清華大學
Tsinghua University

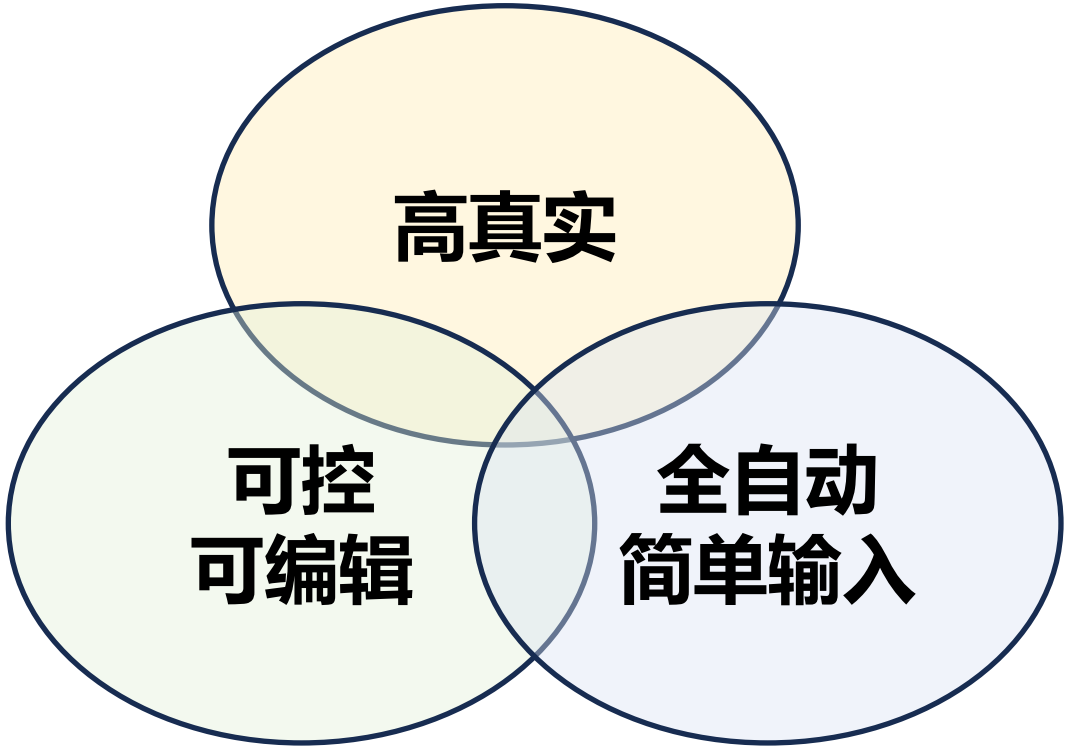
动作与视点可控人体视频生成

刘烨斌

4/15/2025

数字人目标与疑问

如何实现高真实感、又交互友好可控，且自动生成的数字人？



2D？3D

基于视频数据驱动的3D数字化身

数据驱动的人体数字化身生成（2011’）

基本动作库

首个基于多视点视频数据库的人体Avatar生成， F. Xu, et al., Video-based Characters - Creating New Human Performances from a Multi-view Video Database, in SIGGRAPH 2011

端到端数字化身生成

- 给定人体骨架参数μ和表情参数ß，生成人物身体及脸部图象

NeRF视点生成：

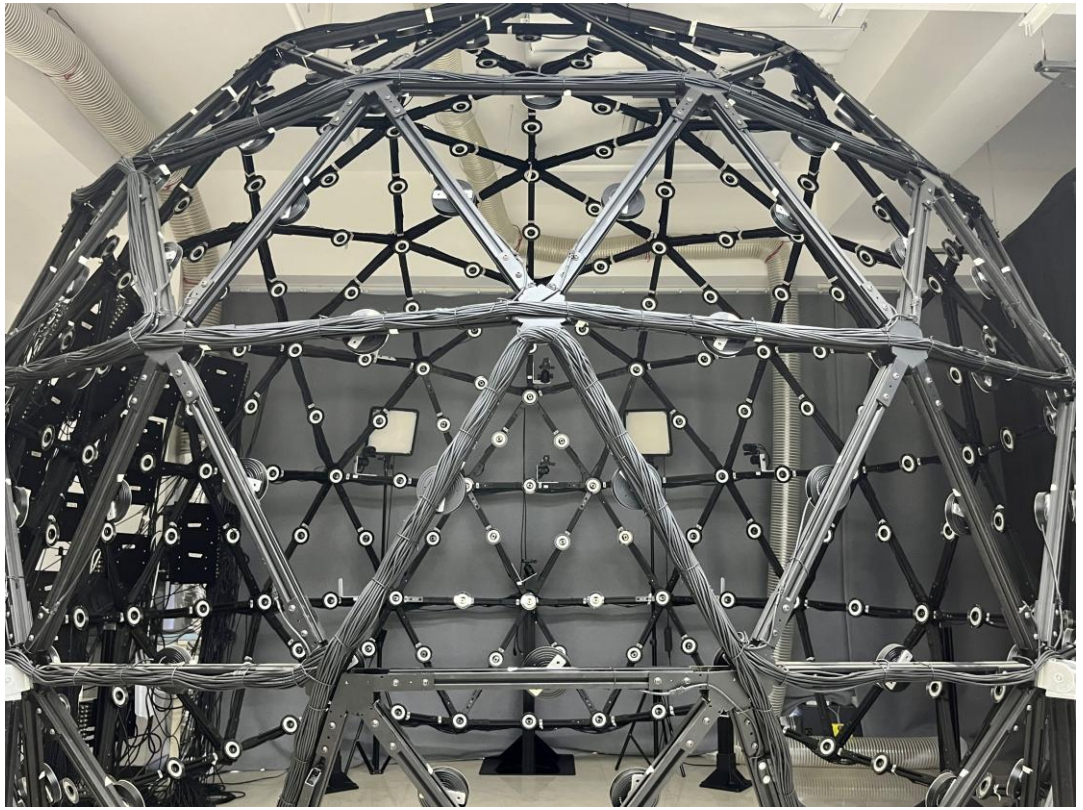
$$(x,y,z,\theta,\phi) \rightarrow F_\Theta \rightarrow (RGB\sigma)$$

