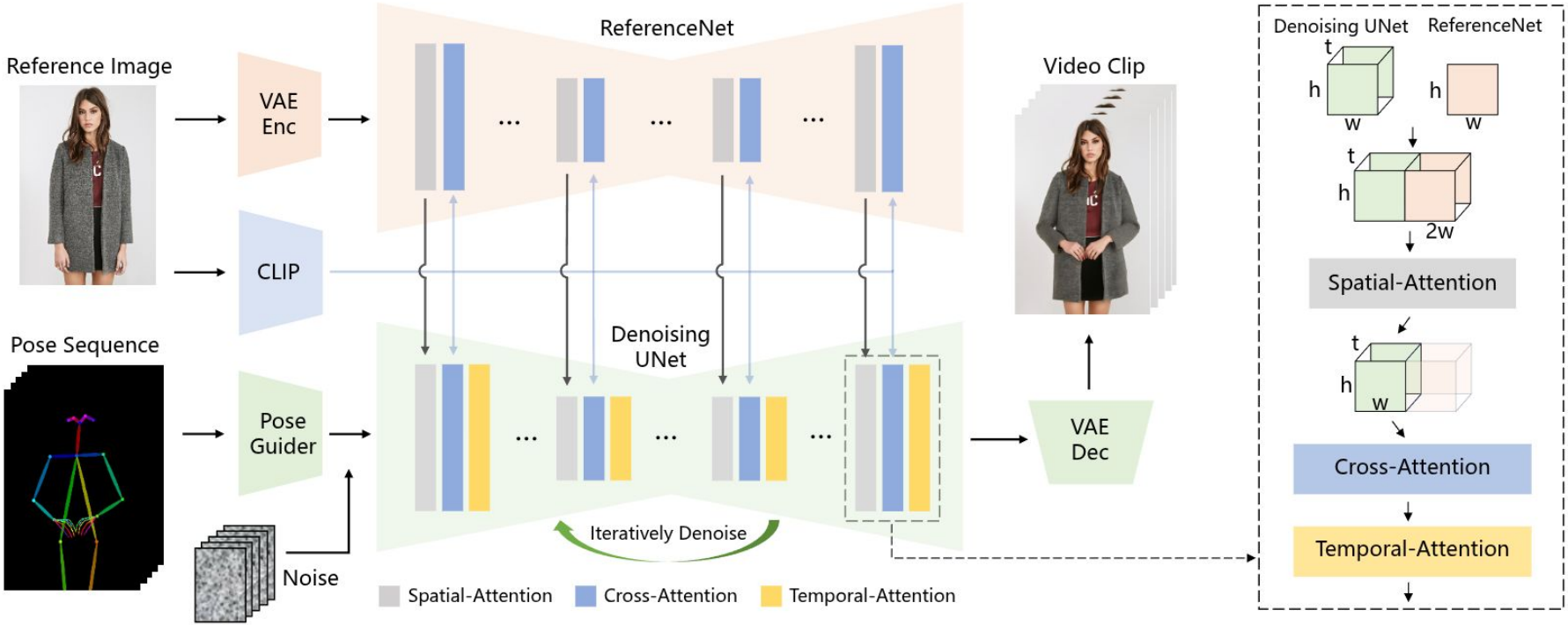
AnimateAnyone



AnimateAnyone



AnimateAnyone



MimicMotion (ICML 2025)

StableAnimator (CVPR 2025)

MimicMotion
(ICML 2025)

