之江实验室
ZHEJIANG LAB

浙江大学
ZHEJIANG UNIVERSITY

基于姿态的个性化可控的武打动作生成方法

目录 CONTENTS

一 研究背景

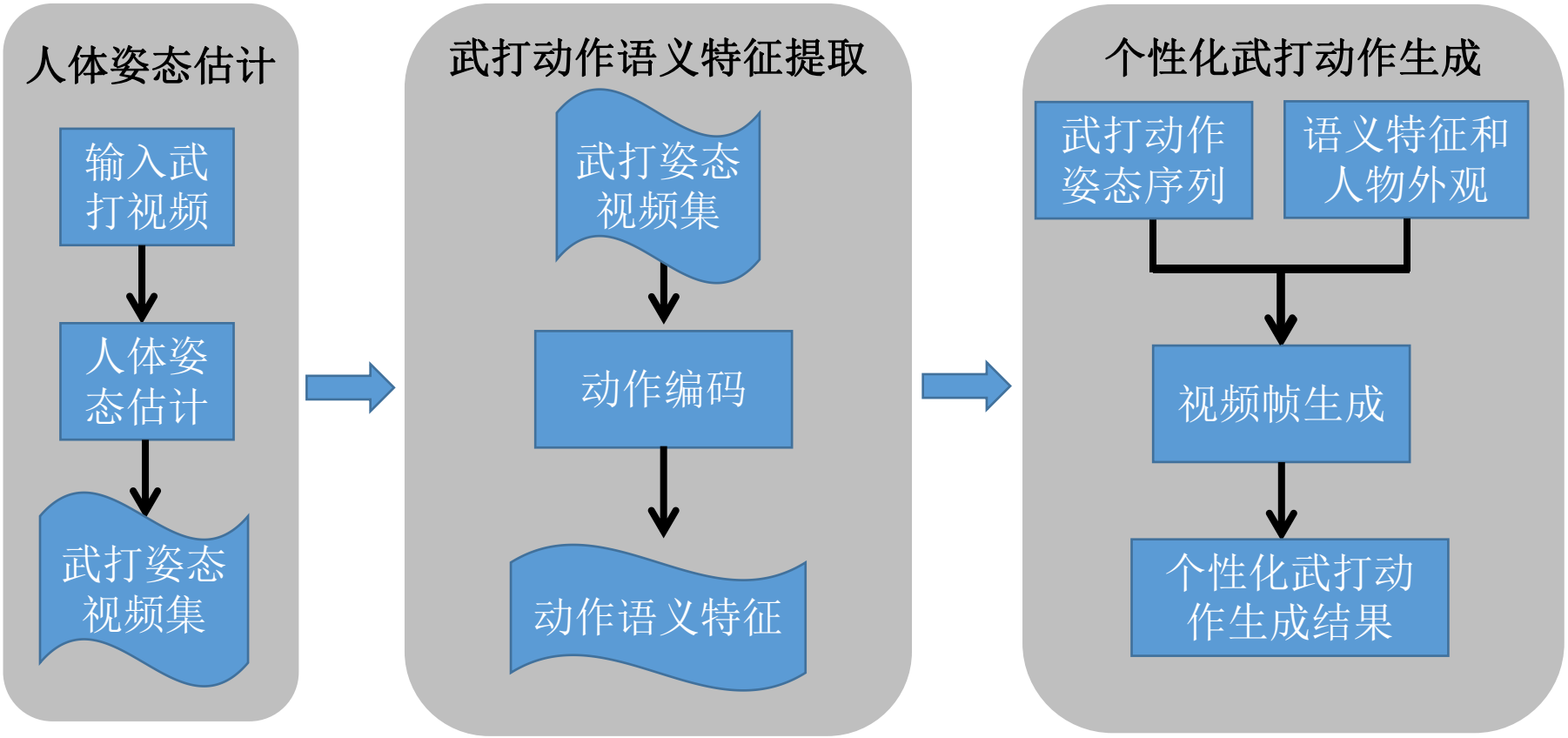
□ 基于姿态的个性化可控的武打动作生成

- ➢ 基于姿态的武打动作生成
  - 通过学习人体结构和运动特征，实现从特征表示到武打动作视频帧间的映射
- ➢ 语义驱动的武打动作生成
  - 给定一段关于武打动作的语义特征嵌入，生成相应的个性化武打动作视频
- ➢ 研究难点
  - 寻找足以表达、刻画人体结构和人物动作的特征表示
  - 寻找根据语义特征表示、姿态和个性化外观生成武打动作视频帧的方法