数字人新动作新表情生成

$$(x,y,z,\theta,\phi,ß,\mu) \rightarrow F_\Theta \rightarrow (RGB\sigma)$$

3D数字化身捕捉设备

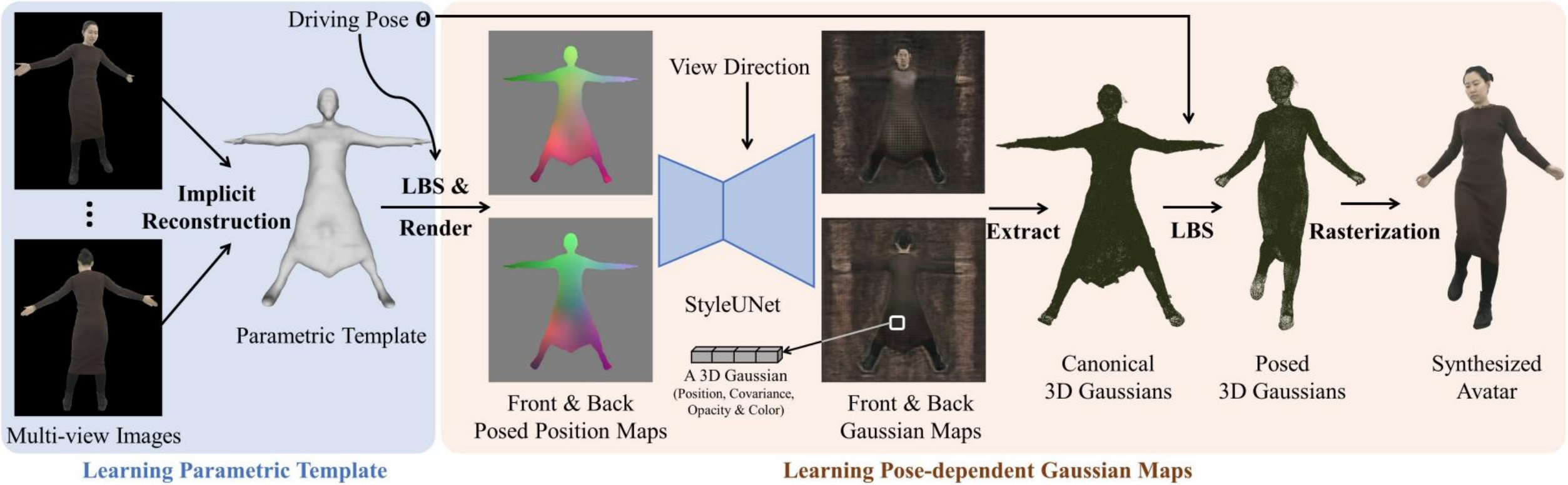
多视点视频相机阵列



变光照多视点视频相机阵列

3D高斯数字化身

3D高斯表征的全息数字化身生成方法



Li et al. Animatable Gaussians: Learning Pose-dependent Gaussian Maps for High-fidelity Human Avatar Modeling, **CVPR 2024**

3D高斯数字化身

Li et al. Animatable Gaussians: Learning Pose-dependent Gaussian Maps for High-fidelity Human Avatar Modeling, **CVPR 2024**

3D高斯数字化身

3D高斯数字化身

8 Views Input

Xu et al. Gaussian Head Avatar: Ultra High-fidelity Head Avatar via Dynamic Gaussians, **CVPR 2024**

3D高斯数字化身

3D高斯数字化身

带头发建模的3D高斯数字化身

驱动

生成

发丝模型

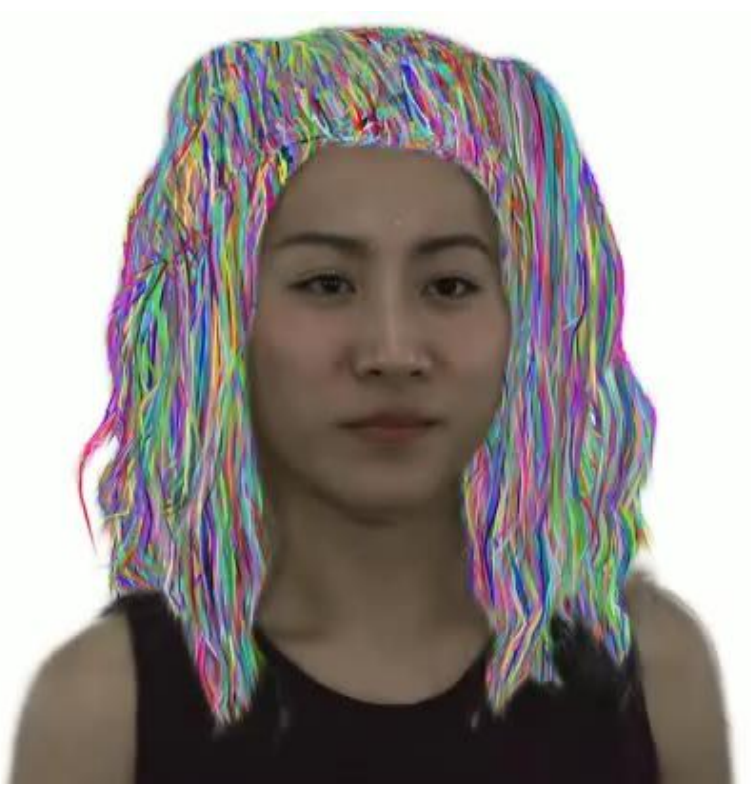
Liao et al. , HHAvatar: Gaussian Head Avatar with Dynamic Hairs， Arxiv 2024