MimicMotion

Reference Image

Video Frames

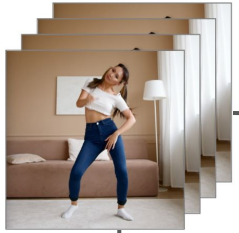
Inputs

MimicMotion

Reference Image

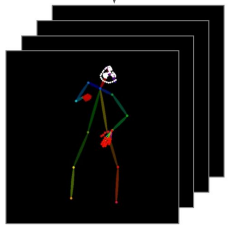
VAE Encoder

Video Frames

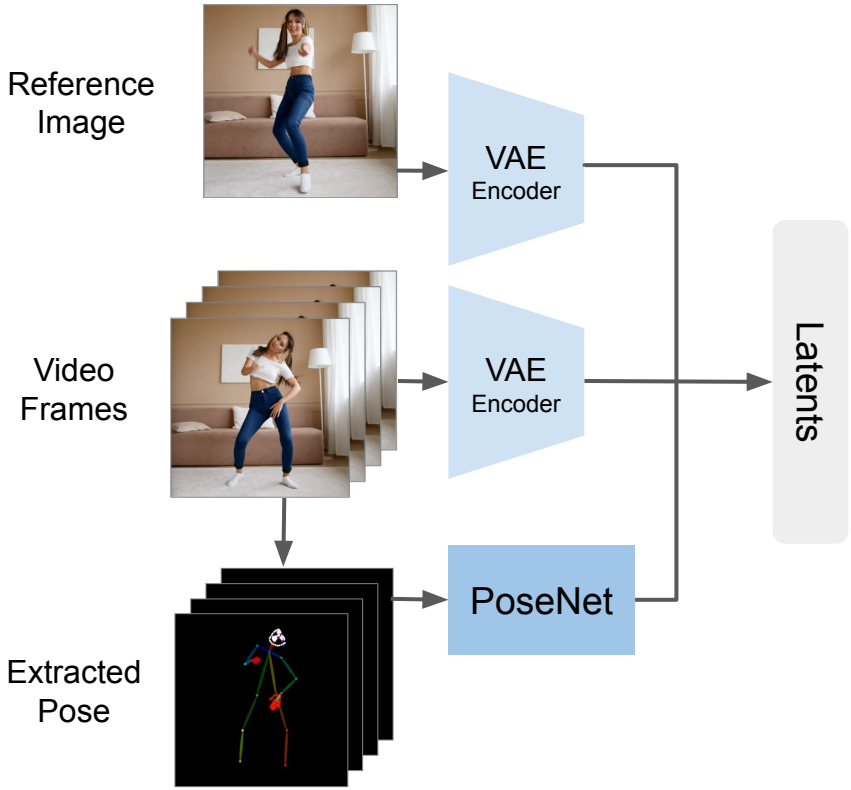