二 研究概述

基于个性化可控的武打动作视频生成

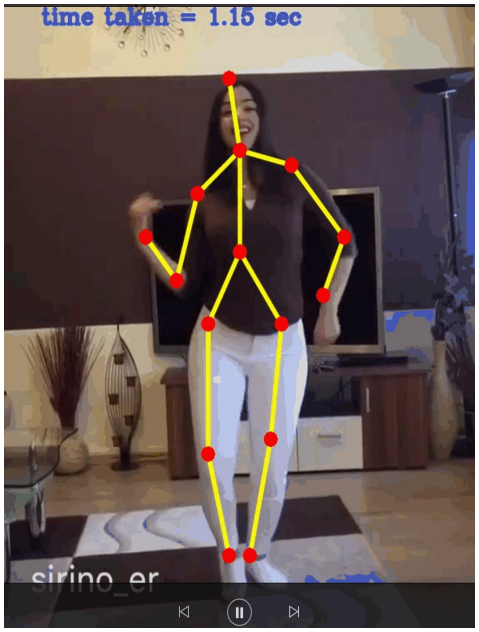
- 研究思路：基于语义特征以及个性化人物外观，学习从姿态到真实视频帧的映射

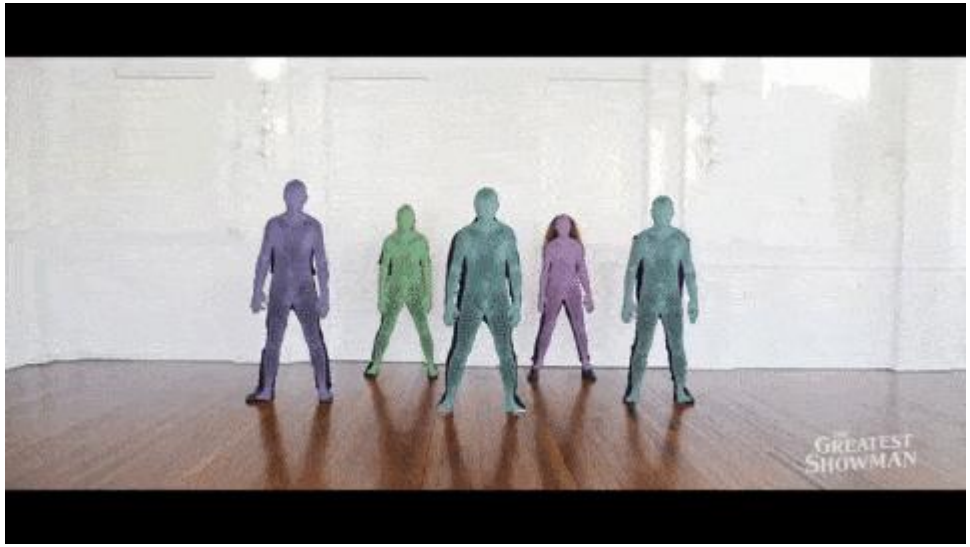


二 研究概述

武打视频人体姿态估计

- 基于骨架的2D人体姿态估计
  - 可控、紧凑、易提取
  - 生成视频时，仅用骨架，纹理特征需依靠生成网络填补，生成质量易不理想
- 3D人体姿态估计
  - 3D人体姿态标注困难
  - 相较骨架姿态，3D人体姿态到2D视频帧的生成质量更加可控