3D高斯数字化身

带头发建模的3D高斯数字化身

驱动

生成



发丝模型

Liao et al. , HHAvatar: Gaussian Head Avatar with Dynamic Hairs， Arxiv 2024

3D高斯数字化身

3D高斯数字化身

3D高斯数字化身

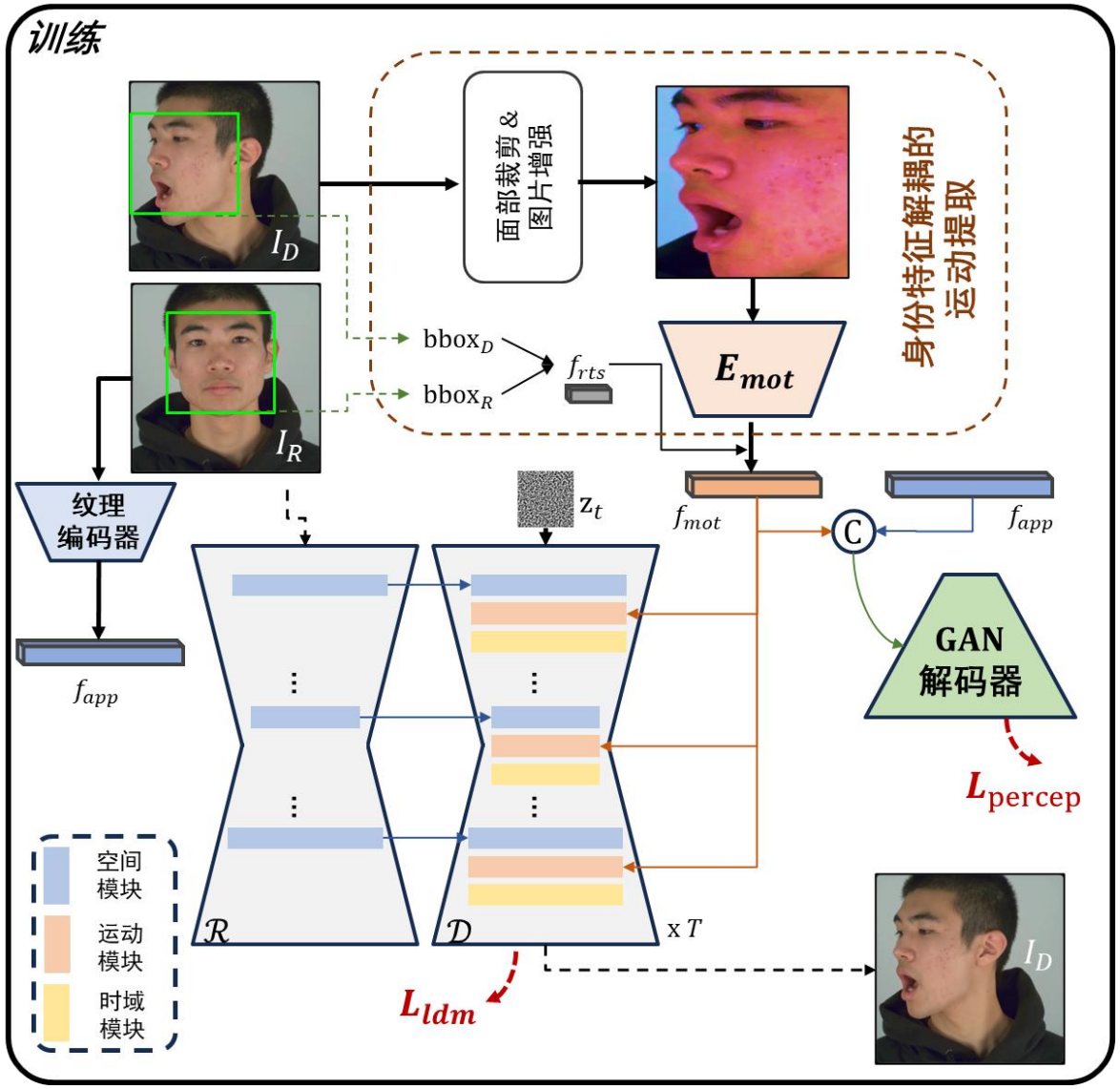
- 优点
  - 具有显式3D表征，三维一致性高
  - 可控性可编辑性强
  - 生成与渲染速度快
- 缺点
  - 依赖多视点视频采集，单视点或单图像效果较差
  - 复杂表情驱动下人脸效果差
  - 复杂动作驱动下人体效果差（数据先验少，穿模以及纹理模糊）

复杂人脸表情视频生成

Zhao et al. X-NeMo: Expressive Neural Motion Reenactment via Disentangled Latent Attention[C]. ICLR, 2025.

复杂人脸表情视频生成

- **运动表征：可优化的隐式运动表征**
  - 直接从视觉信号捕捉细粒度运动信息
- **控制方式：基于注意力机制控制**
  - 实现语义级别的表情迁移
- **解耦表征学习**
  - 图像信号的增强，破坏驱动图片与目标图片的外观一致性，获得精准运动
  - 双解码器（扩散模型+GAN）：GAN一步解码生成图像，便于使用感知损失函数引导编码器聚焦于面部局部区域，稳定注意力模块对于隐式运动分布的建模



Zhao et al. X-NeMo: Expressive Neural Motion Reenactment via Disentangled Latent Attention[C]. ICLR, 2025.

复杂人脸表情视频生成

Zhao et al. X-NeMo: Expressive Neural Motion Reenactment via Disentangled Latent Attention[C]. ICLR, 2025.