VAE Encoder

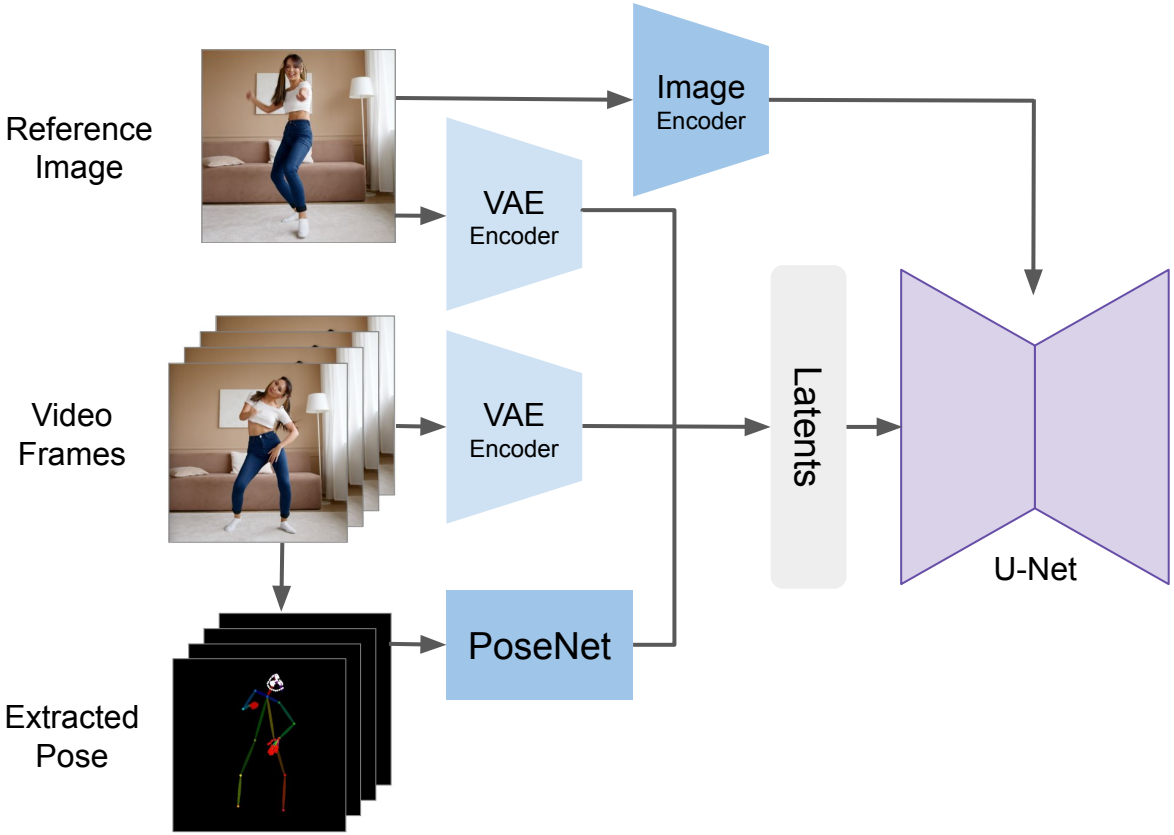
Extracted Pose



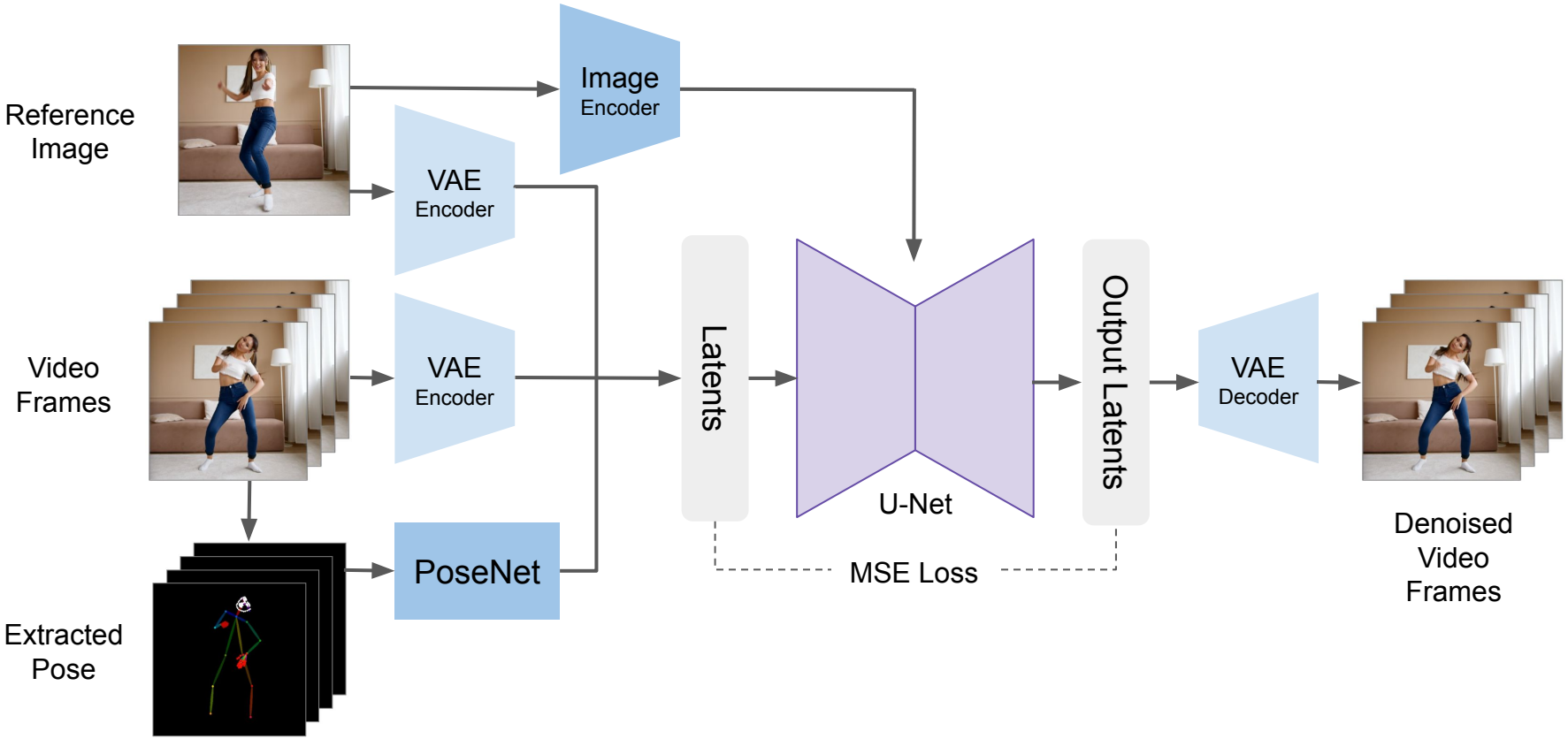
MimicMotion



MimicMotion



MimicMotion

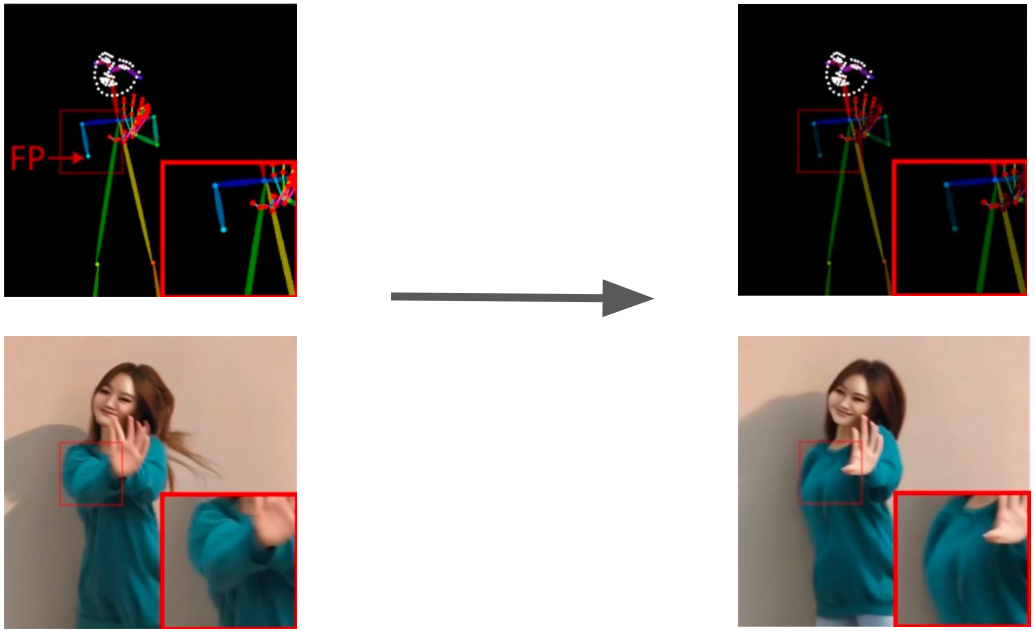


MimicMotion

