北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

Text-to-Anything 之搭积木方法浅析

汇报人：关晓宇　　时间：2024/10/21

搭积木传奇 美妙至极

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

Siggraph 2024

Text to Crowd Animation

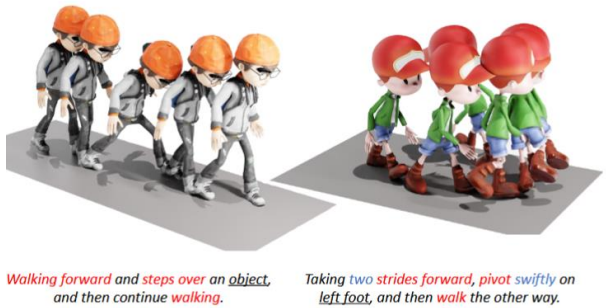
Text to Cloth

a person stretches out his two arms and dances

a person walks forward while holding out their arms for balance

Text to Motion

CVPR 2024



Text To Motion

Arxiv

Text to Crowd Animation

Text to 2Dface

搭积木传奇 美妙至极

Siggraph 2024

To the best of our knowledge, we introduce the **first-ever** pipeline that targets at **language-guided generation** of environment compatible scenarios involving a large number of agents navigating in real-time.

Text to Crowd Animation

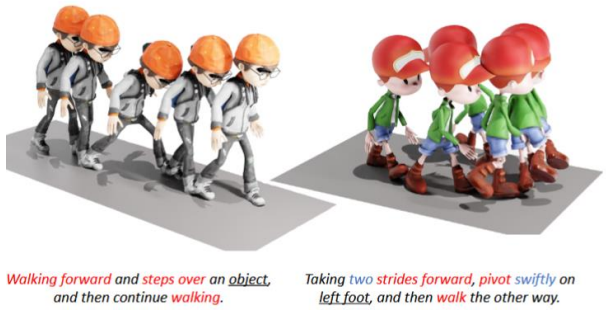
Text to Cloth

a person stretches out his two arms and dances

a person walks forward while holding out their arms for balance

Text to Motion

CVPR 2024



Text To Motion

Arxiv

Text to Crowd Animation

Text to 2Dface

搭积木传奇 美妙至极

Siggraph 2024

To the best of our knowledge, we introduce the **first-ever** pipeline that targets at **language-guided generation** of environment compatible scenarios involving a large number of agents navigating in real-time.

Text to Crowd Animation

We propose a **first text-driven garment generation** pipeline with high-quality garment sewing patterns and physically based textures.

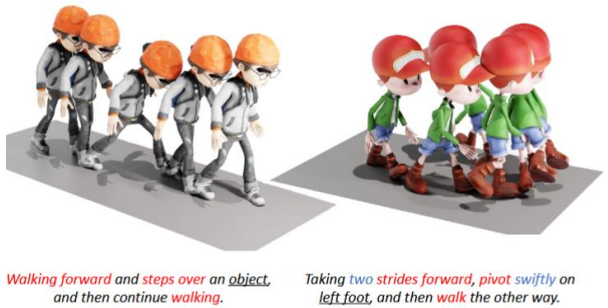
Text to Cloth

a person stretches out his two arms and dances

a person walks forward while holding out their arms for balance

Text to Motion

CVPR 2024



Text To Motion

Arxiv

Text to Crowd Animation

Text to 2Dface

搭积木传奇 美妙至极

Siggraph 2024

To the best of our knowledge, we introduce the **first-ever** pipeline that targets at **language-guided generation** of environment compatible scenarios involving a large number of agents navigating in real-time.

Text to Crowd Animation

We propose a **first text-driven garment generation** pipeline with high-quality garment sewing patterns and physically based textures.

Text to Cloth

a person stretches out his two arms and dances

a person walks forward while holding out their arms for balance

Text to Motion

CVPR 2024

MoMask is the **first generative masked modeling framework** for the problem of text-to-motion generation.