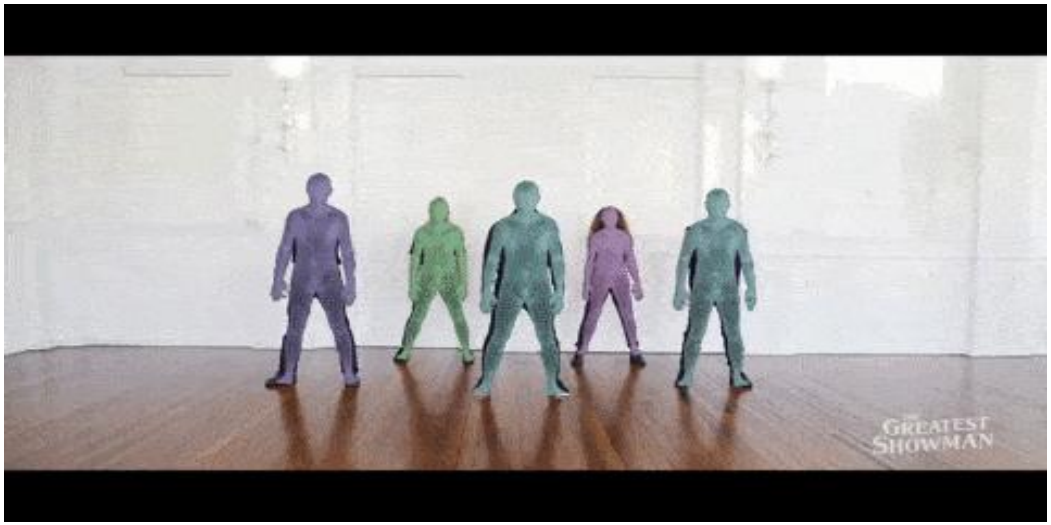




二 研究概述

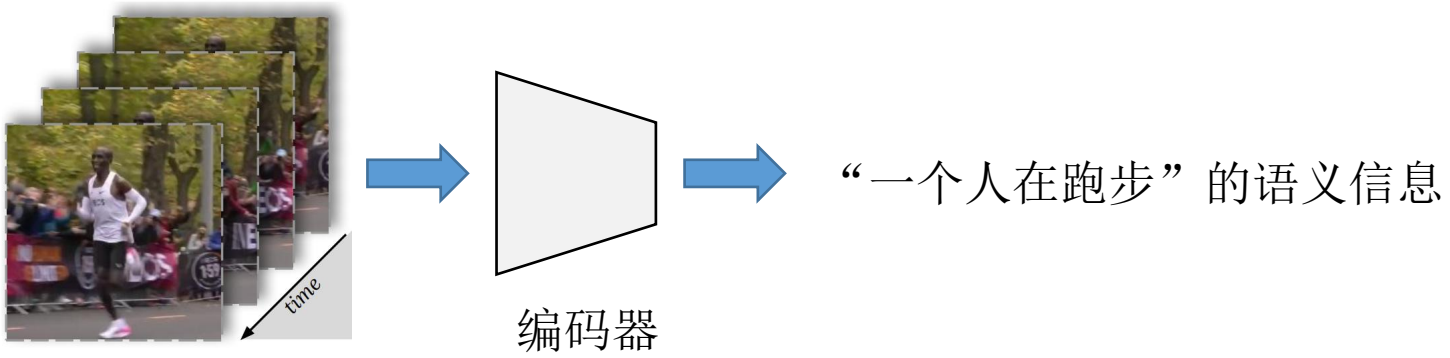
□ 武打视频人体姿态估计

- 面向武打动作视频生成需求，以视频中的3D人体姿态估计为研究场景，围绕如何建立更为有效的3D人体姿态估计算法开展研究。
- 研究难点
  - 2D场景下，被遮挡的人体部位视觉信息难以还原，人体姿态估计不理想
  - 3D人体姿态标注困难