1. The extracted pose sequence is **NOISY**
2. **DISTORTION** in certain regions
3. Unable to generate **LONG** videos

MimicMotion

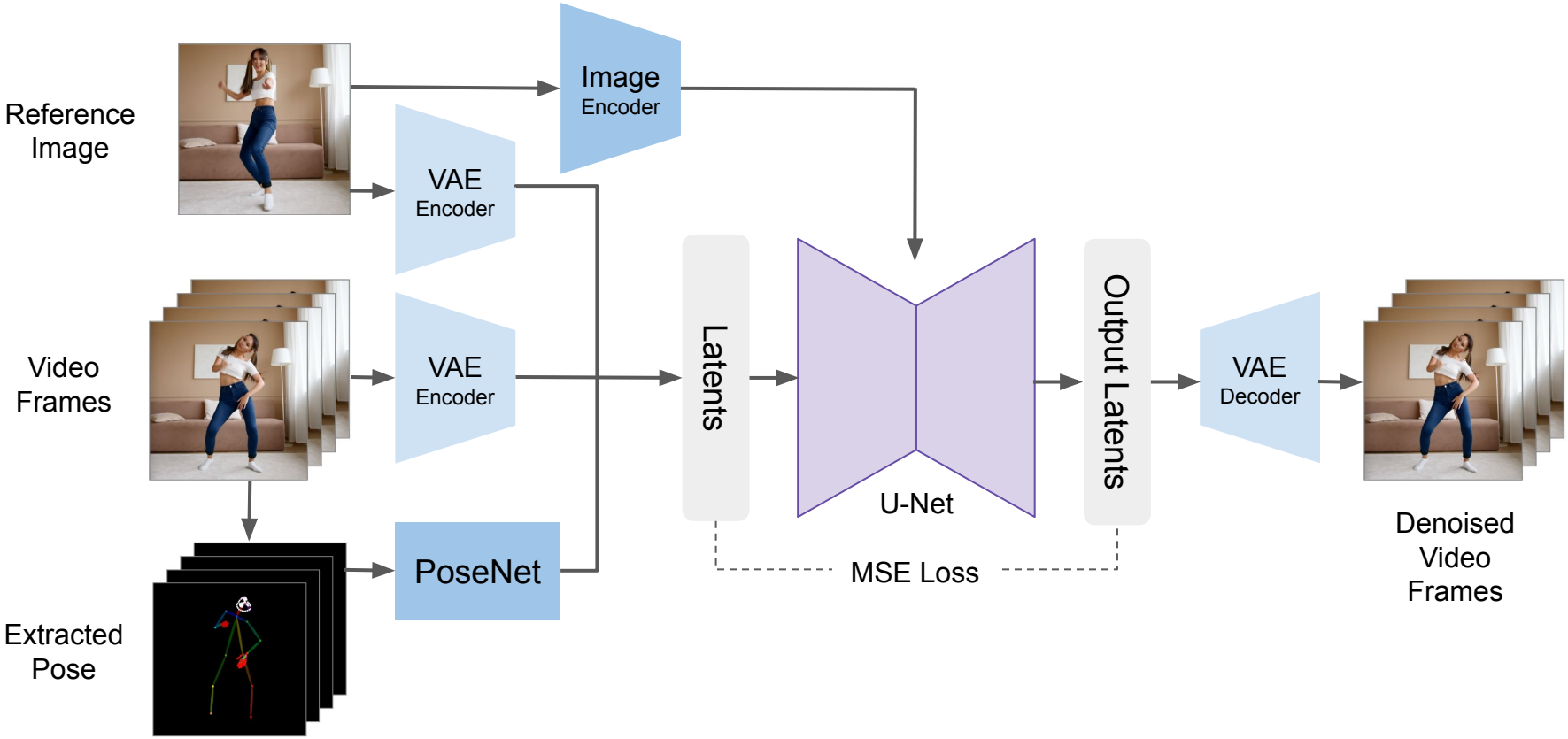
1. The extracted pose sequence is **NOISY**