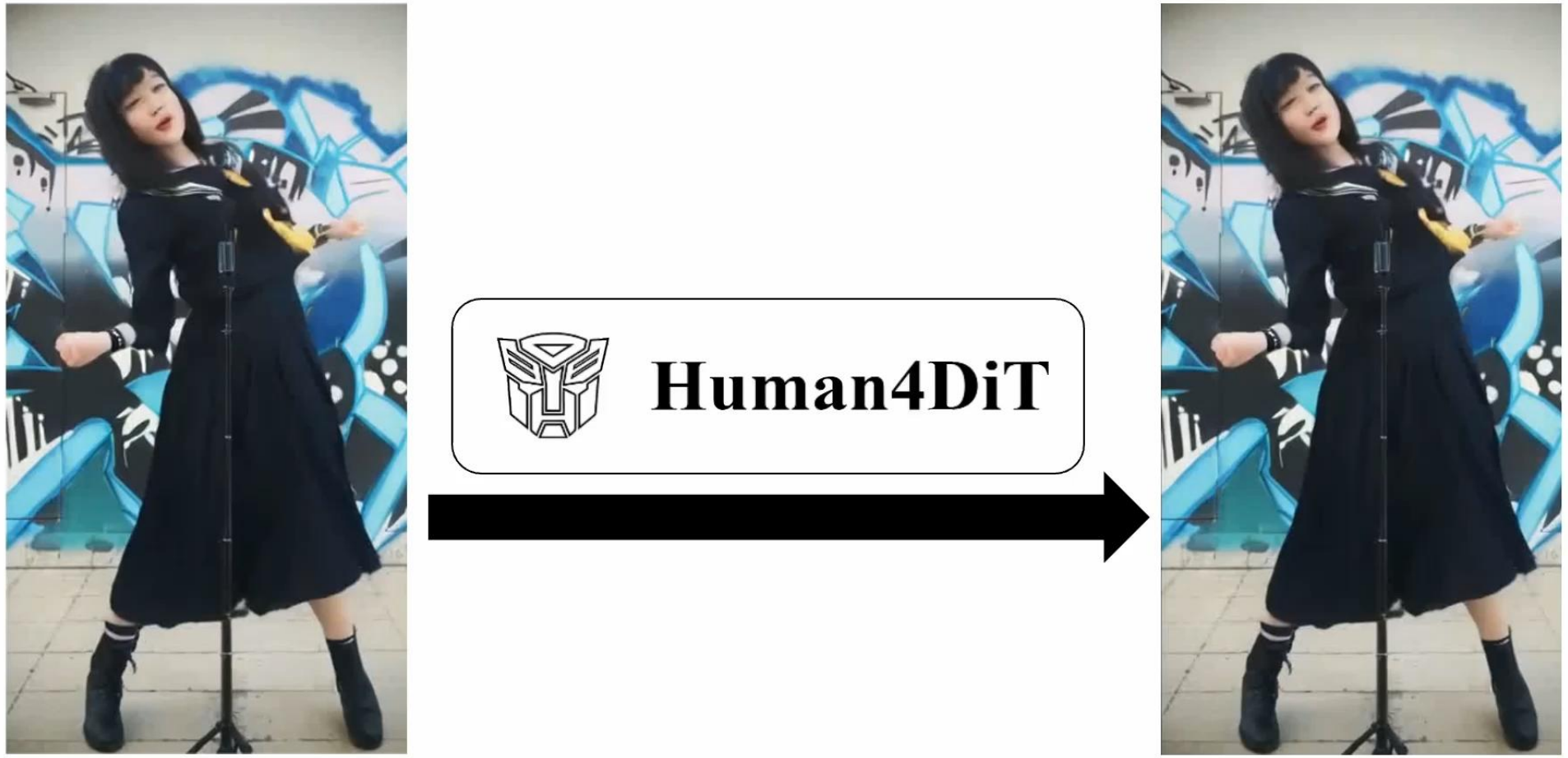
复杂人脸表情视频生成

- 使用学到的1D隐式运动表征，通过低维运动生成实现高维视频生成
- 在带情绪标签人脸视频数据集（MEAD）上训练的情感条件视频生成效果

参考图片

Zhao et al. X-NeMo: Expressive Neural Motion Reenactment via Disentangled Latent Attention[C]. ICLR, 2025.

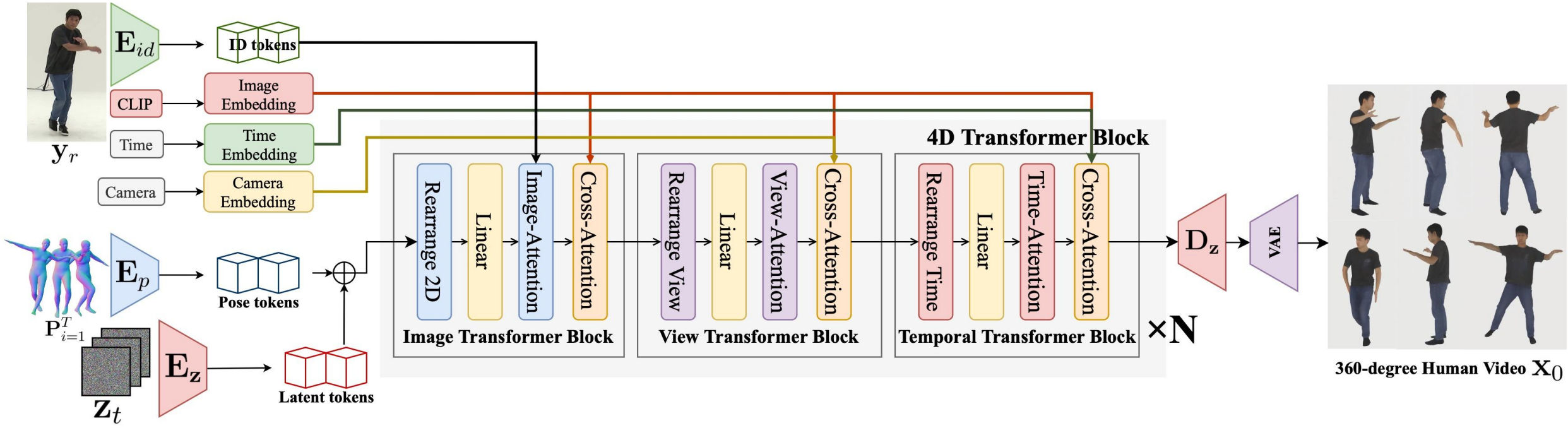
动作与视点可控人体视频生成



Shao et al. , Human4DiT: Free-view Human Video Generation with 4D Diffusion Transformer, **SIGGRAPH Asia 2024, Journal Track**

动作与视点可控人体视频生成

- 将3D人体形态模版作为控制条件，以DiT框架为基础，设计了4D时空扩散变换器，学习角度可控视频生成



Shao et al. , Human4DiT: Free-view Human Video Generation with 4D Diffusion Transformer,
SIGGRAPH Asia 2024，Journal Track