Text To Motion

Walking forward and steps over an object, and then continue walking.

Taking two strides forward, pivot swiftly on left foot, and then walk the other way.

Arxiv

Text to Crowd Animation

Text to 2Dface

搭积木传奇 美妙至极

Siggraph 2024

To the best of our knowledge, we introduce the **first-ever** pipeline that targets at **language-guided generation** of environment compatible scenarios involving a large number of agents navigating in real-time.

Text to Crowd Animation

We propose a **first text-driven garment generation** pipeline with high-quality garment sewing patterns and physically based textures.

Text to Cloth

a person stretches out his two arms and dances

a person walks forward while holding out their arms for balance

Text to Motion

CVPR 2024

MoMask is the **first generative masked modeling framework** for the problem of text-to-motion generation.

Text To Motion

Arxiv

To our best knowledge, it is the **first text-guided** 2D-based talking face generation framework.

Text to Crowd Animation

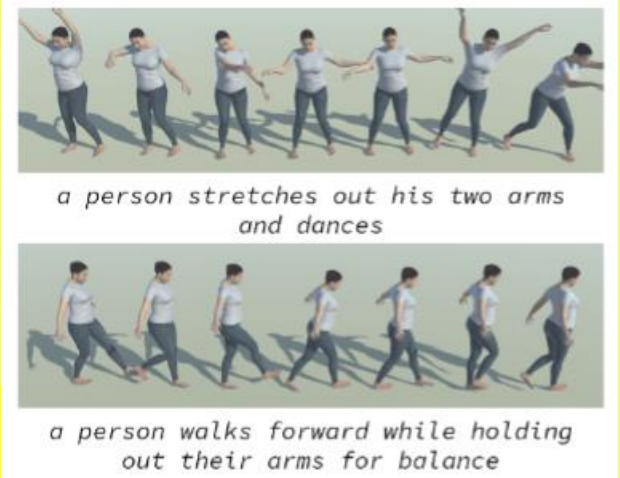
Text to 2Dface

搭积木传奇 美妙至极

Siggraph 2024

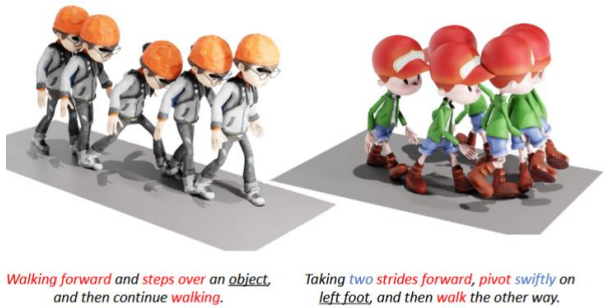
Text to Crowd Animation

Text to Cloth



Text to Motion

CVPR 2024



Text To Motion

Arxiv

Text to Crowd Animation

Text to 2Dface

积木速览目录

1）看商品图　　**看实现效果**

2）看看品牌　　**看架构**

3）查下成分　　**看实现细节**

4）货比三家　　**对比（包括行文思路、实验设置等）**

5）总结　　**积木搭建法则**

Text-Guided Synthesis of Crowd Animation 【腾讯+港大】

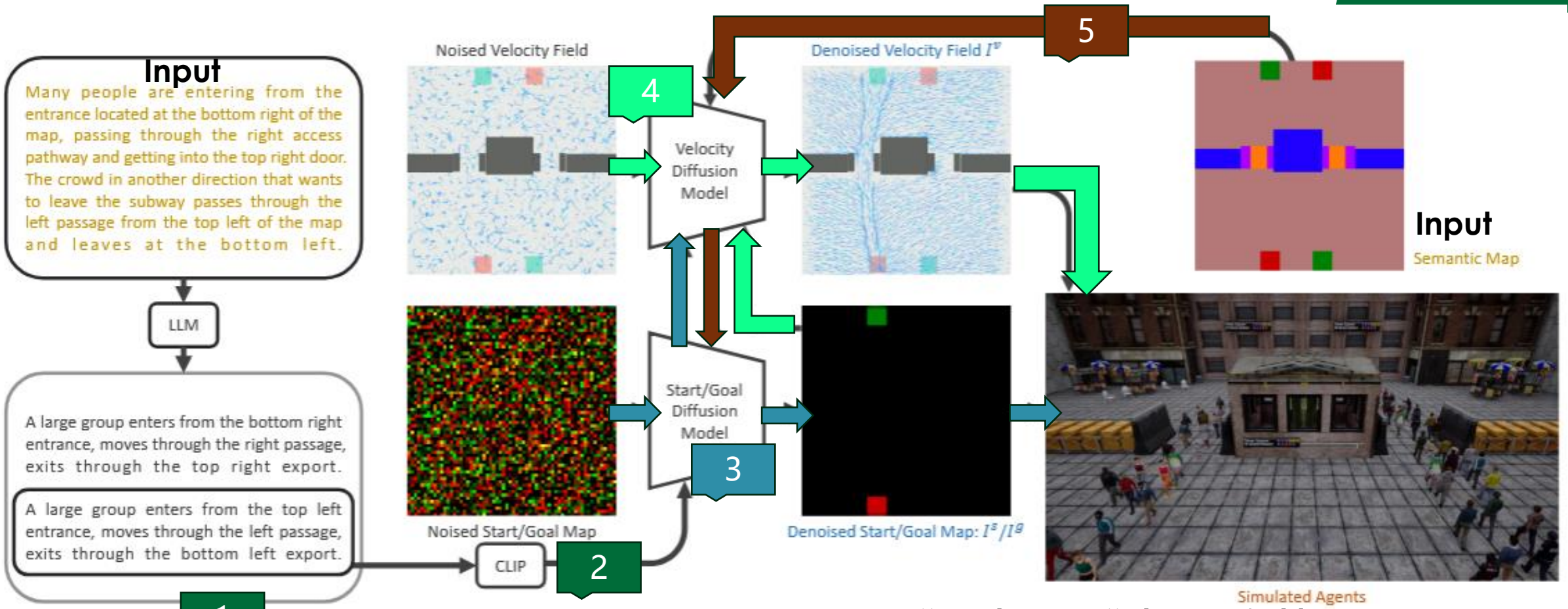
Input:
There are hundreds of humans escaping through the passage from the open space above the map to the exit below.