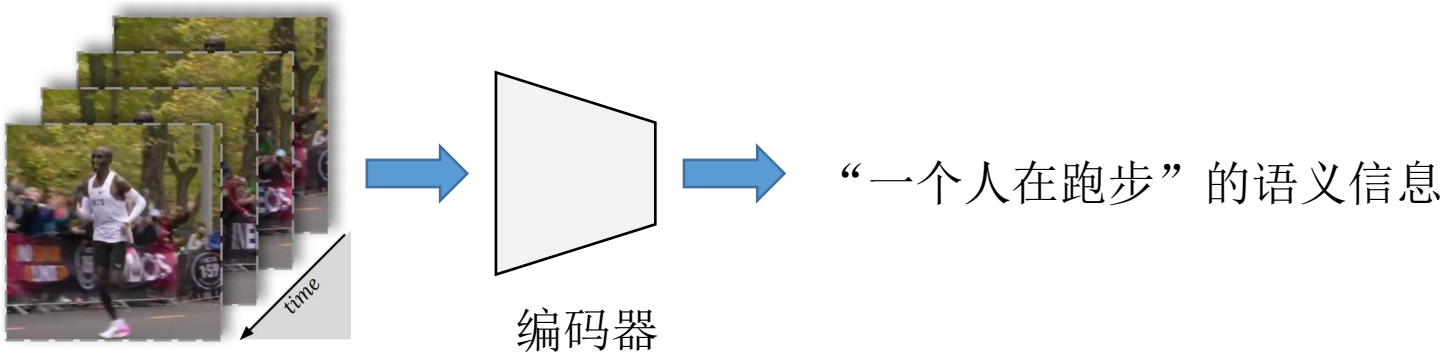
武打动作视频语义提取

- 给定一段视频，提取人物武打动作的语义信息。



武打动作视频语义提取

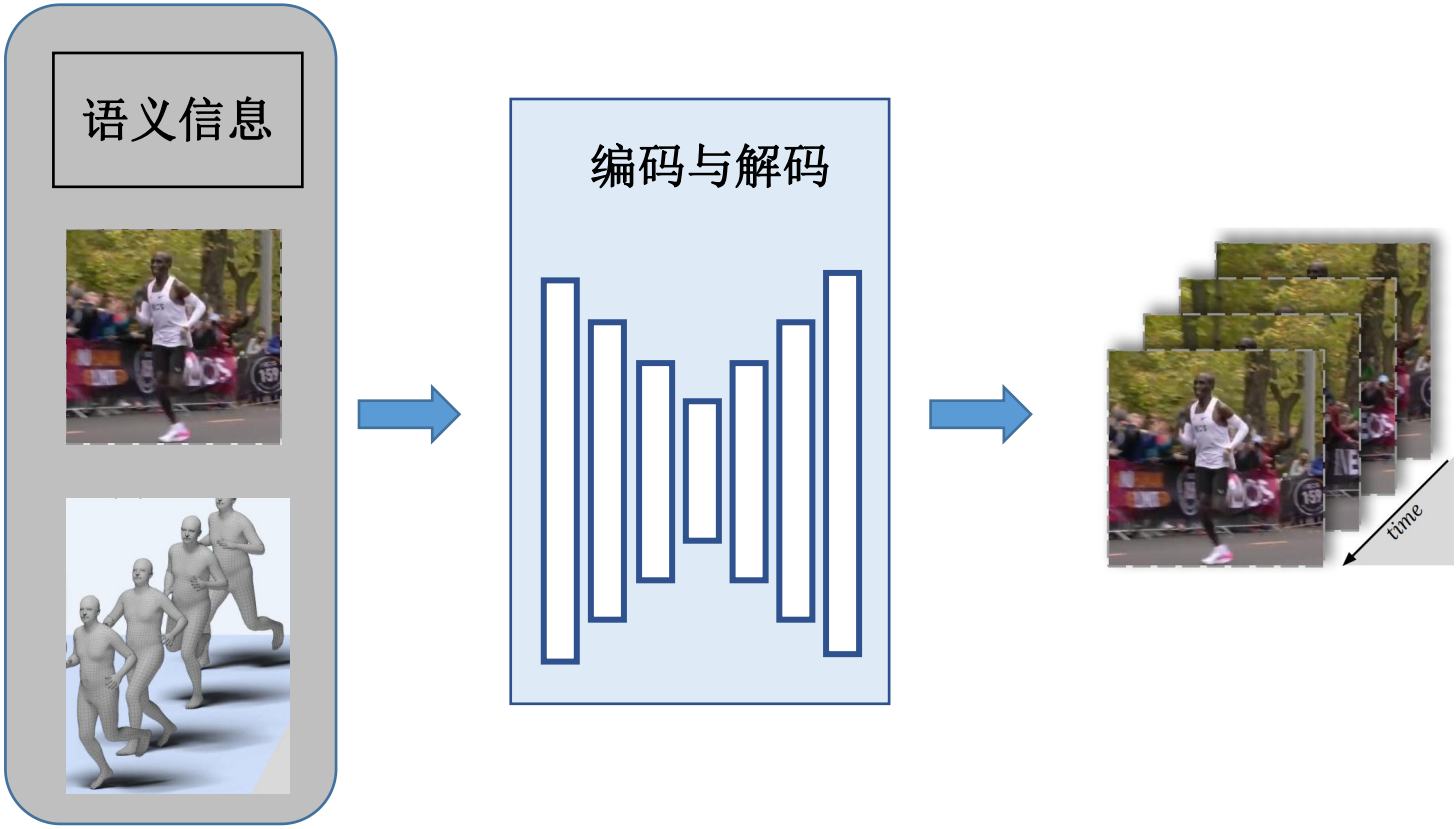
- 面向武打动作视频语义提取，以特征嵌入为研究场景，围绕如何建立有效的编码器提取视频中人物的动作语义特征。
- 研究难点
  - 缺少动作相关的语义标签
  - 抽取的语义特征易受视频背景影响



