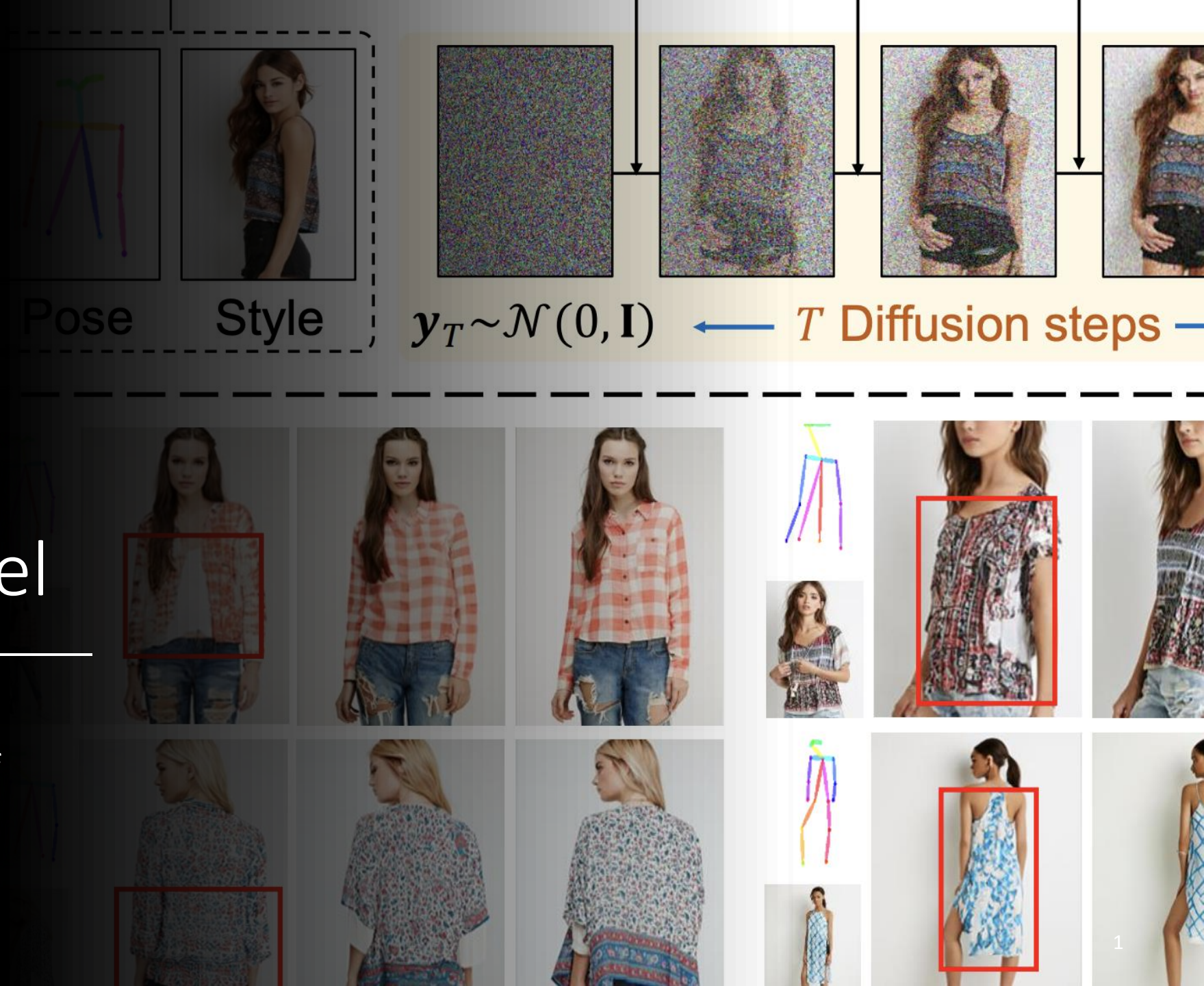
Person **Video** Synthesis via Denoising Diffusion Model

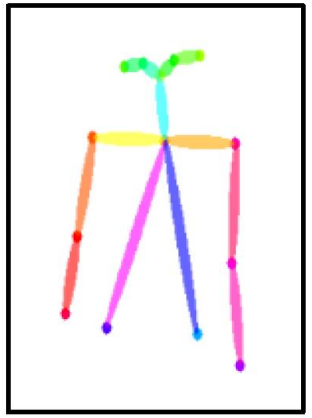
Manu S Pillai, Prudvi Kamtam,
Mukund Dhar and Adeel Yousaf