Confidence-Aware Pose Guidance

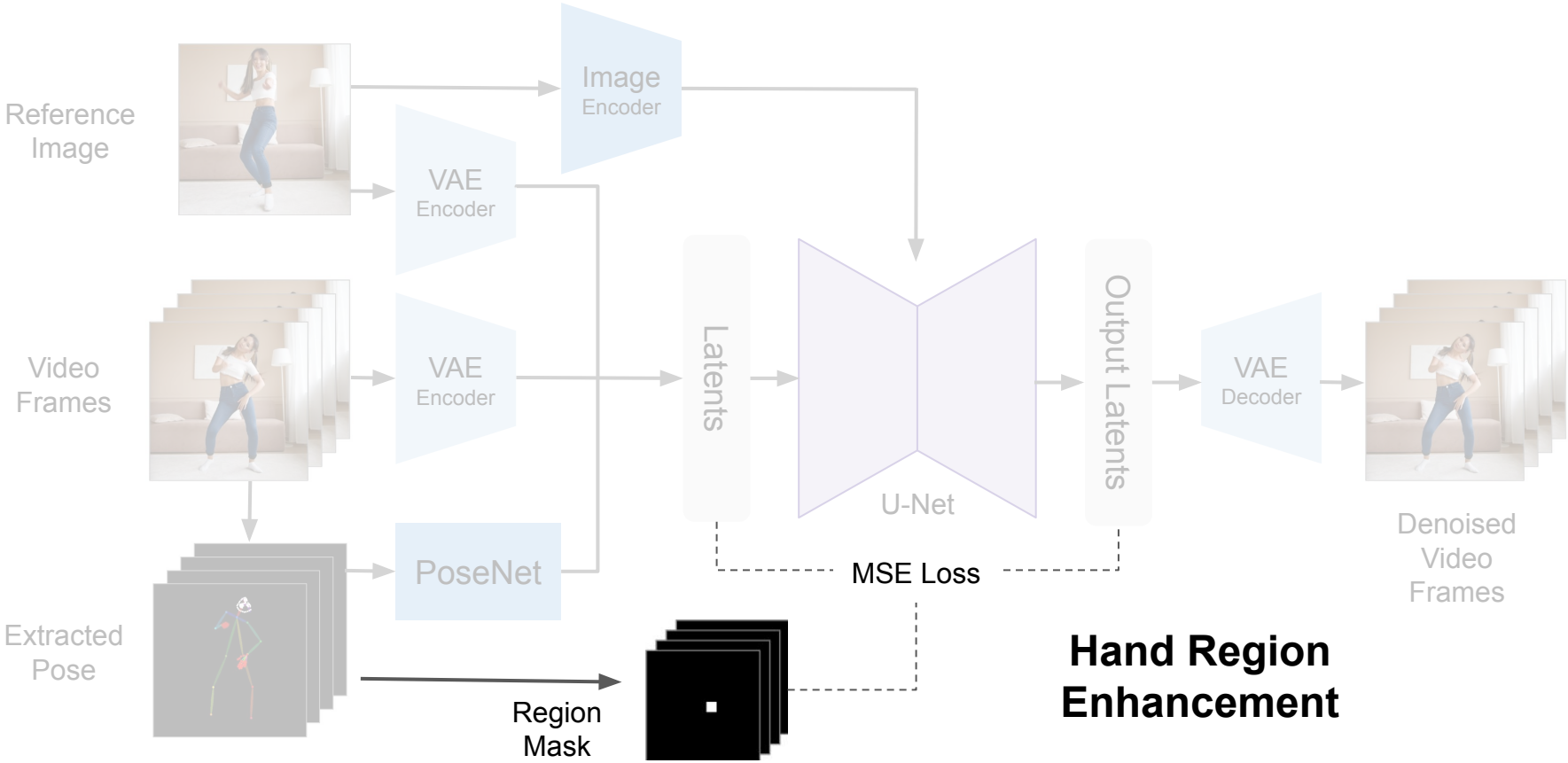
MimicMotion

2. **DISTORTION** in certain regions



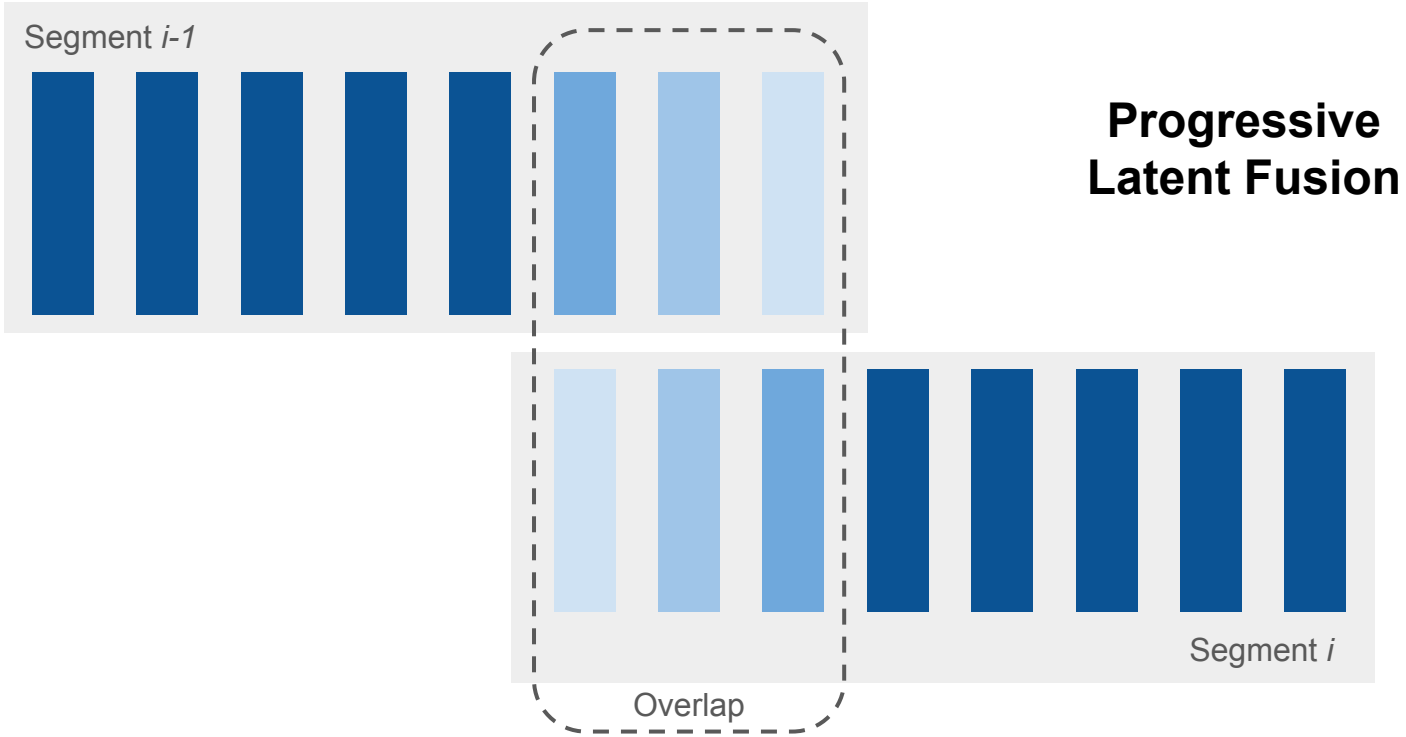
MimicMotion

2. **DISTORTION** in certain regions



MimicMotion

3. Unable to generate **LONG** videos



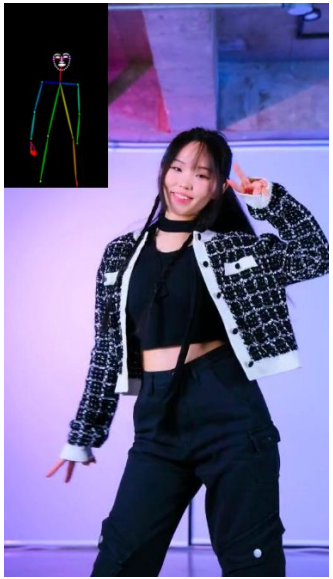
MimicMotion

1. The extracted pose sequence is **NOISY** → **Confidence-Aware Pose Guidance**
2. **DISTORTION** in certain regions → **Hand Region Enhancement**
3. Unable to generate **LONG** videos → **Progressive Latent Fusion**

StableAnimator
(CVPR 2025)