研究概述

个性化可控武打动作生成

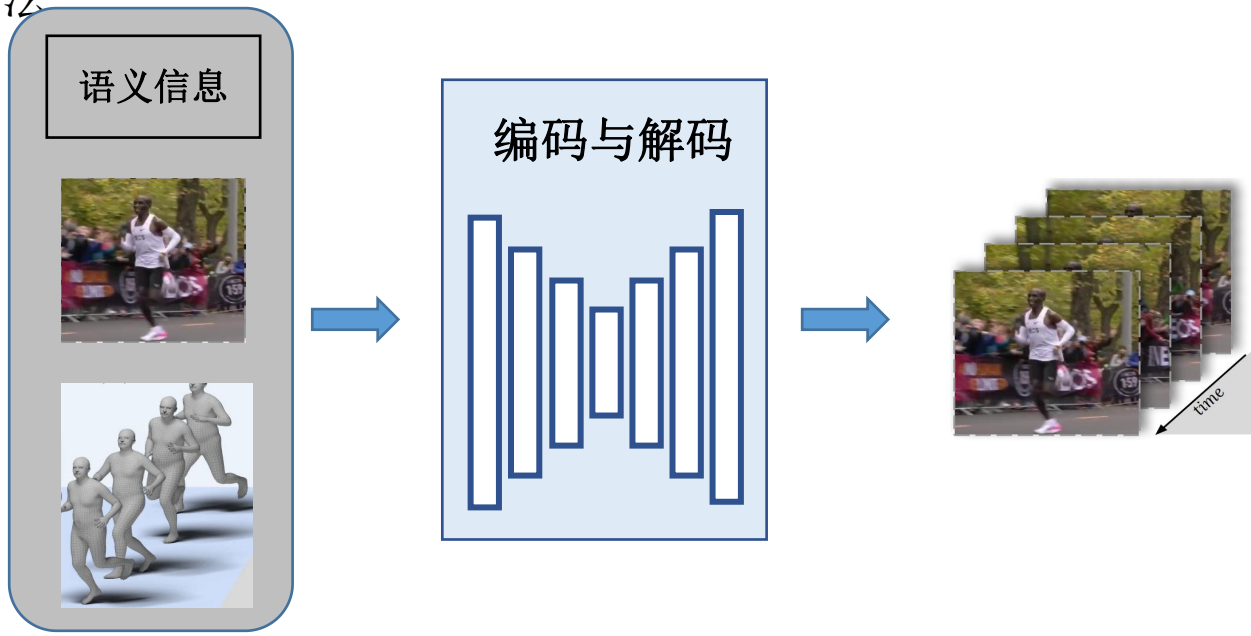
- 给定语义信息和人物外观，生成相应的个性化武打动作视频。



研究概述

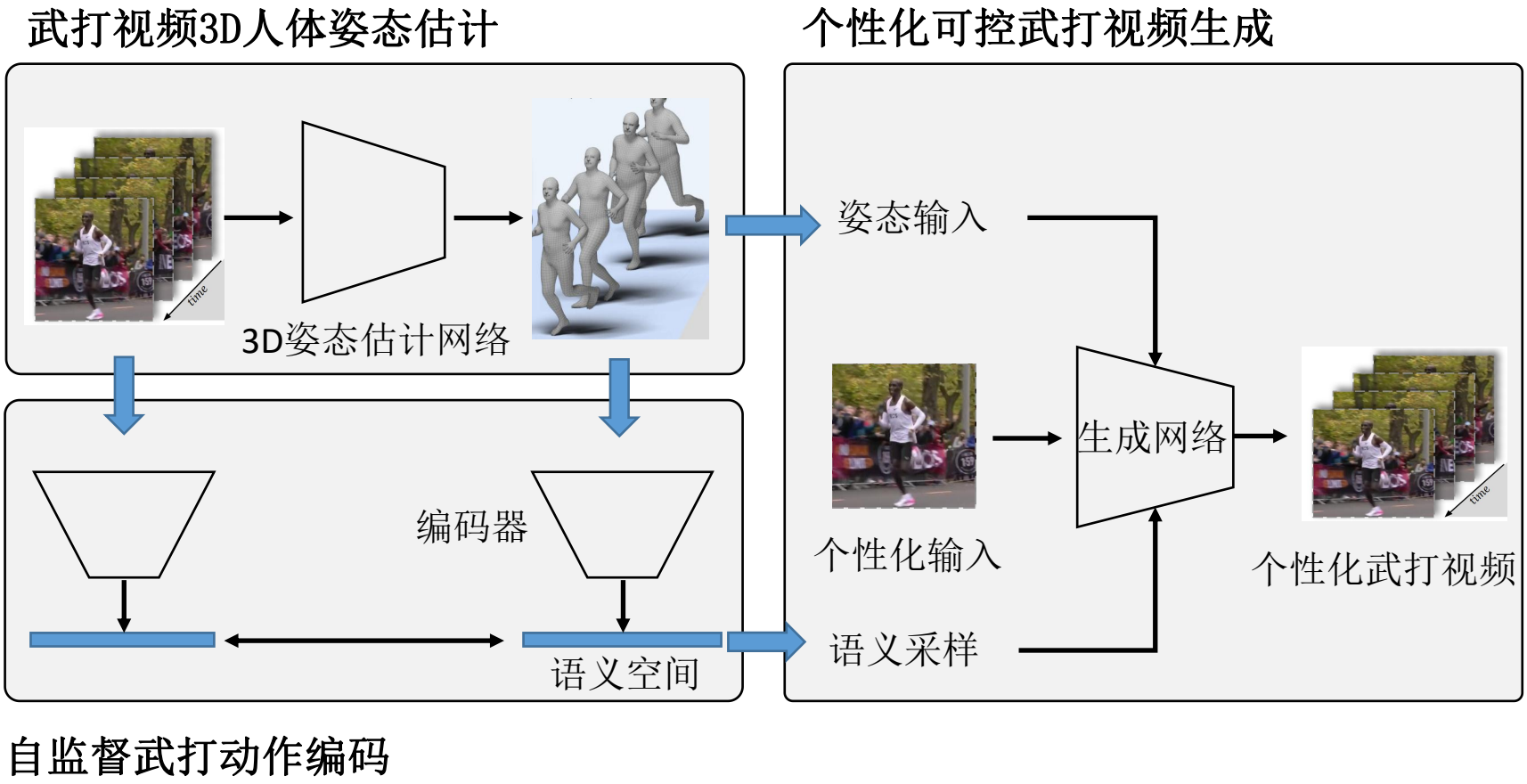
个性化可控武打动作生成

- 面向个性化可控武打动作生成需求，以视频生成为研究场景，围绕如何建立有效的个性化的武打动作视频生成方法开展研究。
- 研究难点
  - 寻找根据语义特征、人物外观以及人体姿态的特征表示生成武打动作视频帧的方法