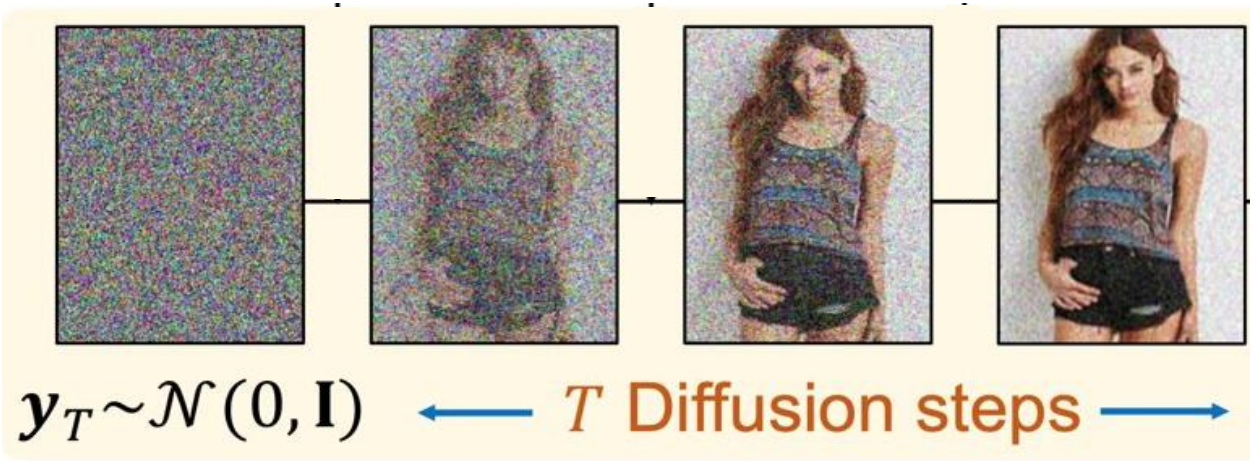


Diffusion based Image Synthesis

x_s

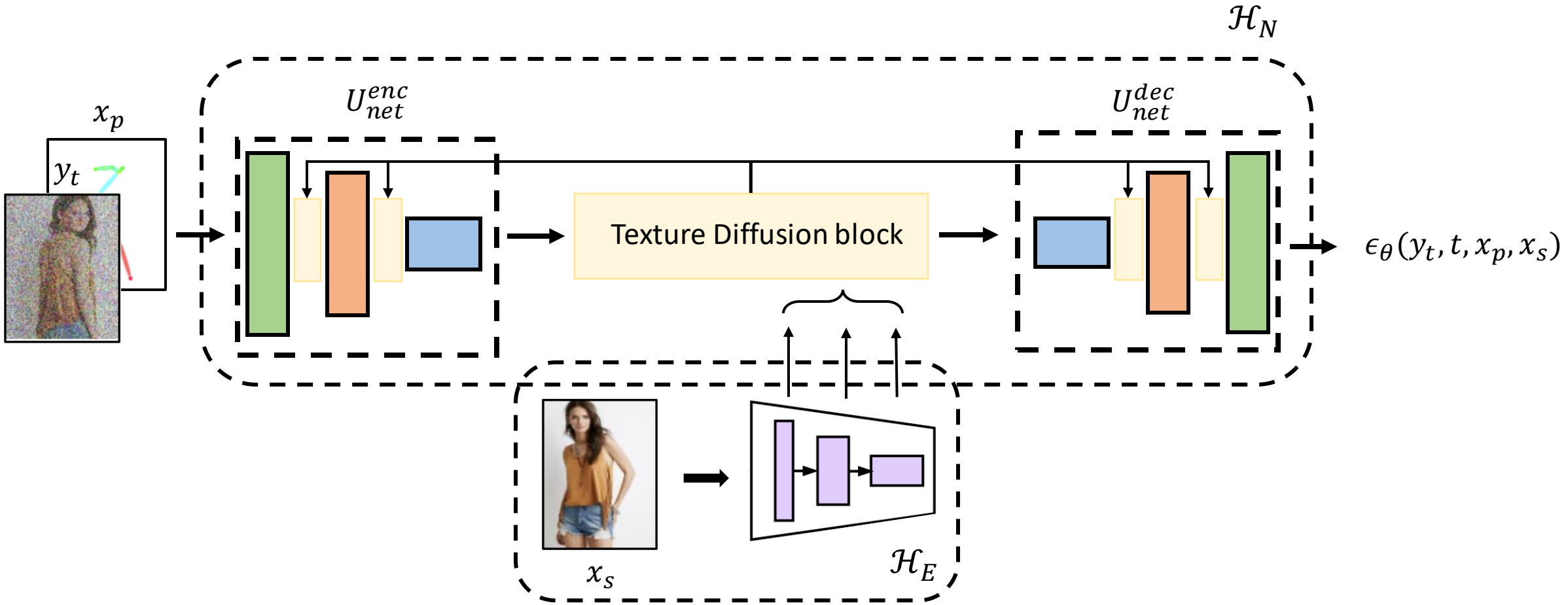
x_p



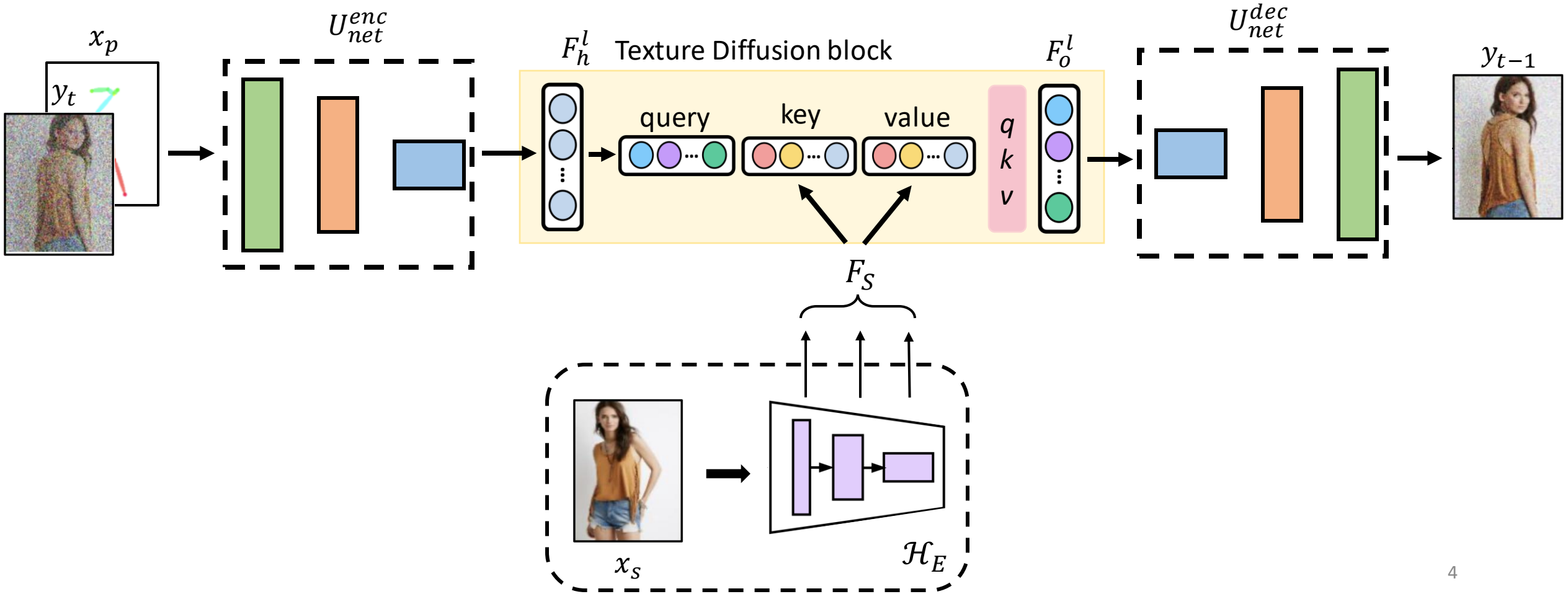


Output

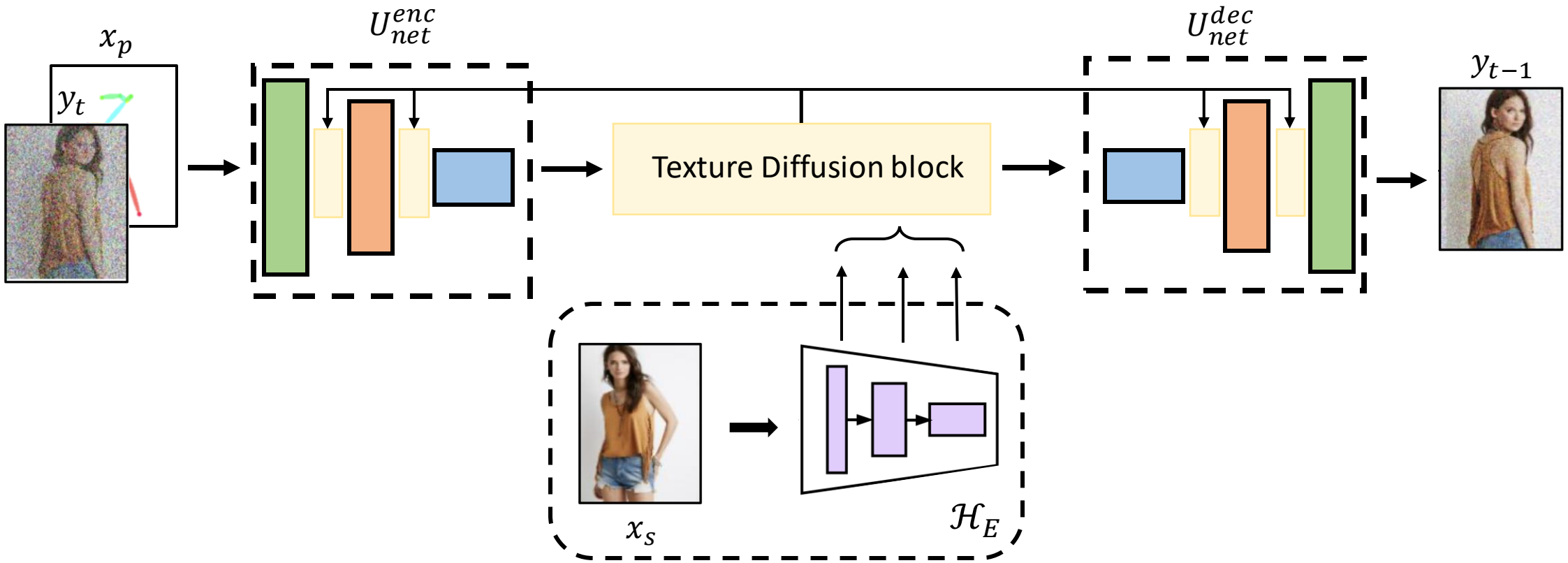
Person Image Synthesis - PIDM

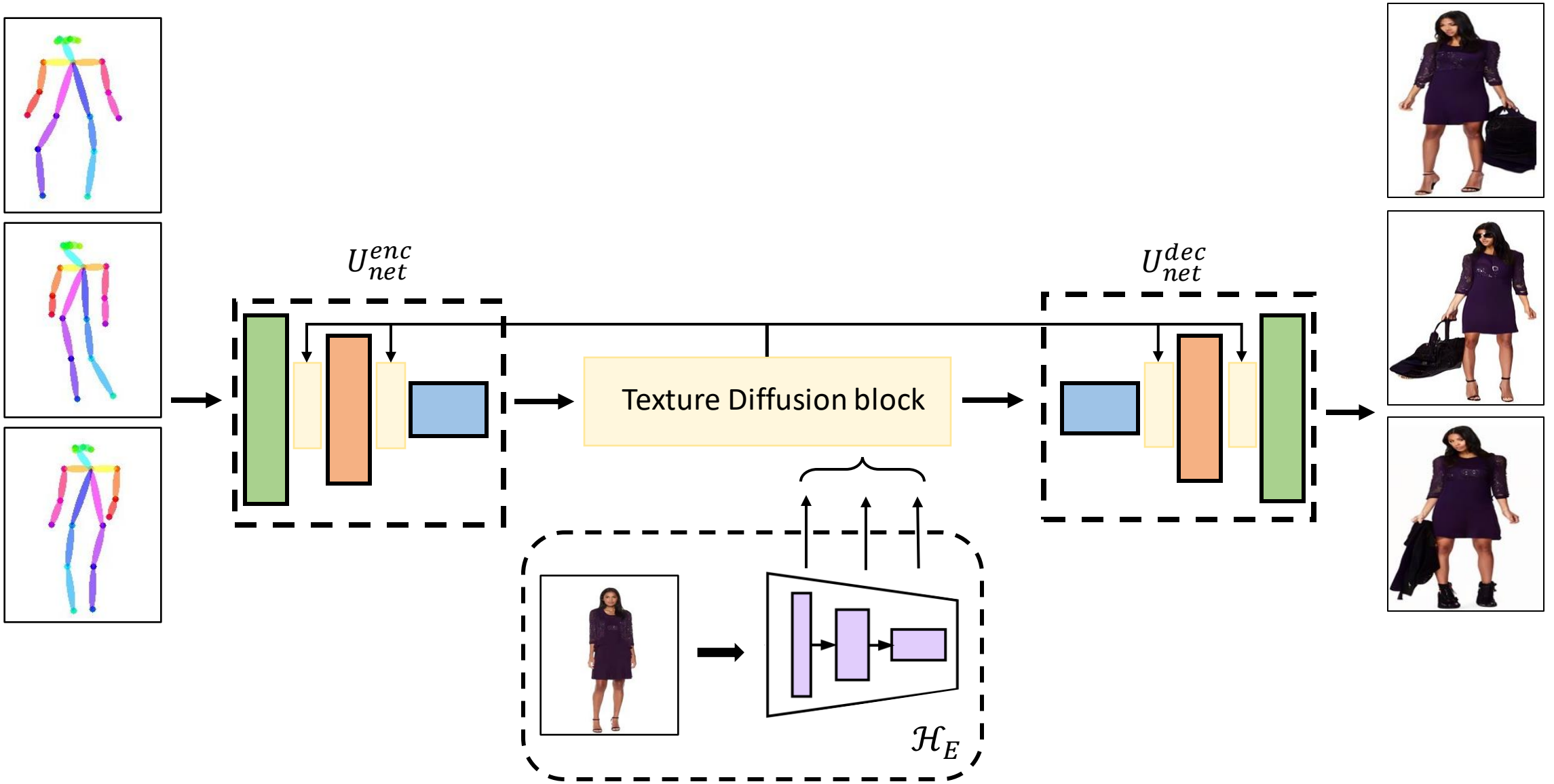