StableAnimator

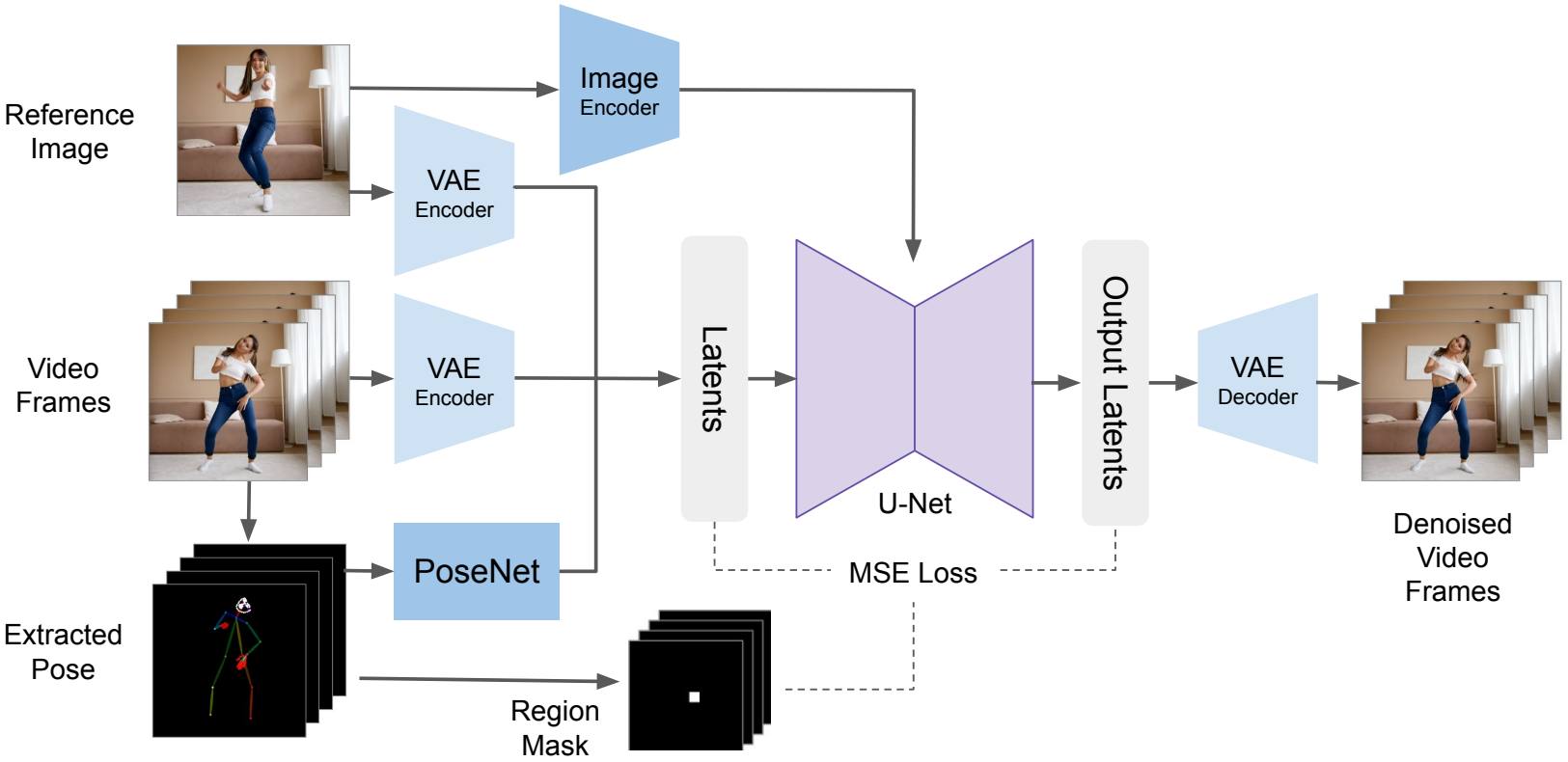
Bad Identity Consistency



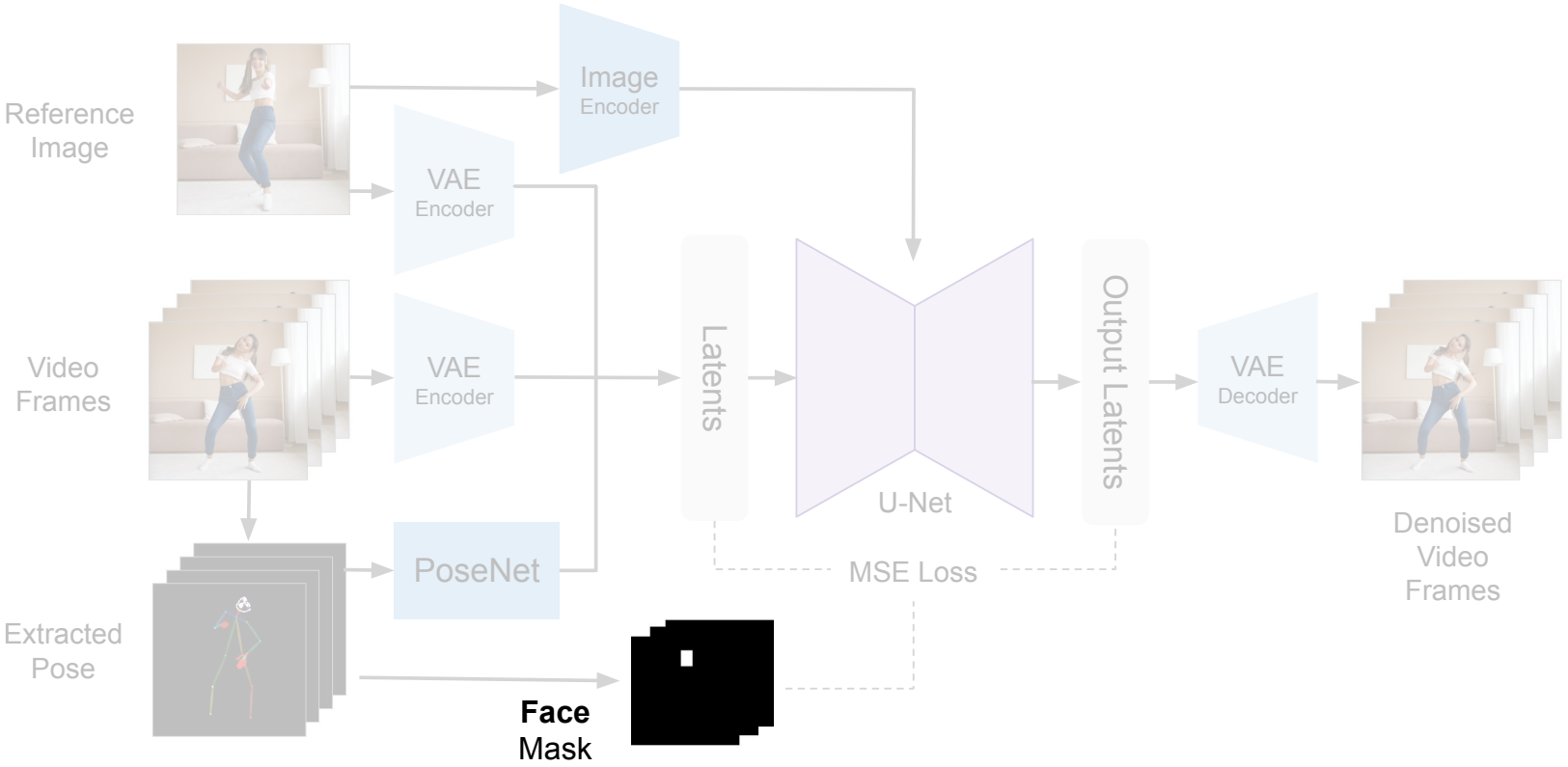
Reference Image

MimicMotion
(ICML 2025)