研究方案

整体框架

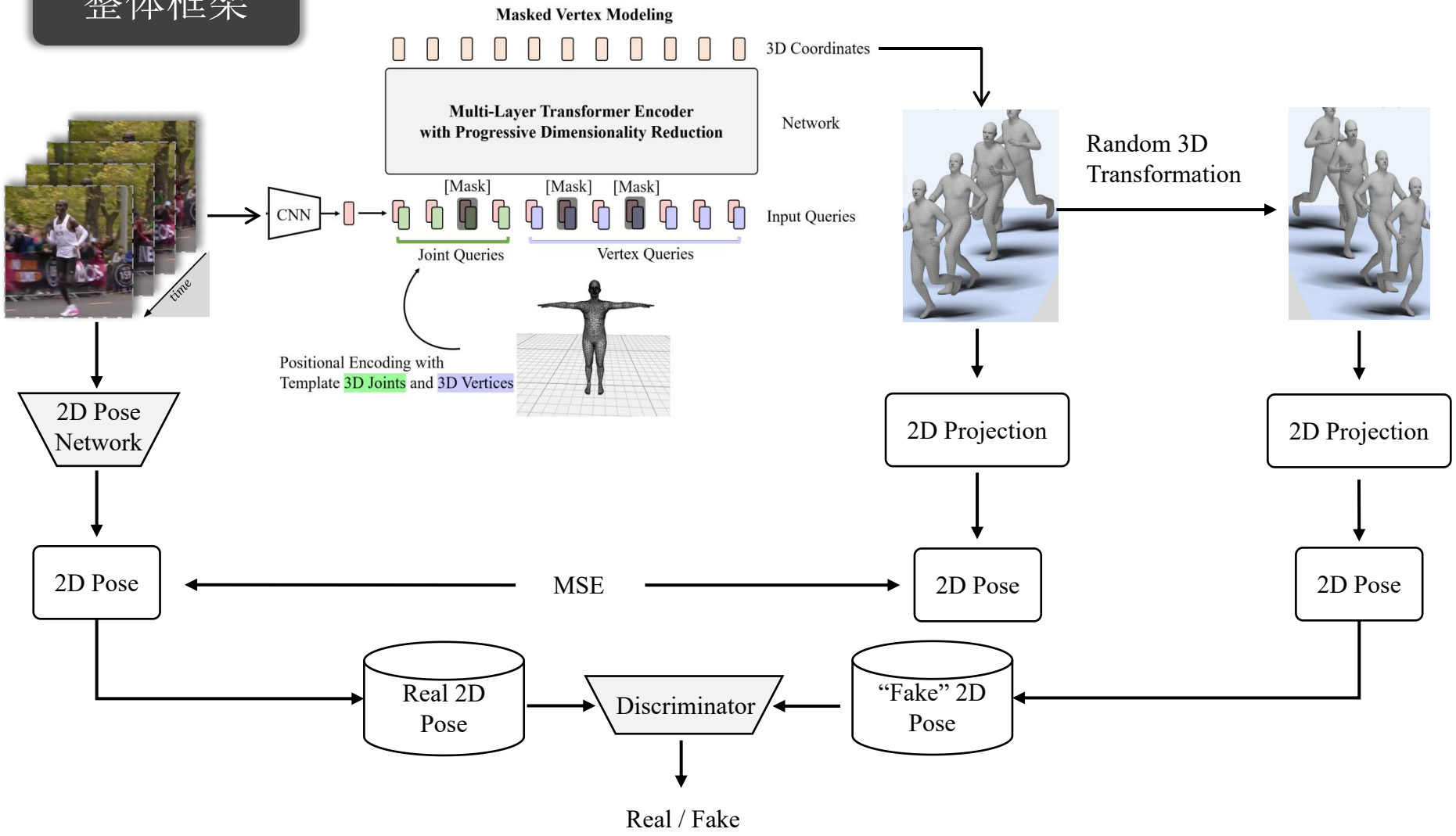

武打视频3D人体姿态估计
个性化可控武打视频生成
time
3D姿态估计网络
姿态输入
编码器
生成网络
个性化输入
个性化武打视频
time
语义空间
语义采样
自监督武打动作编码

三 研究方案

武打视频3D人体姿态估计

- 基于Transformer的无监督3D人体姿态估计方法
  - 研究动机
    - 现有视频3D人体姿态估计方法聚焦于回归姿态和形状参数
    - 现有方法依赖于大量的3D姿态真值标签
    - 如何在不依赖任何参数模型和3D标签情况下进行3D人体姿态估计
  - 设计具有渐进降维的多层Transformer编码器
  - 设计掩码向量建模目标，对顶点-顶点和顶点-关节的相互作用进行建模，以实现更好的重建
  - 直接并行输出人体关节和网格顶点的三维坐标
  - 通过3D-2D的投影误差以及判别器生成真实3D姿态