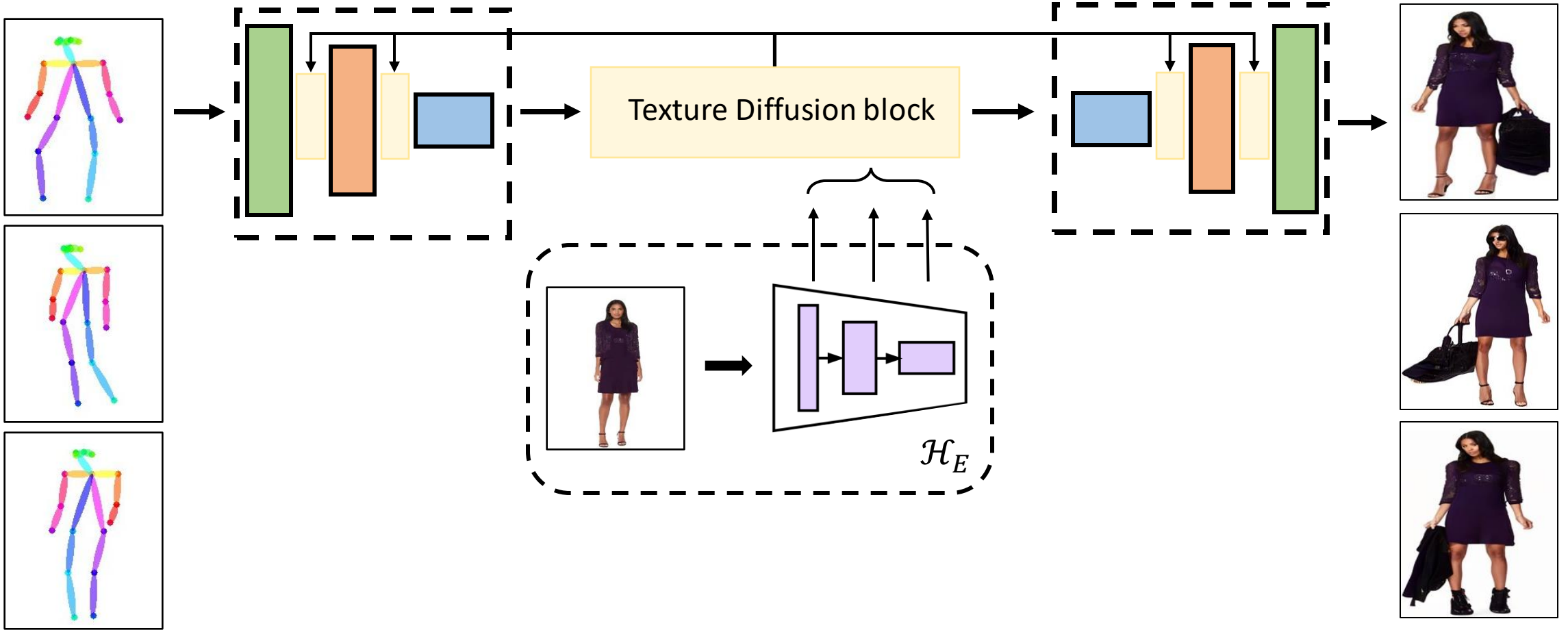
Texture Diffusion block

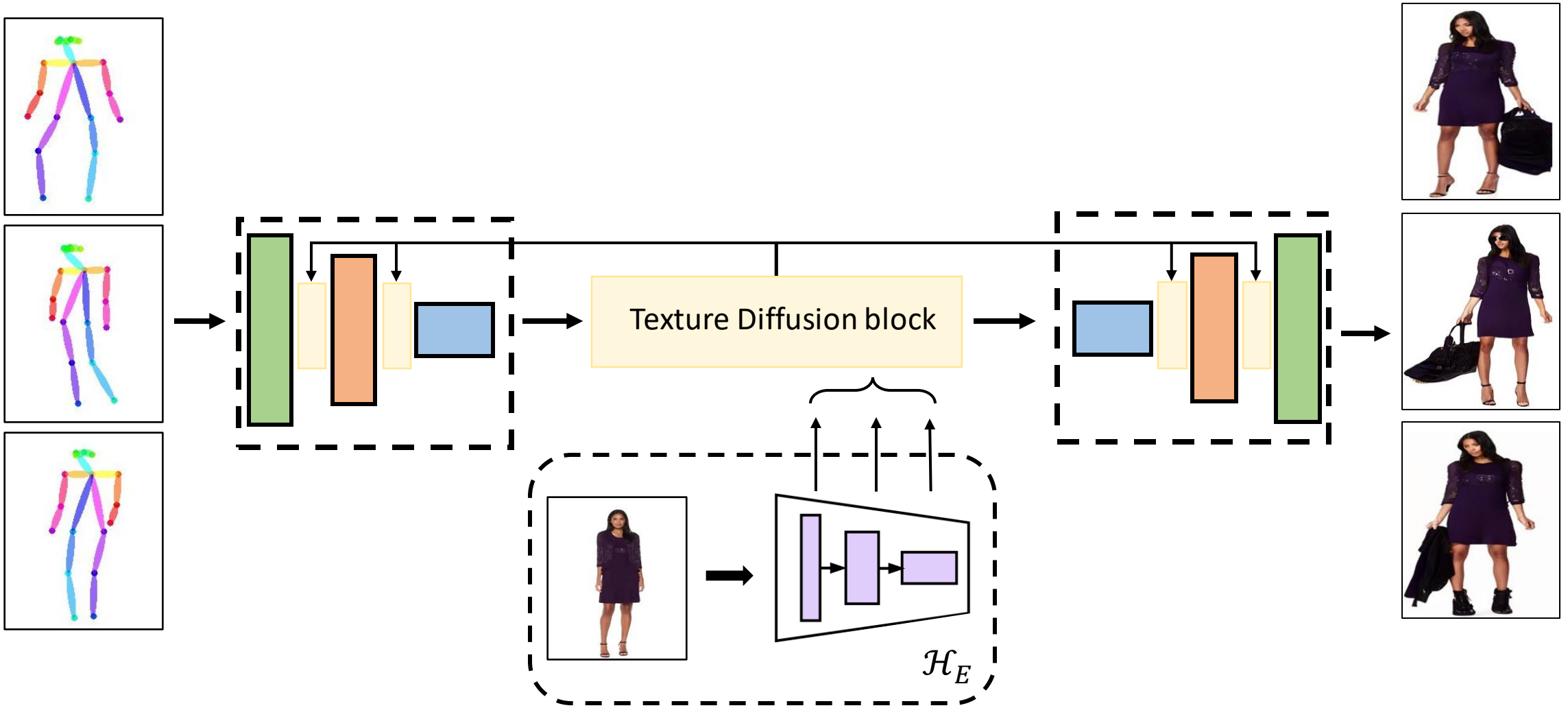


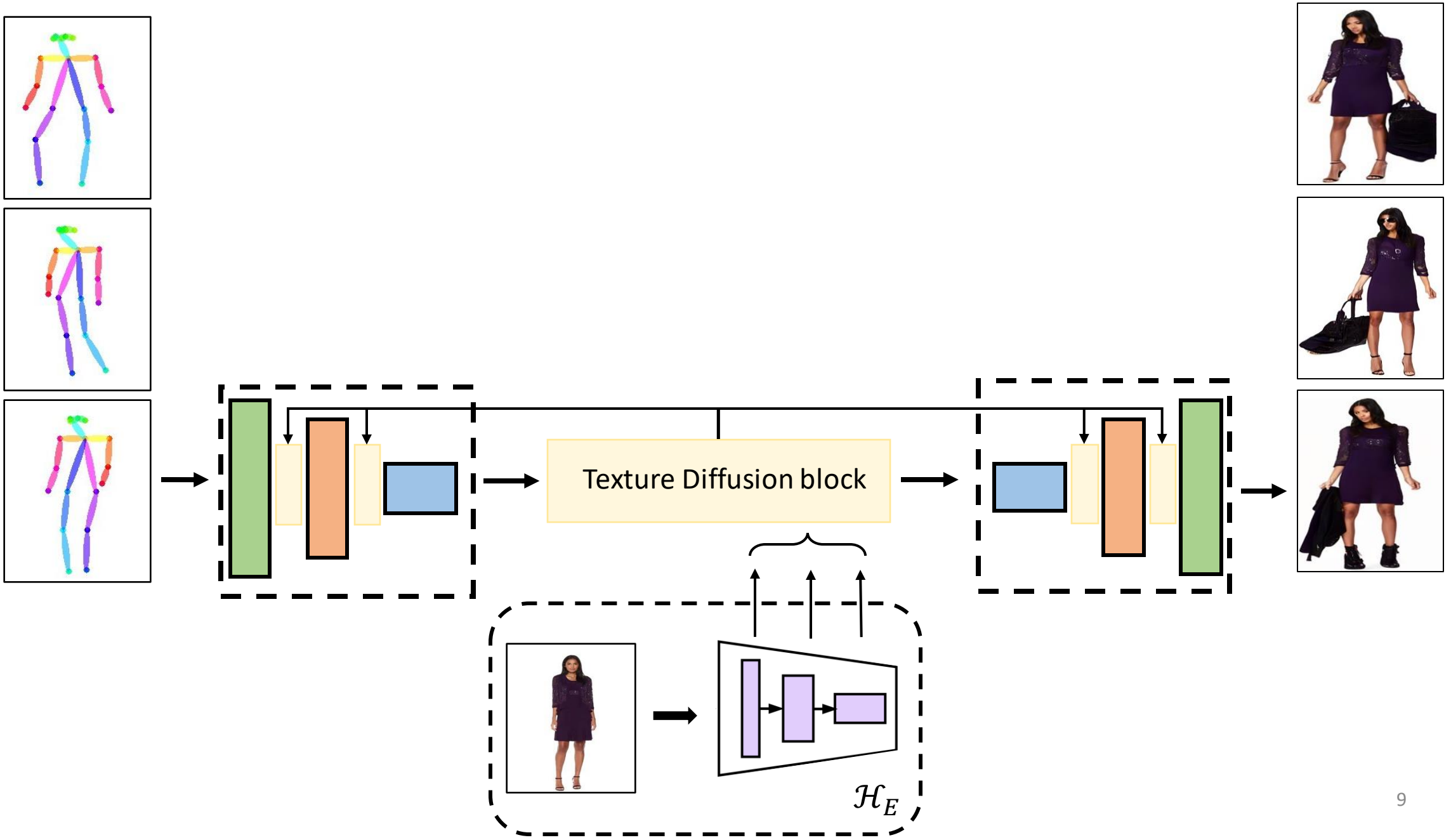
Texture Diffusion block




U_{net}^{enc}
U_{net}^{dec}
Texture Diffusion block
$\mathcal{H}_E$


Texture Diffusion block
$\mathcal{H}_E$