动作与视点可控人体视频生成

- 将3D人体形态模版作为控制条件，以DiT框架为基础，设计了4D时空扩散变换器，学习角度可控视频生成

动作与视点可控人体视频生成

Reference Image　　Free-view Video (motion 1)　　Free-view Video (motion 2)

Shao et al. , Human4DiT: Free-view Human Video Generation with 4D Diffusion Transformer,
SIGGRAPH Asia 2024，Journal Track

动作与视点可控人体视频生成

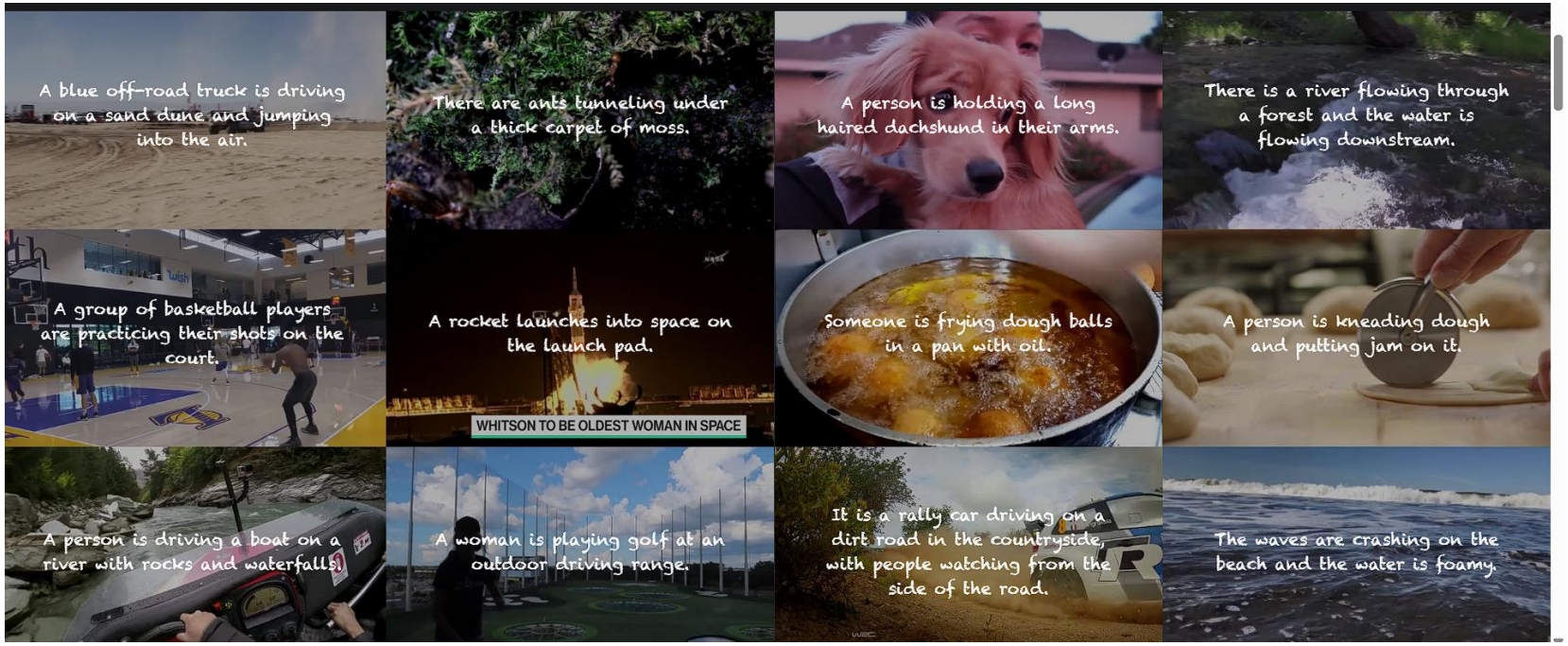
三维动态人体的自由视点视频生成

多图输入

动作与视点可控人体视频生成

- 从开源数据集中构建起一套关于人体运动的数据处理管线，**筛选并标注**了近**两百万**的人体视频数据，包括**多人场景、多视角、camera运动等**

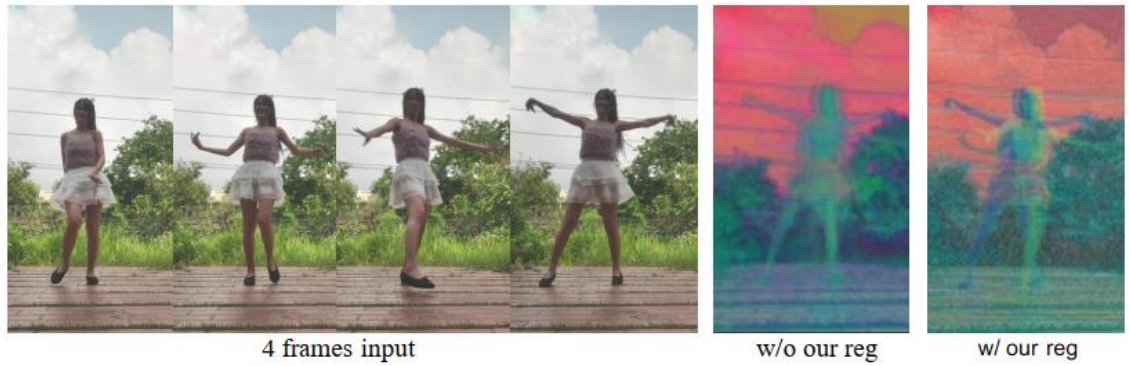
Word cloud of short captions.