整体框架

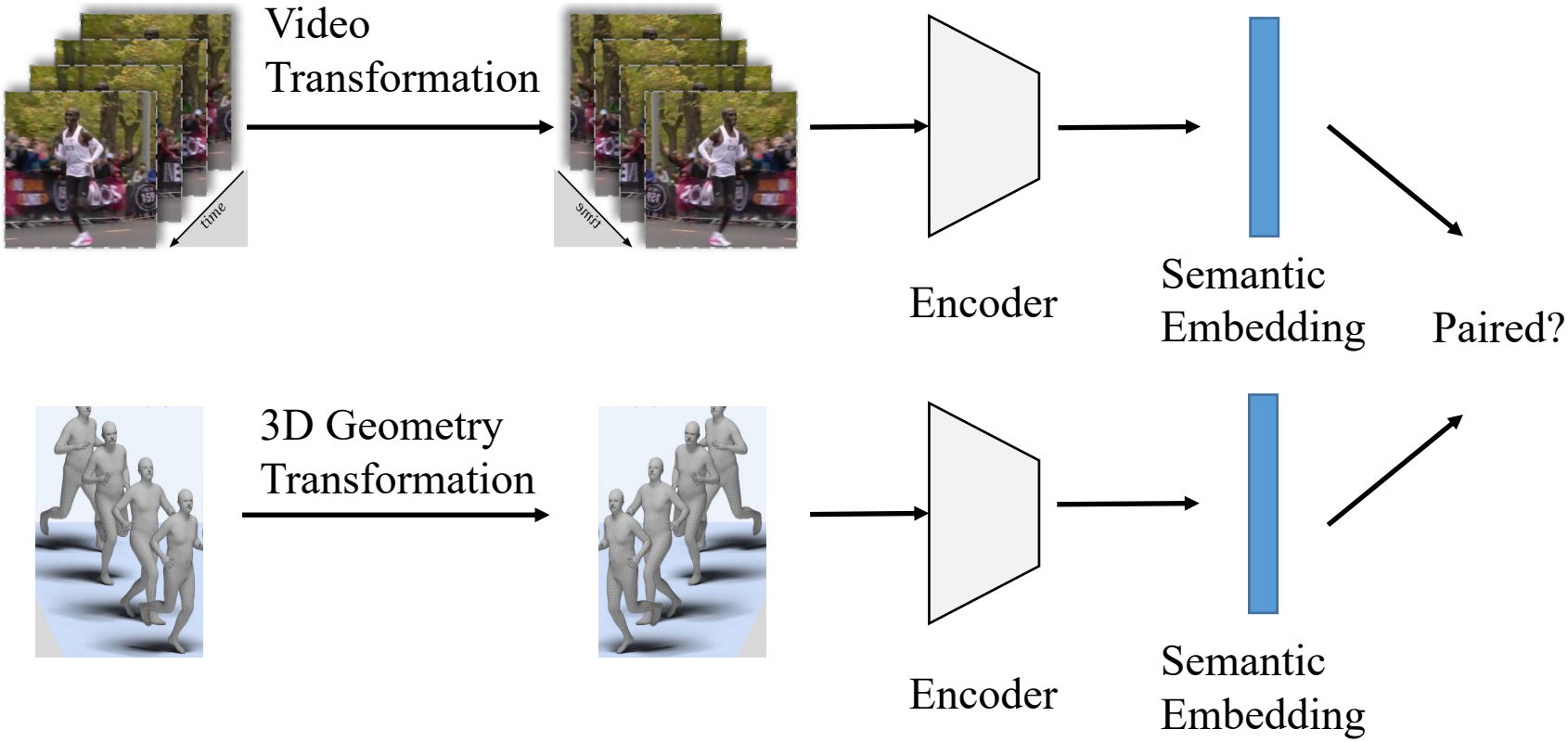

Masked Vertex Modeling
3D Coordinates
Multi-Layer Transformer Encoder
with Progressive Dimensionality Reduction
Network
[Mask]
[Mask]
[Mask]
Input Queries
Joint Queries
Vertex Queries
CNN
time
Positional Encoding with
Template 3D Joints and 3D Vertices
Random 3D
Transformation
2D Pose
Network
2D Projection
2D Projection
2D Pose
MSE
2D Pose
2D Pose
Real 2D
Pose
Discriminator
“Fake” 2D
Pose
Real / Fake

三 研究方案

自监督武打动作编码

- 基于自监督对比学习的武打动作语义特征编码方法
  - 研究动机
    - 根据给定的成对武打视频与3D姿态，生成具有语义信息的编码
  - 设计具有武打动作语义信息的特征嵌入提取方法
  - 将视频流与3D姿态流映射到统一的语义空间
  - 采用自监督对比表示学习

整体框架


Video
Transformation
time
time
Encoder
Semantic
Embedding
Paired?
3D Geometry
Transformation
Encoder
Semantic
Embedding