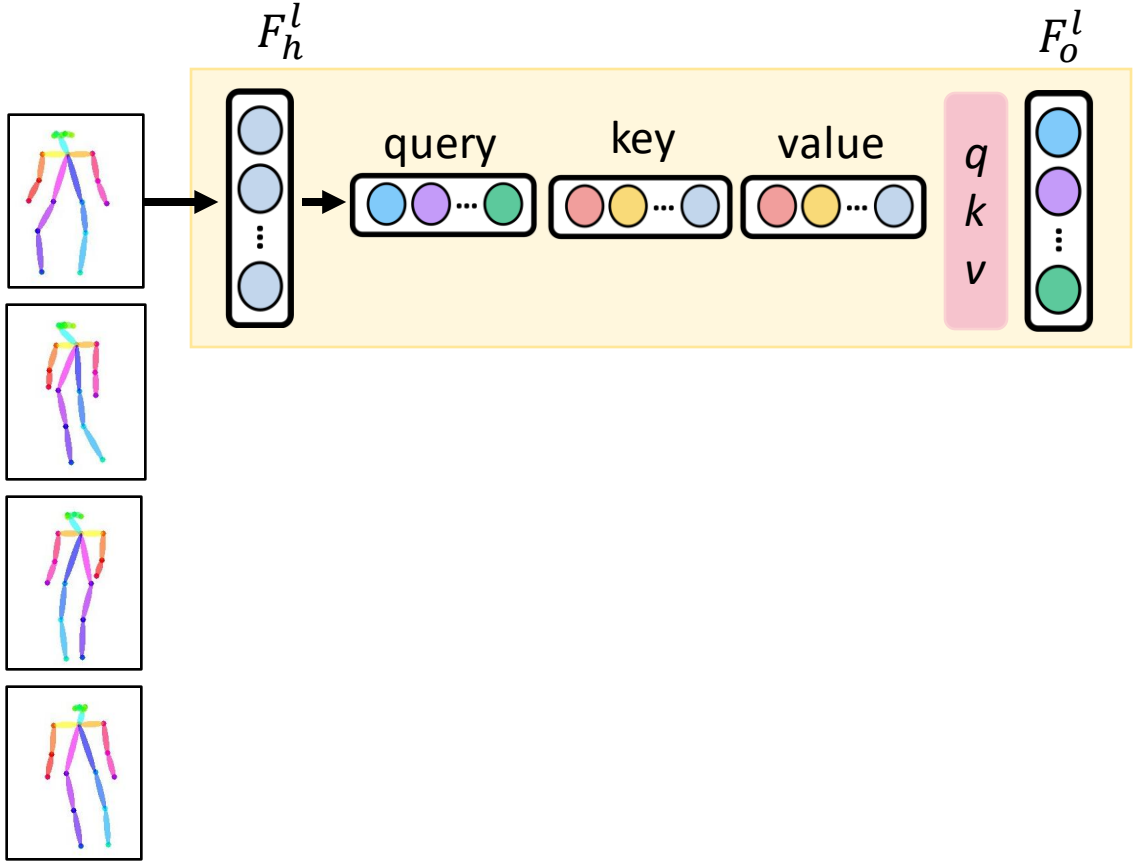

Texture Diffusion block
$\mathcal{H}_E$


Texture Diffusion block
$\mathcal{H}_E$

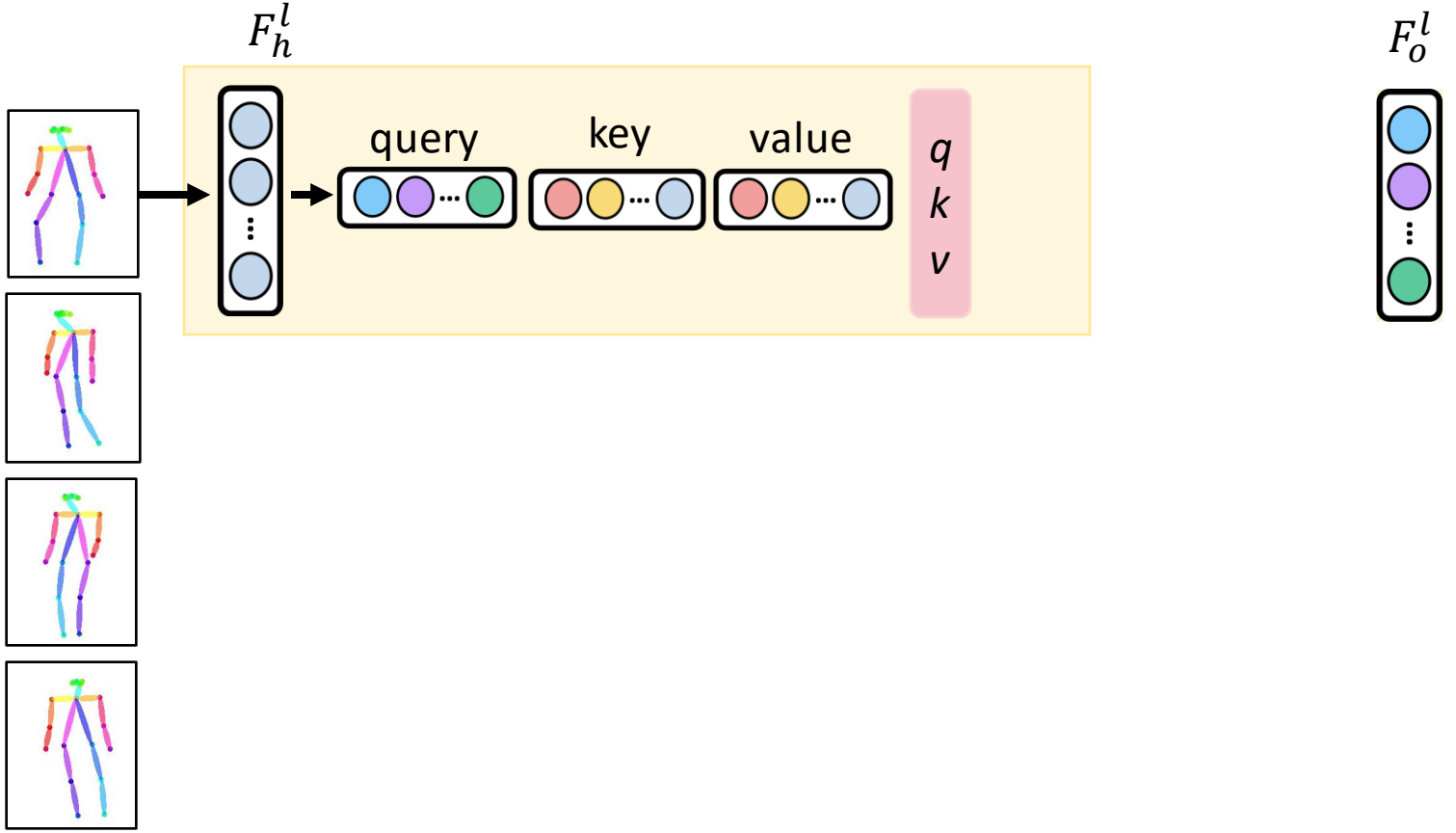
Results



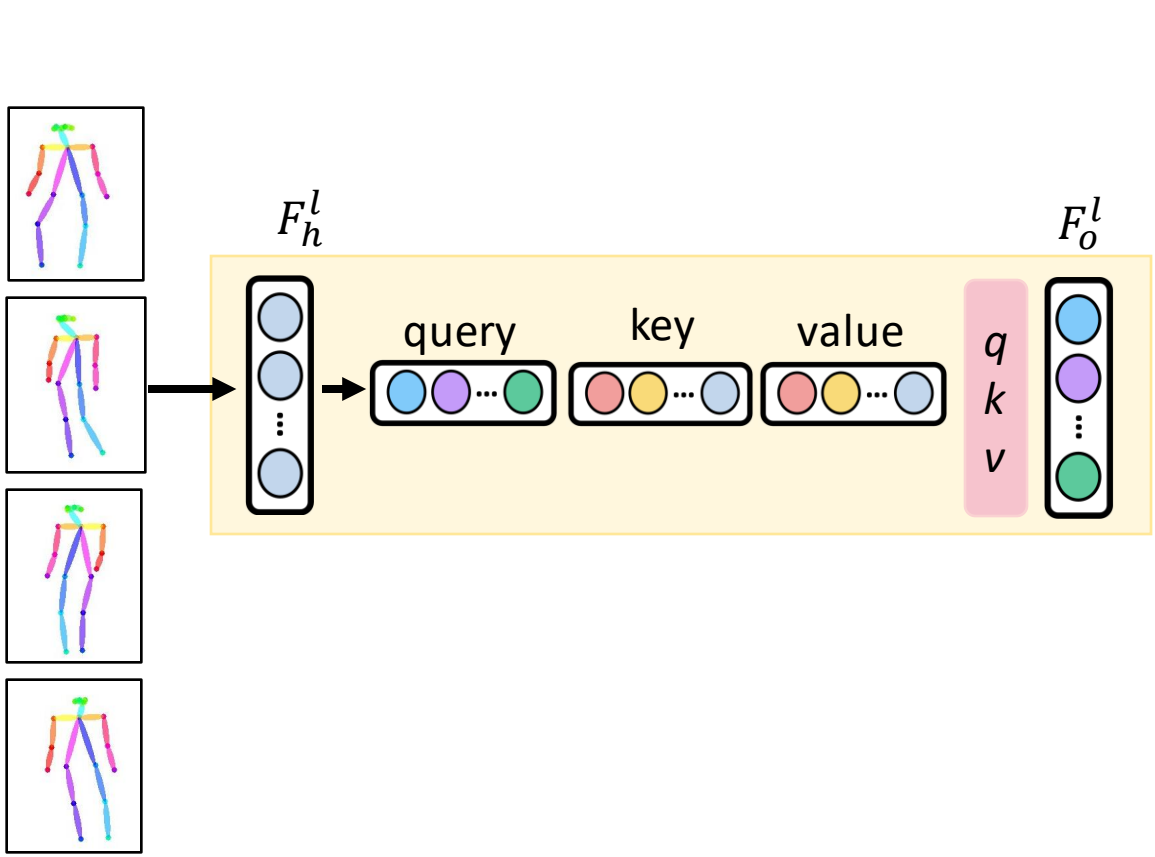

Texture Diffusion block

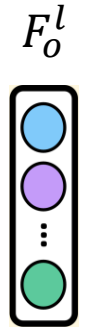


Texture Diffusion block