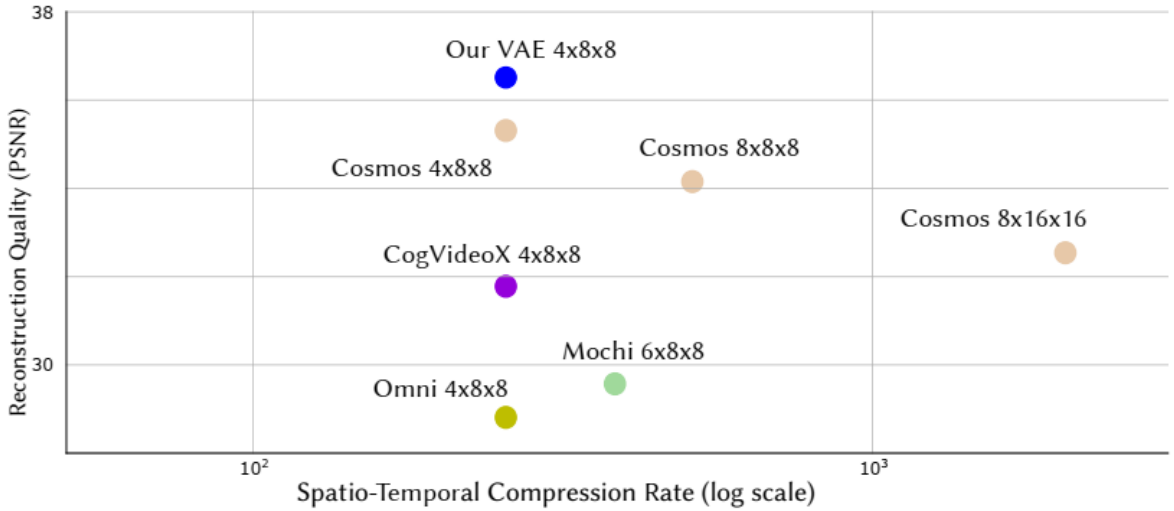


动作与视点可控人体视频生成

- 针对人体的快速动作视频专门训练了Video VAE，**性能领先**于当前所有开源的VAE模型

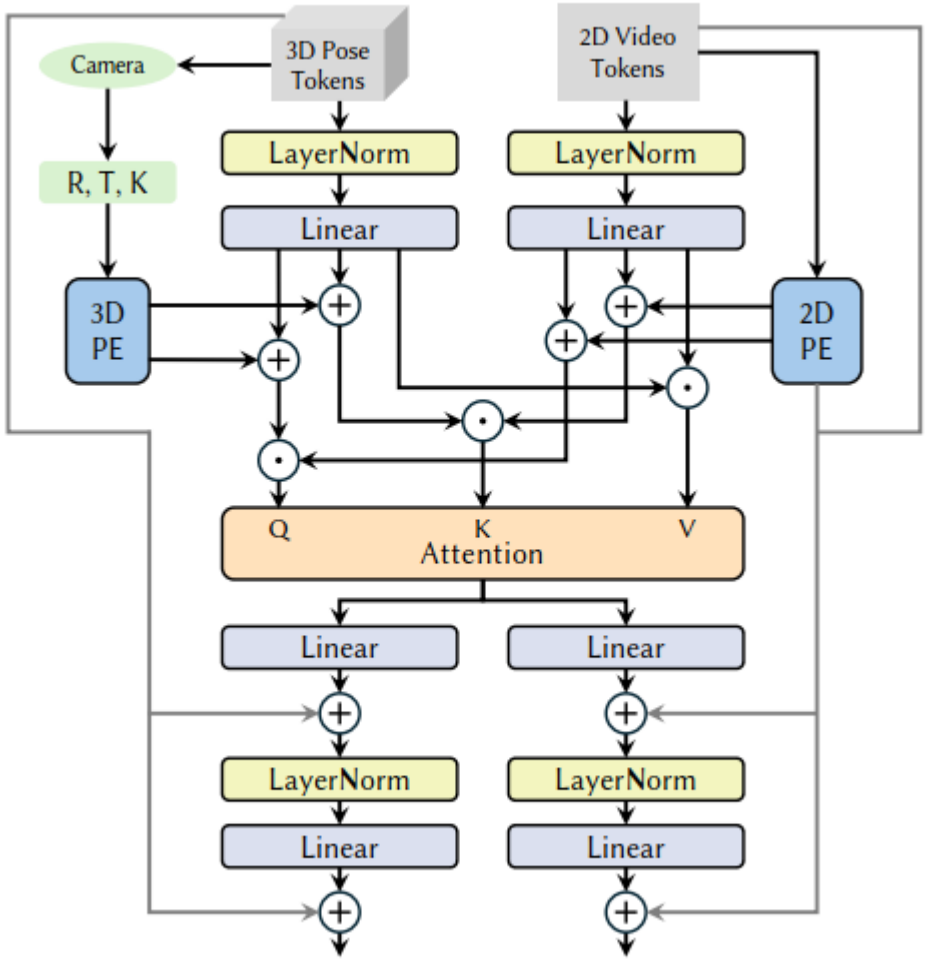
VAE编码效果





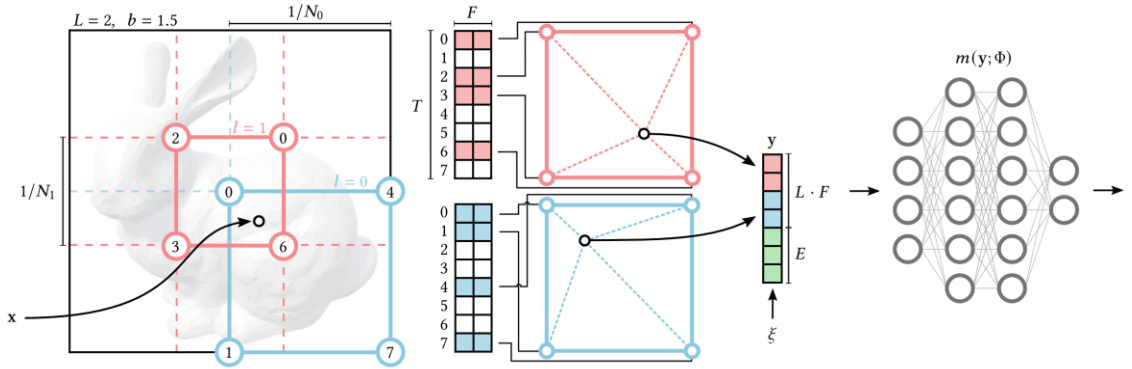
动作与视点可控人体视频生成

- 提出Interspatial Attention，借助NeRF当中的隐式编码思想，无需渲染直接建立起3D SMPL与2D Video的关联



$$\mathbf{g}_{clip} = \left[x_{clip}, y_{clip}, z_{clip}, w_{clip}\right]^T = \mathbf{M}\mathbf{g},$$