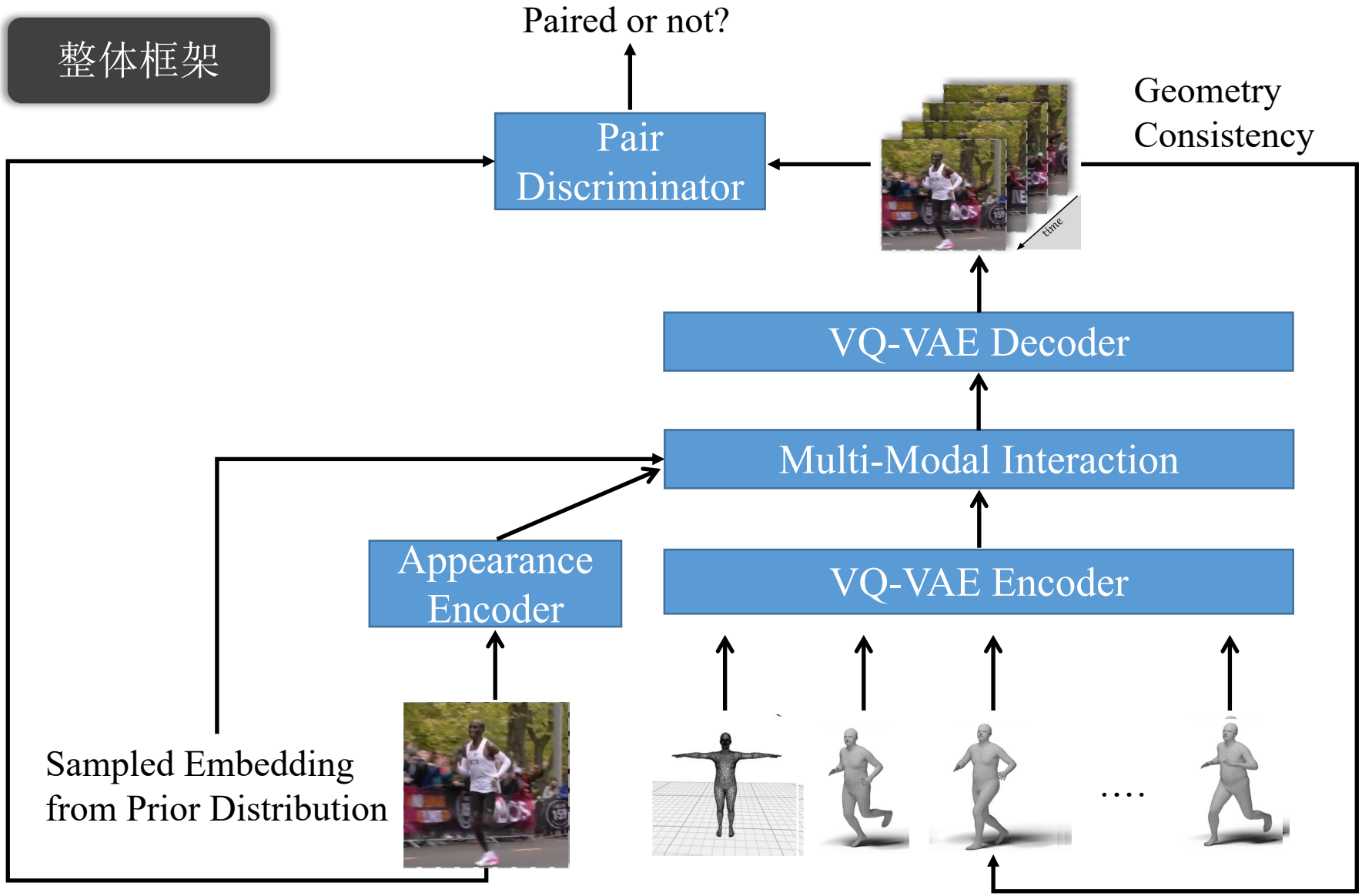
三 研究方案

个性化可控武打视频生成

- 基于人物外观可控的个性化视频生成方法
  - 研究动机
    - 根据给定先验嵌入采样、人物外观以及姿态估计，生成可控的视频片段
  - 设计自回归框架处理跨模态武打动作视频生成
  - 采用VQ-VAE 编码-解码结构，用于视觉token表示学习
  - 将人物外观特征融入视觉特征，用于可控和个性化视频帧生成

三 研究方案

整体框架


Paired or not?
Pair Discriminator
Geometry Consistency
time
VQ-VAE Decoder
Multi-Modal Interaction
Appearance Encoder
VQ-VAE Encoder
Sampled Embedding from Prior Distribution
....

四 预期成果与创新点

- 提出视频三维人体姿态估计方法，实现有效的武打动作三维人体姿态估计框架
- 提出自监督武打动作语义特征编码方法，实现视频流与姿态流的跨模态语义提取
- 提出个性化可控的武打动作生成方法，实现开放环境下基于语义信息的个性化武打动作生成