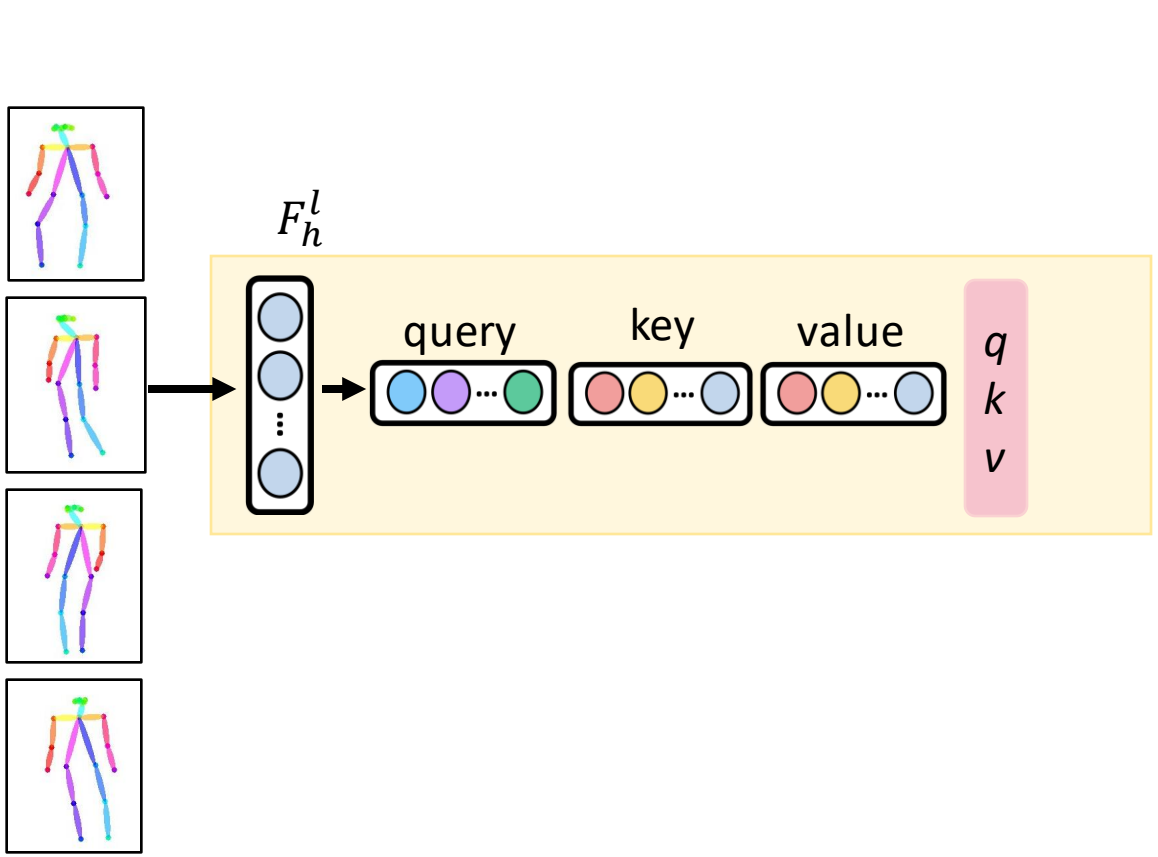


Texture Diffusion block

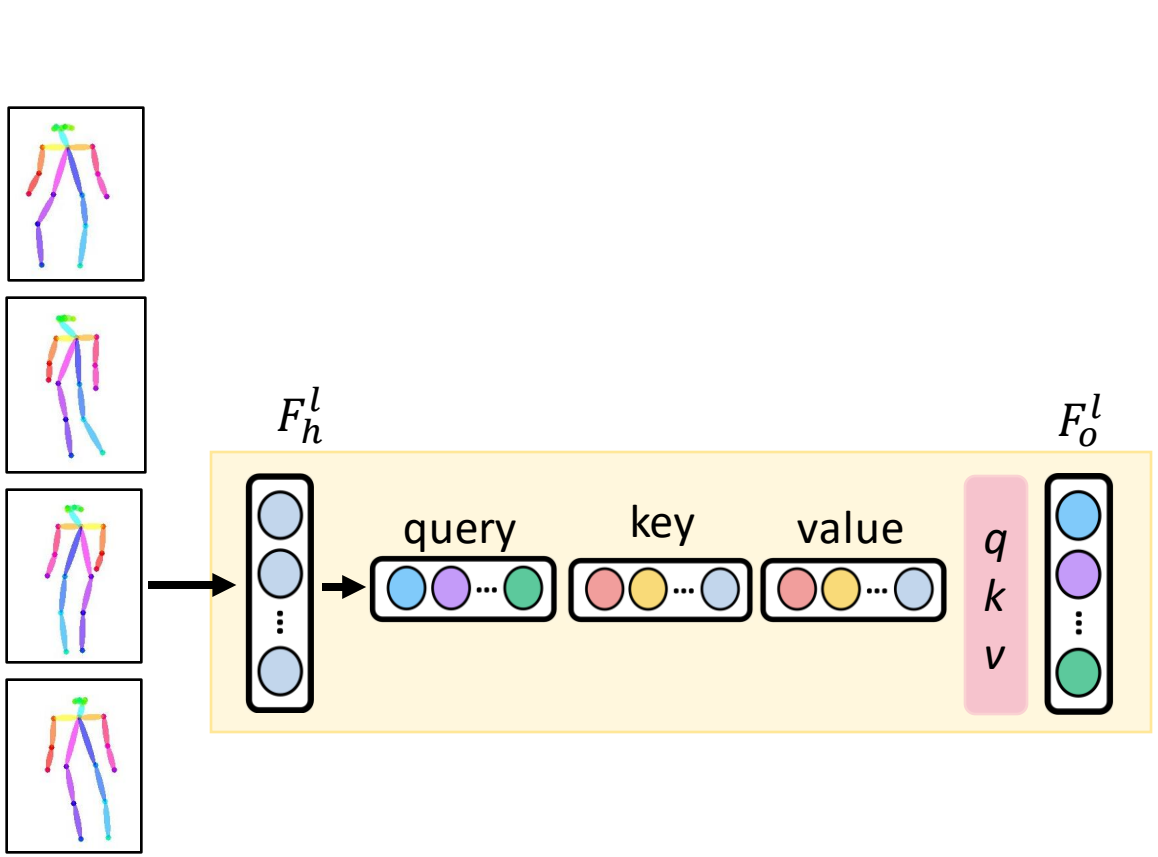




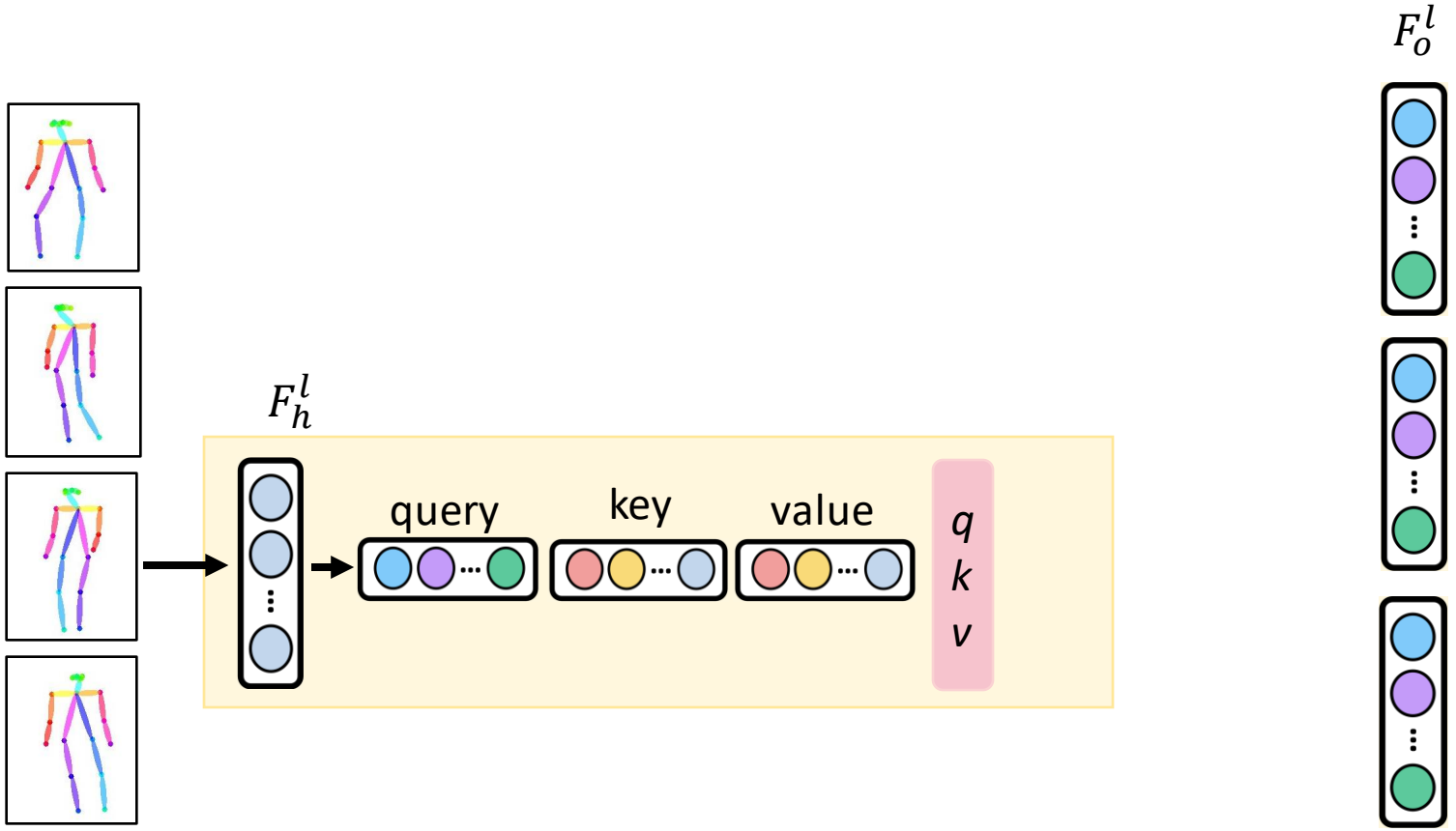
Texture Diffusion block



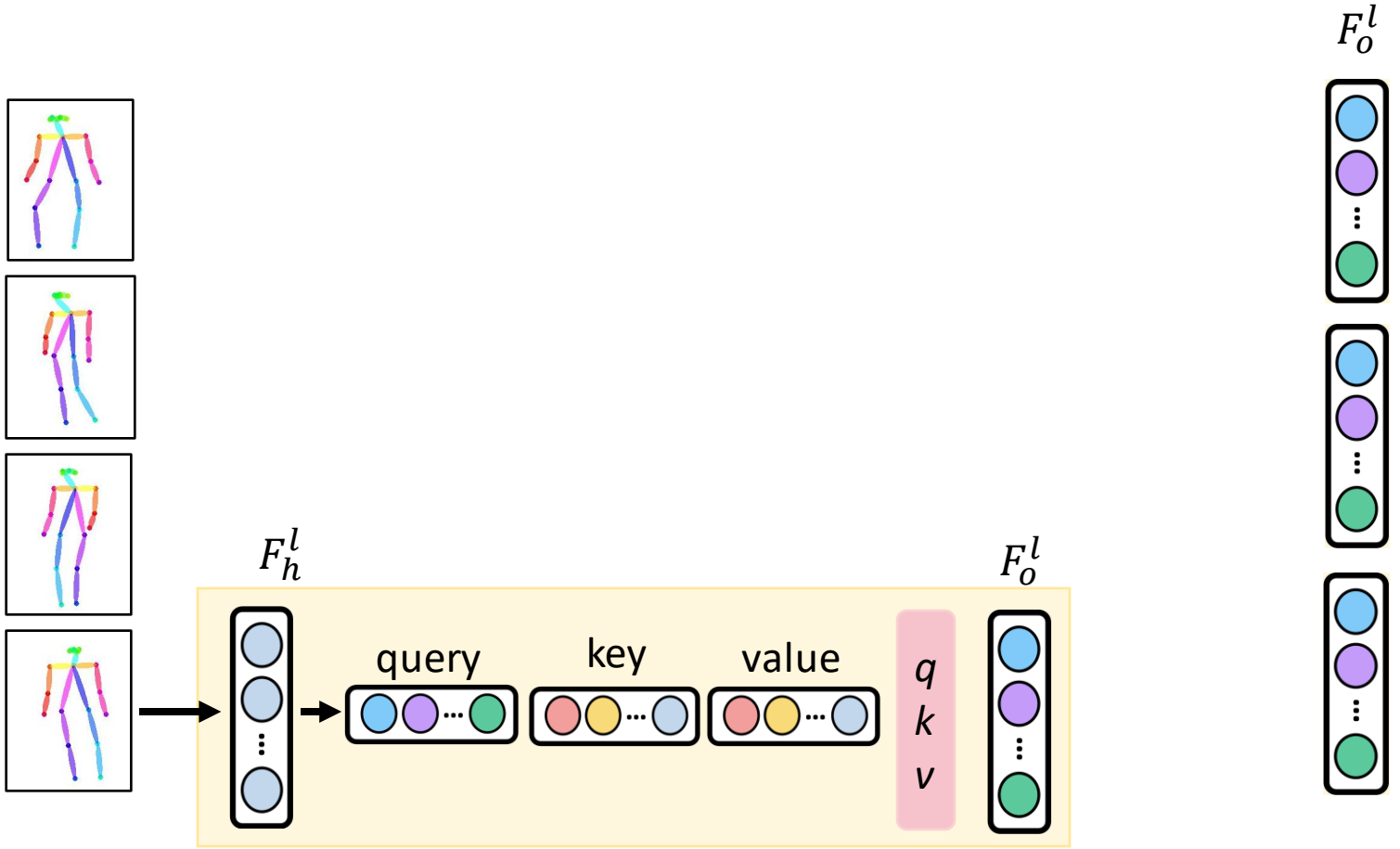
Texture Diffusion block



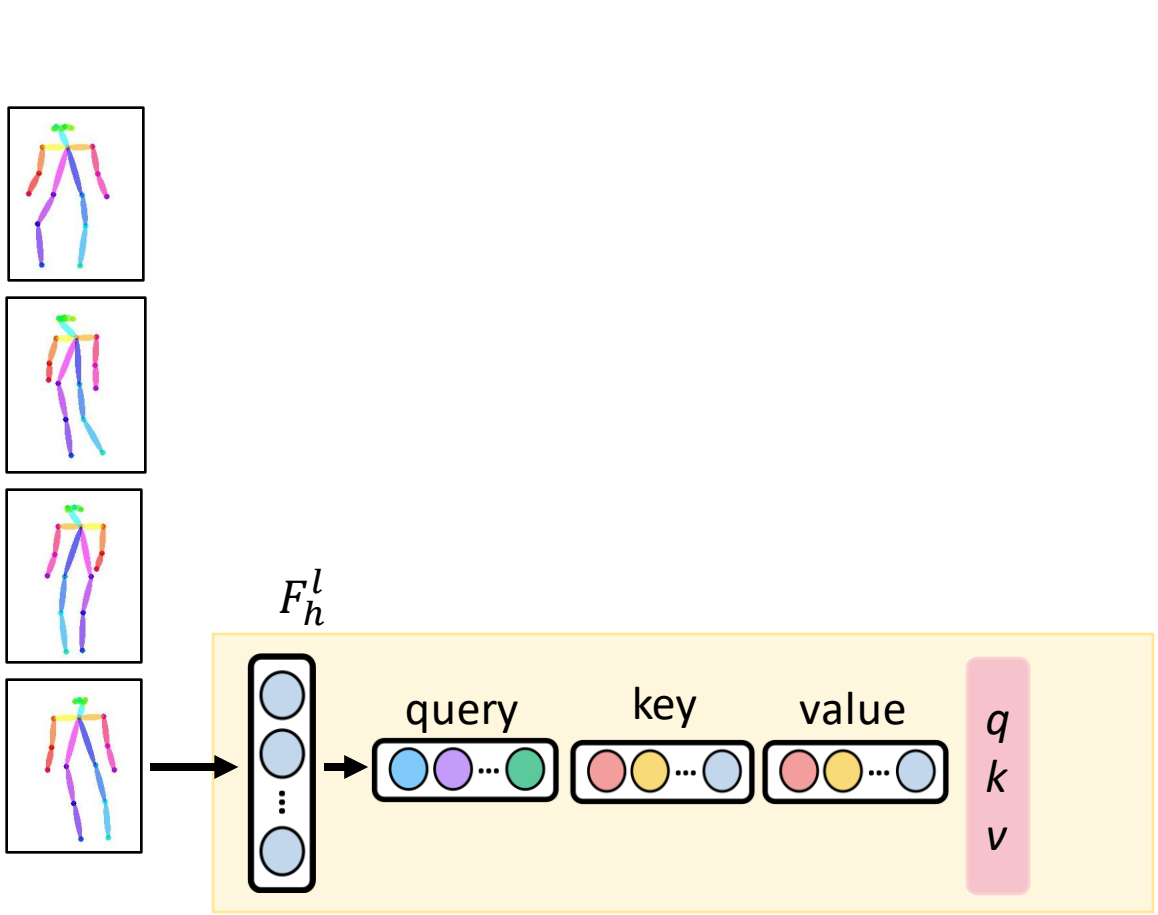
Texture Diffusion block