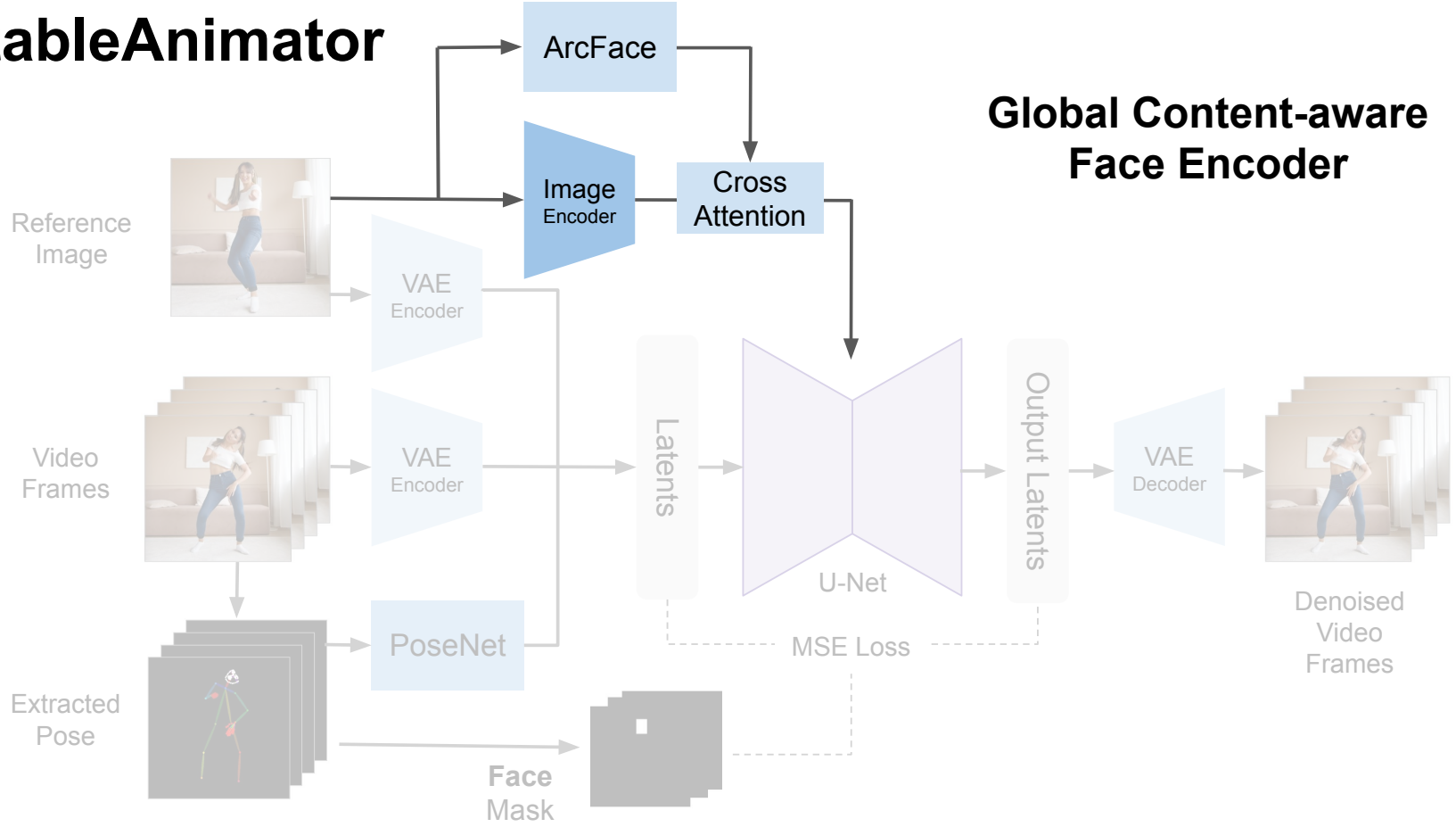
StableAnimator



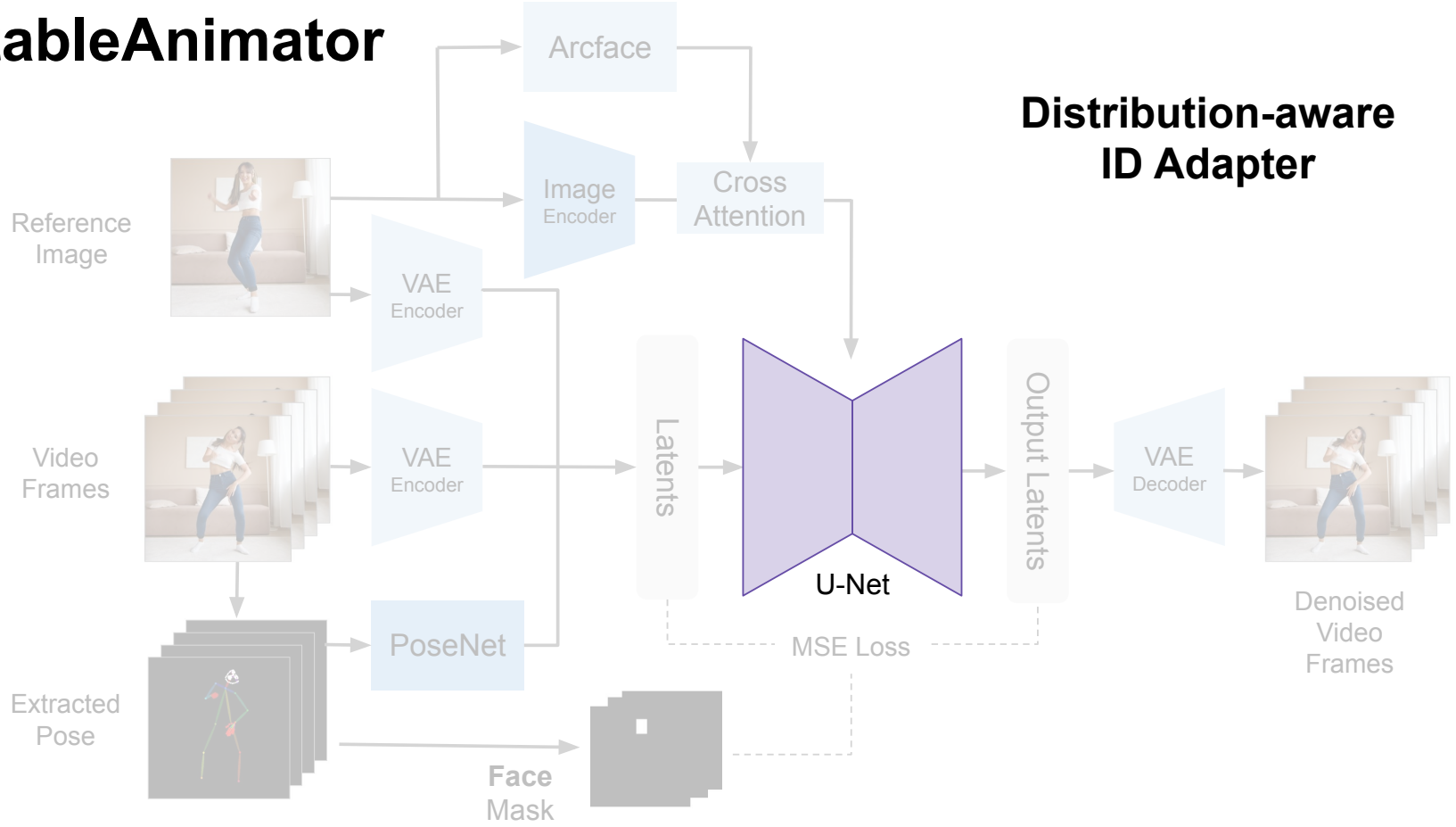
StableAnimator



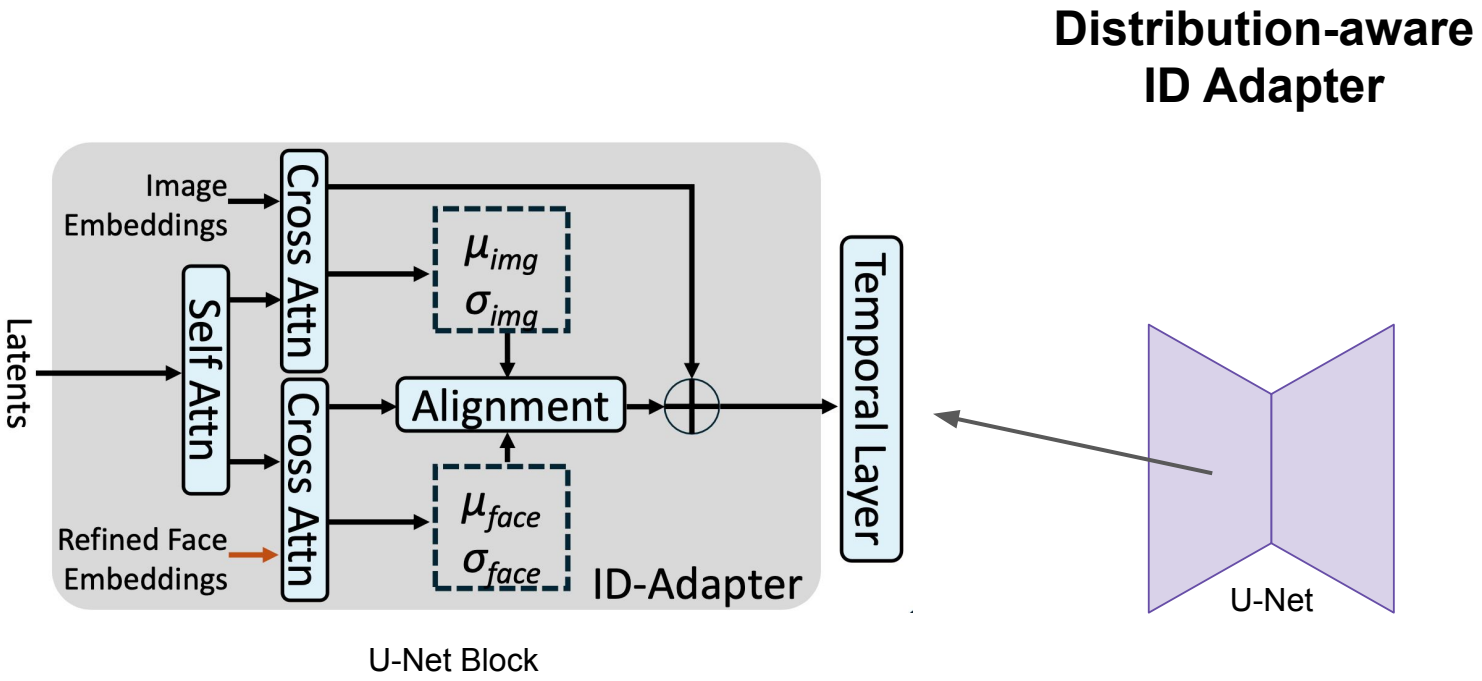
StableAnimator



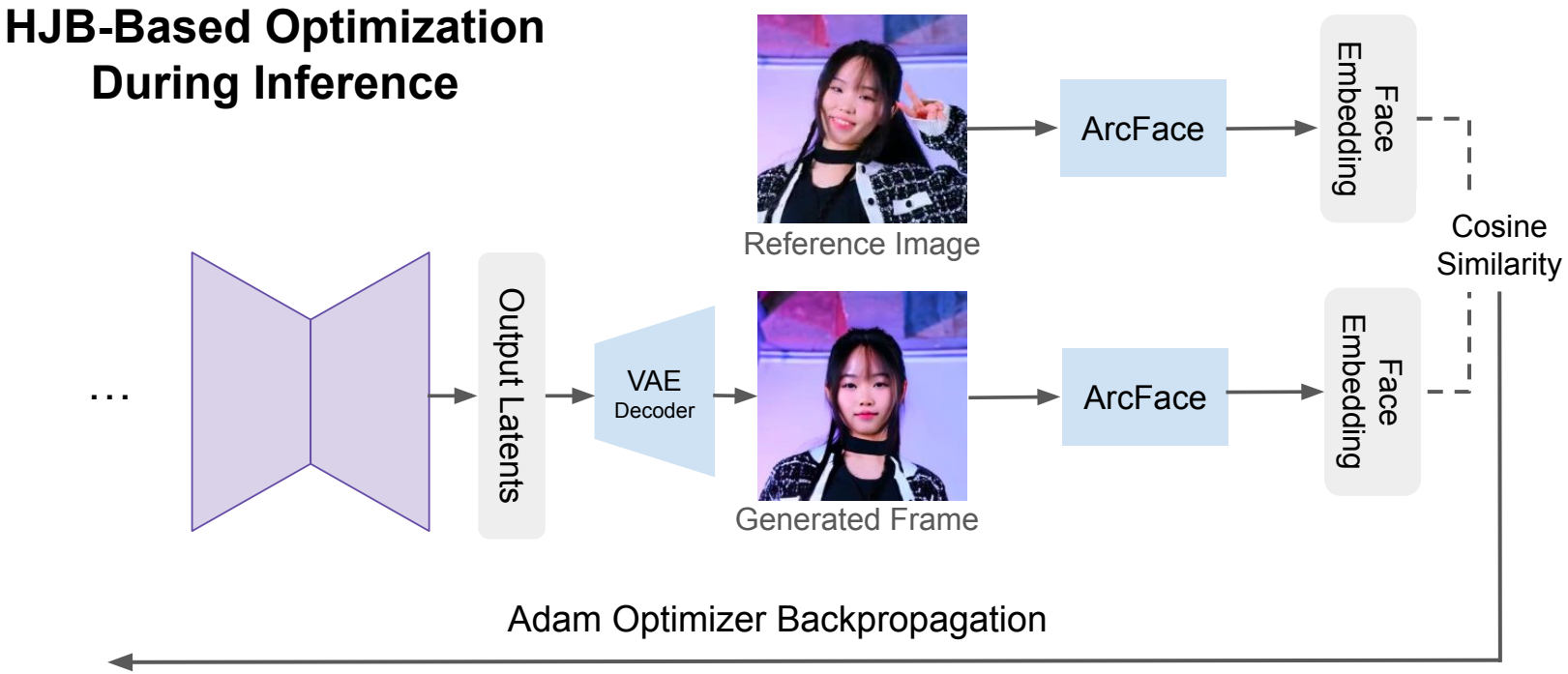
StableAnimator



StableAnimator

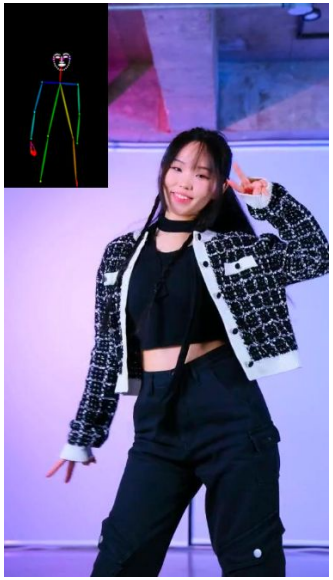


StableAnimator



StableAnimator

Better **Identity** Consistency



Reference Image

MimicMotion

Face Mask

Global Content-aware Face Encoder

Distribution-aware ID Adapter

HJB-Based Optimization

StableAnimator

Q & A