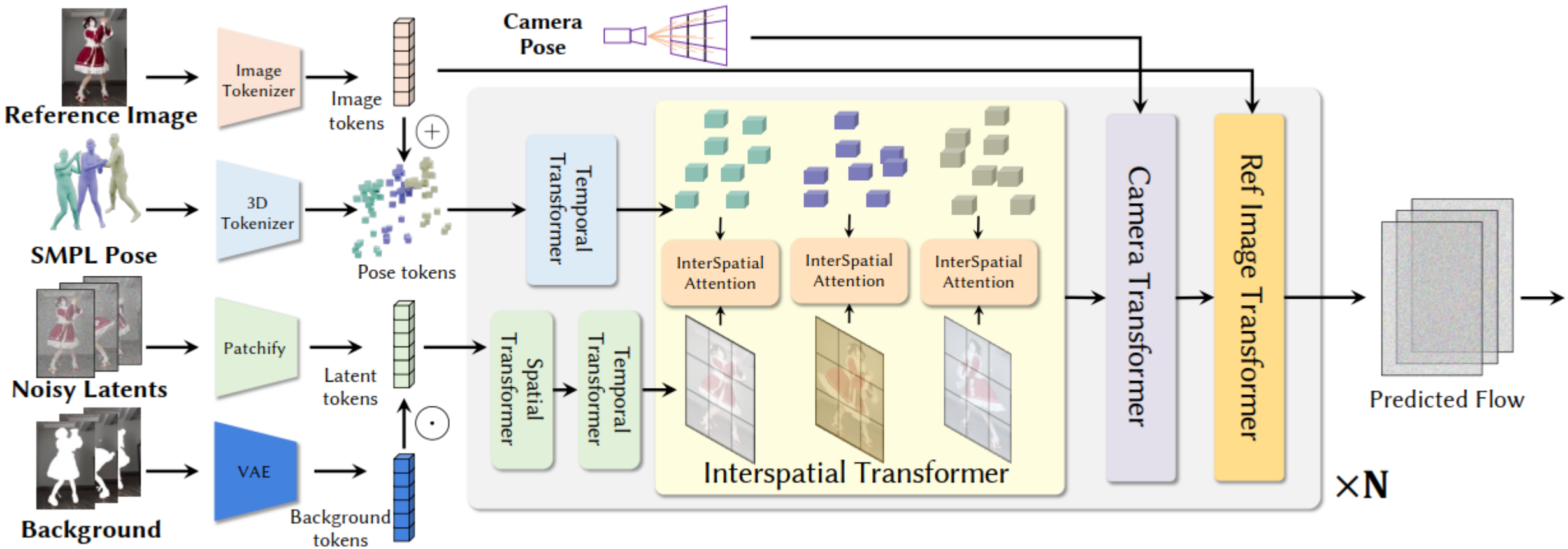
$$\mathbf{g}_{ndc} = \left[\frac{x_{clip}}{w_{clip}}, \frac{y_{clip}}{w_{clip}}, \frac{z_{clip}}{w_{clip}}\right]^T.$$



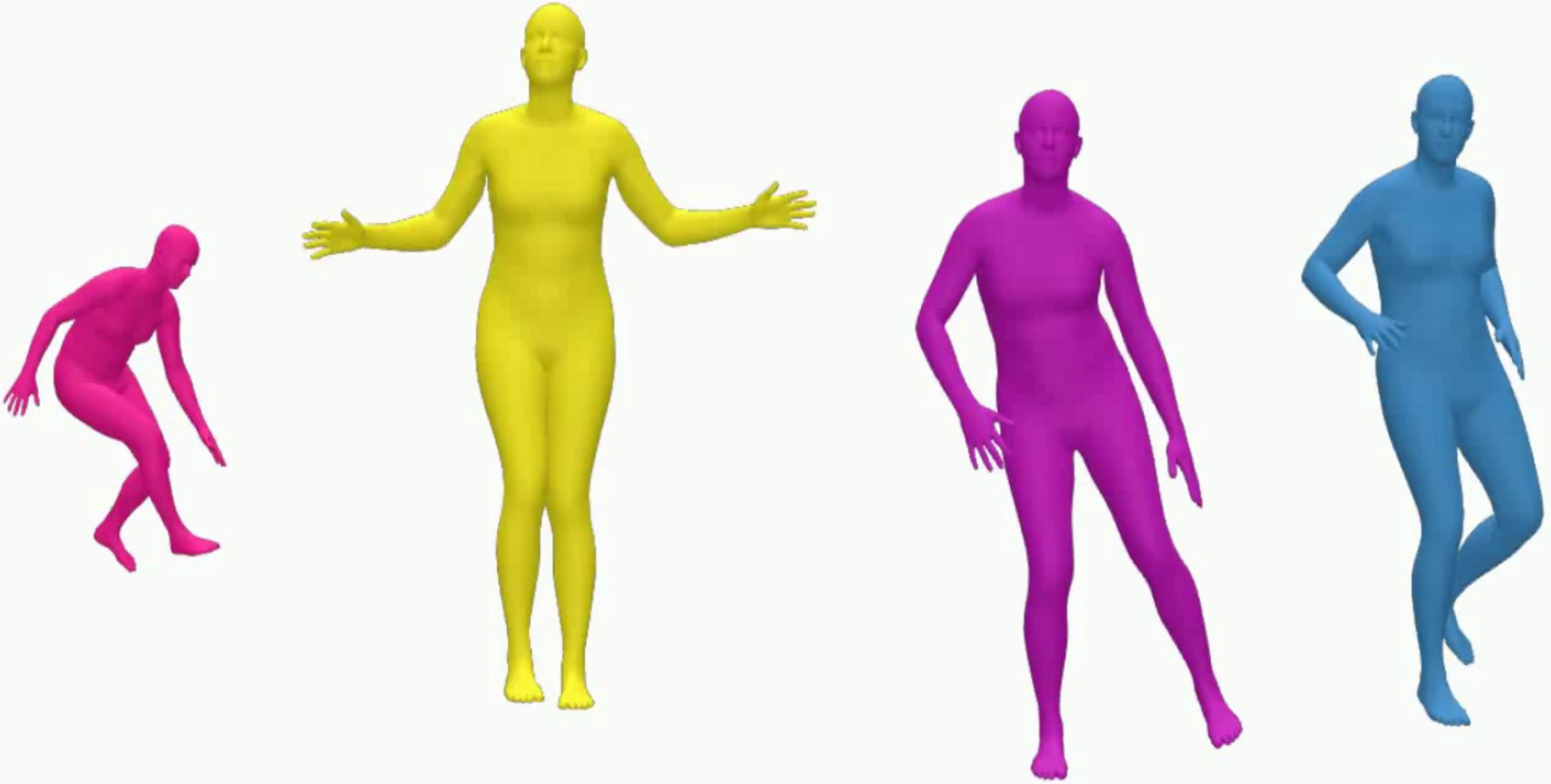
$$\mathbf{z}'_j = \text{ISATTENTION}(\text{Q}(\mathbf{z}_j + \text{PE}(\mathbf{s}_{ndc})), \text{K}(\mathcal{Y}_j + \text{PE}(\mathbf{g}_{ndc})), \text{V}(\mathcal{Y}_j + \text{PE}(\mathbf{g}_{ndc}))).$$