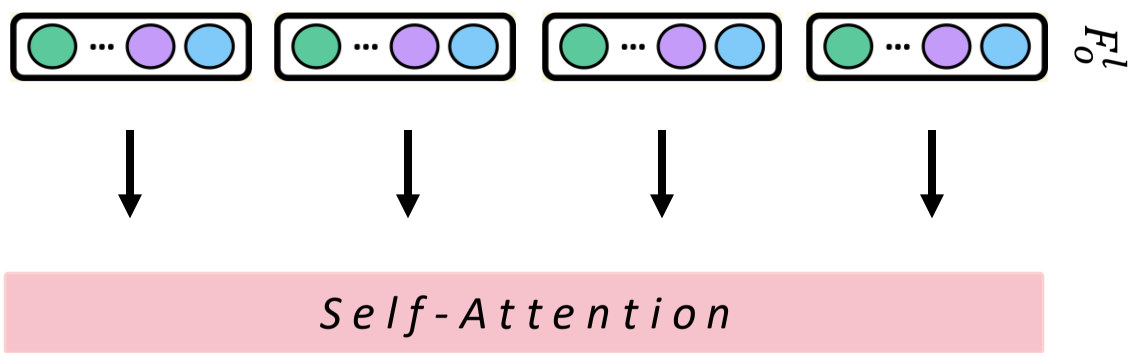


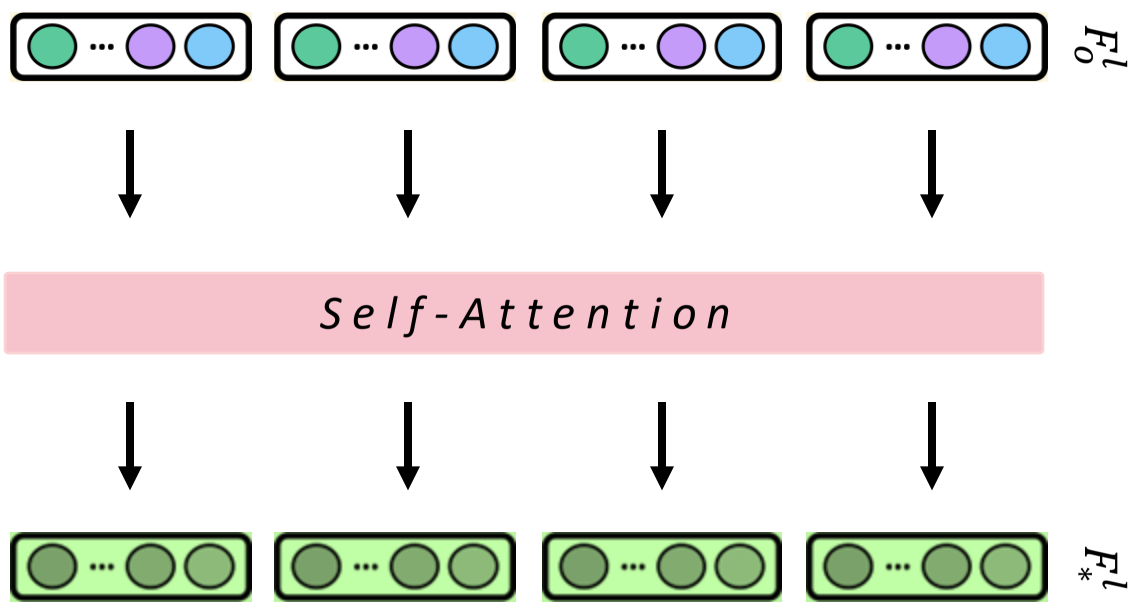
Texture Diffusion block



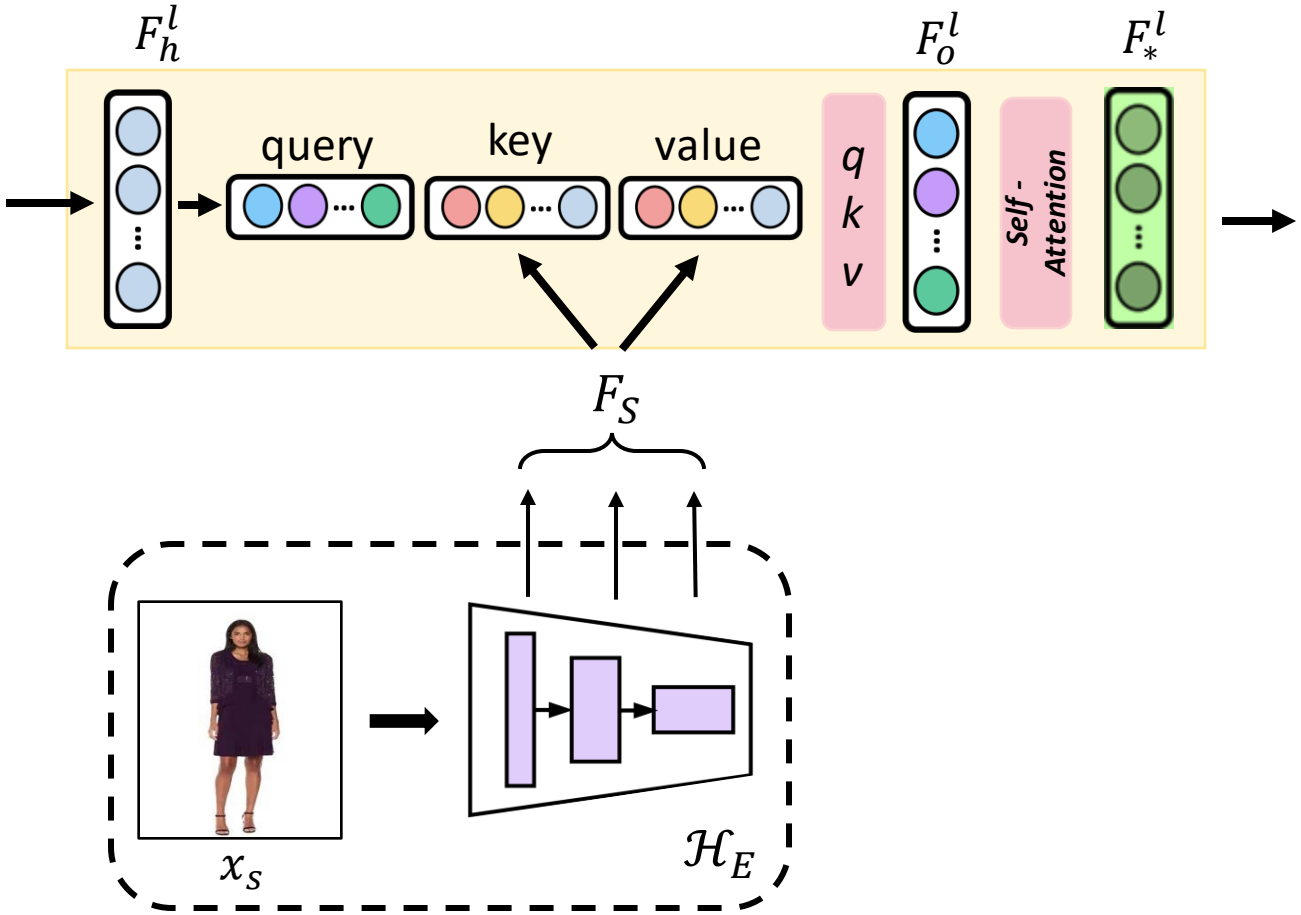
Texture Diffusion block

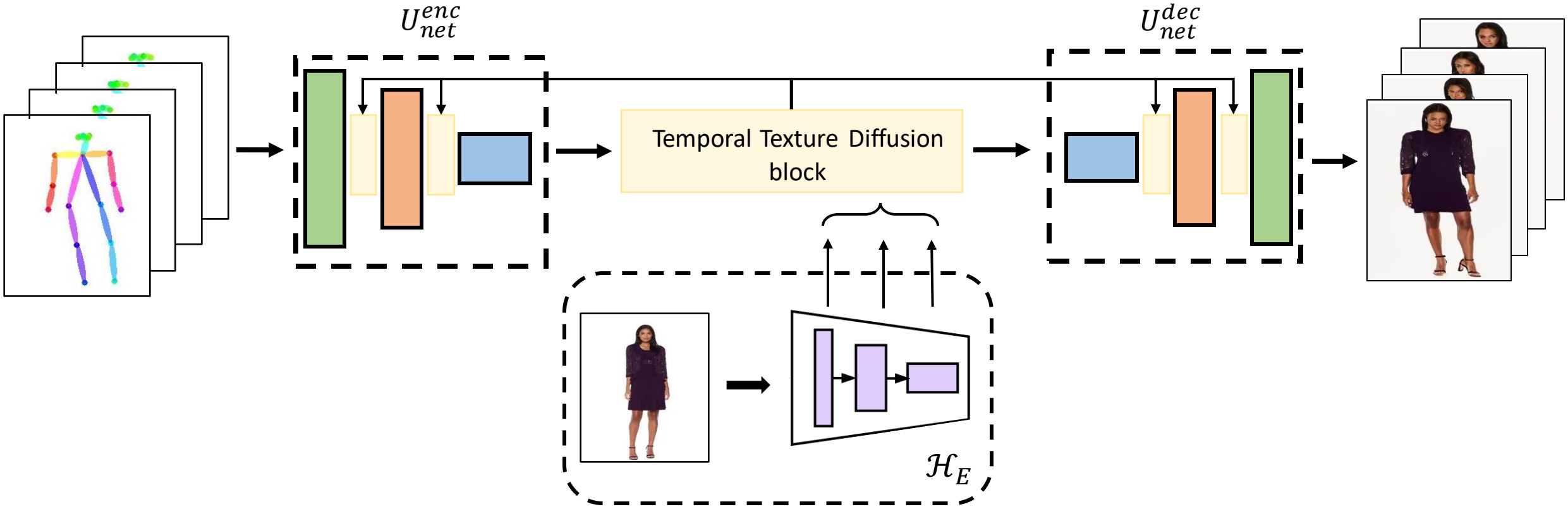



F_o^l
Self-Attention


F_o^l
Self-Attention
F_*^l

Temporal Texture Diffusion block




U_{net}^{enc}
U_{net}^{dec}
Temporal Texture Diffusion block
$\mathcal{H}_E$

Training

- 48 GB Ampere GPU
- Clip Length: 4
- Batch size 2

- Number of total parameters: 186141318 (186M)
- Number of trainable parameters: 5785600 (5.7M)
- Number of non-trainable parameters: 180355718 (180M)