动作与视点可控人体视频生成

- 无需2D渲染，直接高效建立起3D与2D video的关联实现2D、3D统一的扩散变换器大模型

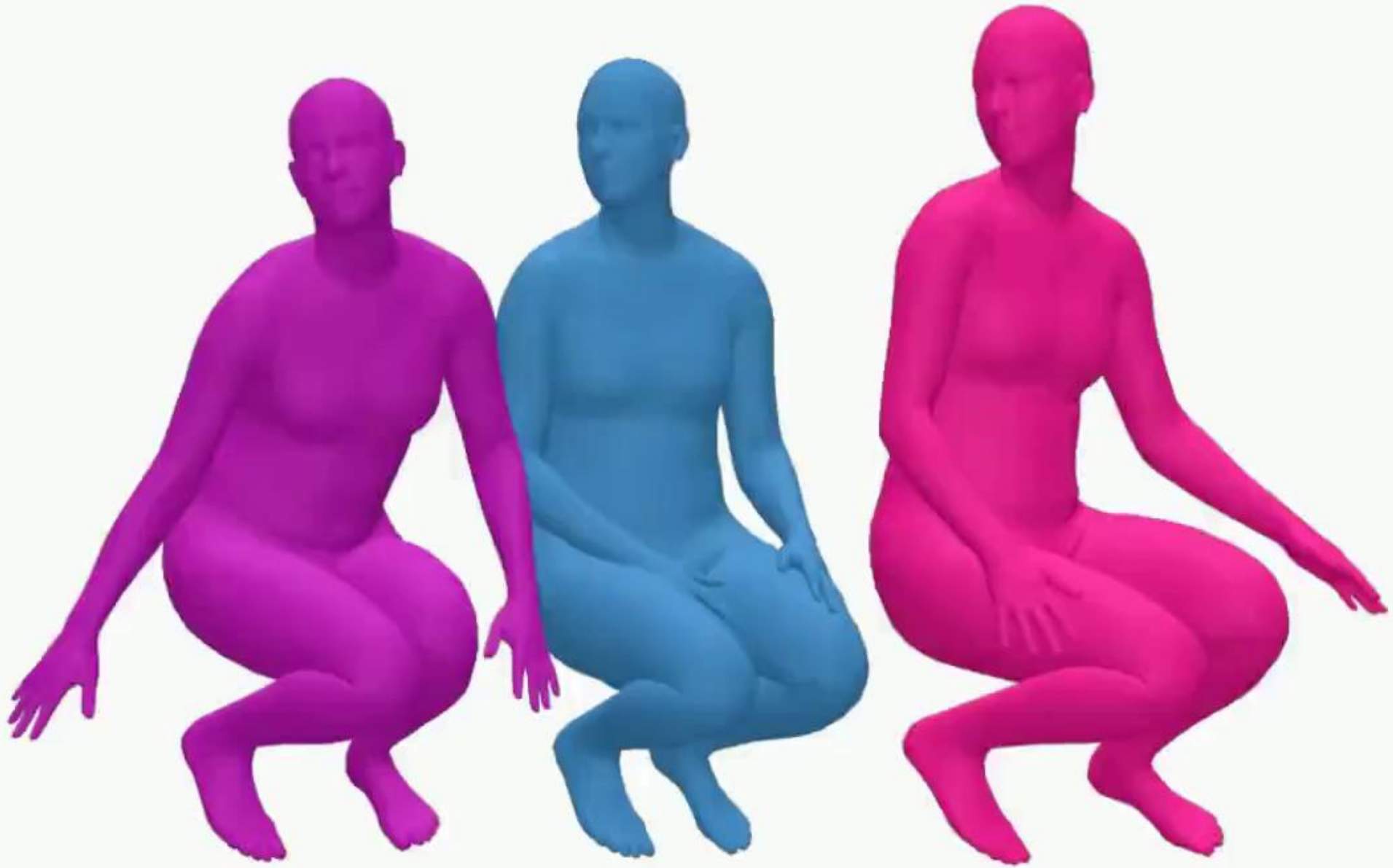


动作与视点可控人体视频生成



动作与视点可控人体视频生成




1.0
0.8
0.6
Z
0.4
0.2
0.0
1.0
0.8
0.6
0.4
Y
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
X

基于视频生成的数字化身

- 优点
  - 泛化性强
  - 支持复杂外观、复杂表情、复杂运动下的动态生成
  - 兼容场景生成
  - 动作表情可控性强
  - 物理较合理，不存在穿模等问题
- 缺点
  - 依赖三维控制信号驱动（未来可**共生生成**，同时动作与外观生成）
  - VR/元宇宙 应用：三维一致性与显式三维表征仍需加强
  - 驱动速度（Diffusion加速）

谢 谢!