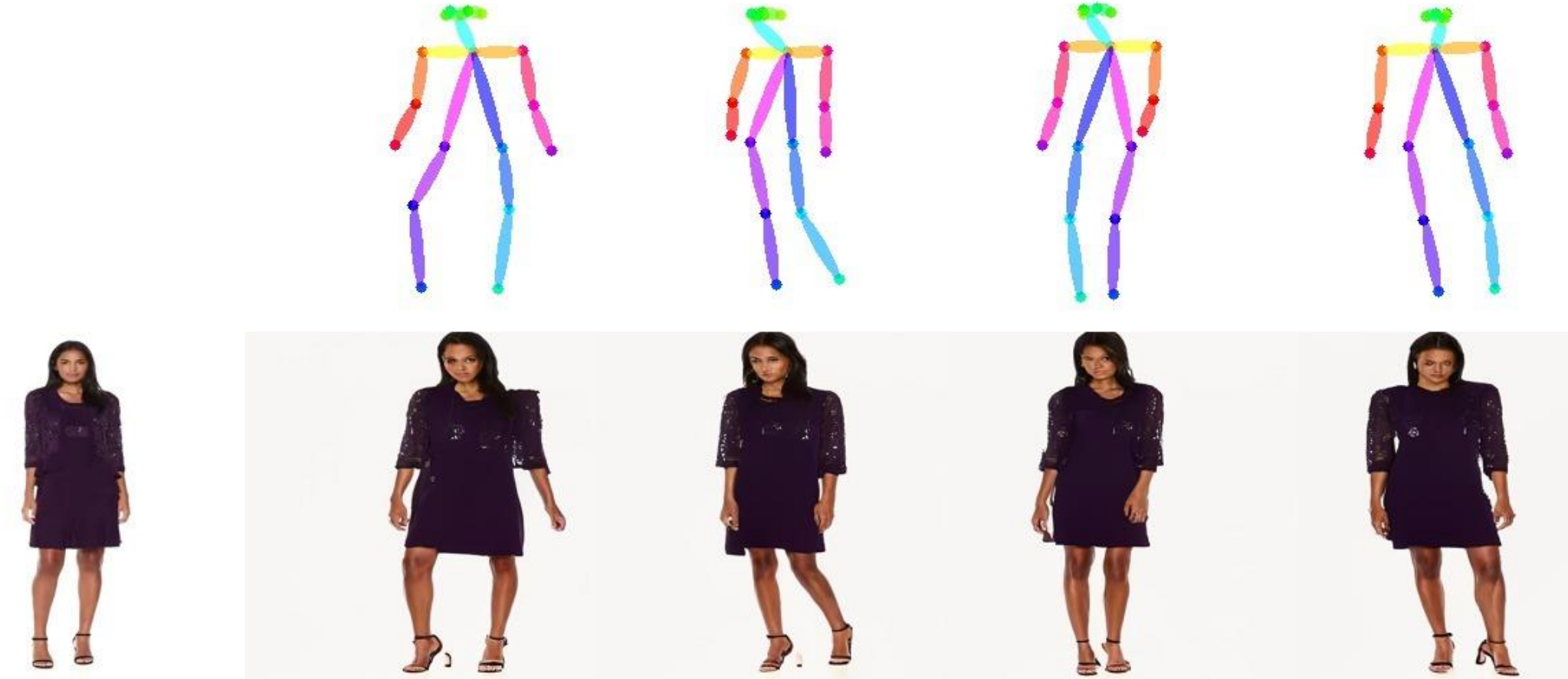
- Training epoch: 1000
- N Steps: 1000

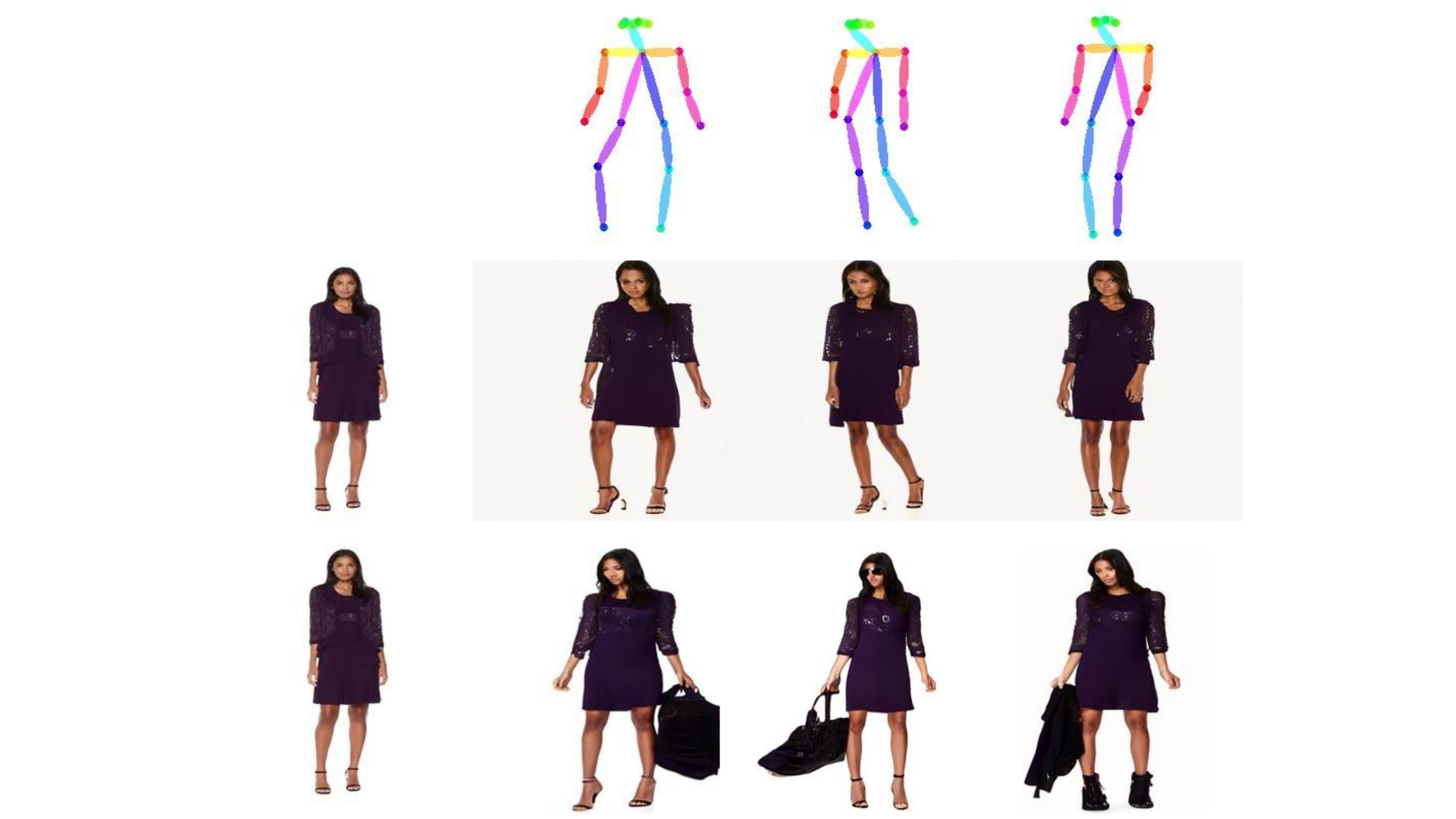

Epoch: 70
Epoch: 200
Epoch: 350
Epoch: 700

Experiments

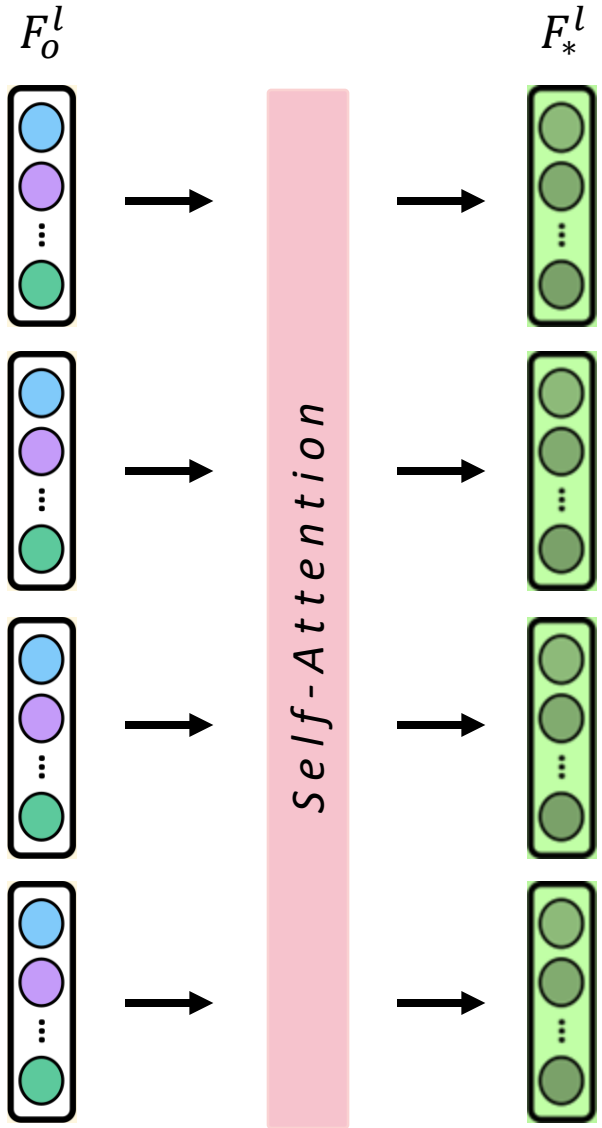
- Sampled with clip length 4
- Starting with frame 0 and step size 20
- Consecutive 4 frames
- Sampled with clip length 4 and looped over the entire video length

Results





Temporal Texture Diffusion Block



Temporal Texture Diffusion Block