Some humans are moving from the right of the map and crossing the crosswalk to get to the left. Some people also enter from the right are walking along the left side of the building and finally exit at the top right of the map. Others get into the map from the top left and move along the right of the other building, then leave at the left.

Many people are entering from the entrance located at the bottom right of the map, passing through the right access pathway and getting into the top right door. The crowd in another direction that wants to leave the subway passes through the left passage from the top left of the map and leaves at the bottom left.

Input:
This garden contains a total of 6 small groups of visitors. The first two groups enter from the main entrance in the top right corner and both of them walk through the upper right passage first. Afterward, one group passes the left side of the circular fountain, then gets through the garden passage and leaves at the bottom left exit. Another group follows a different path where they circle around the upper left rectangular groves counterclockwise and visit the fountain, finally exit through the bottom gate. Another two groups get in from the right gate. One group just visits the fountain and leaves through the bottom left export. Another group firstly moves towards the triangular entertainment area and circles around the lower side of it clockwise, then they pass by the fountain and circle around the upper left groves counterclockwise, finally leave through the left gate. The final two groups enter from the top entrance. One of them passes by the right of the upper left groves, circles around the fountain, and leaves the map through the left gate. Another small group visits the lower side of the fountain first. Then, they get around the triangular entertainment area and exit through the bottom export.

Text-Guided Synthesis of Crowd Animation



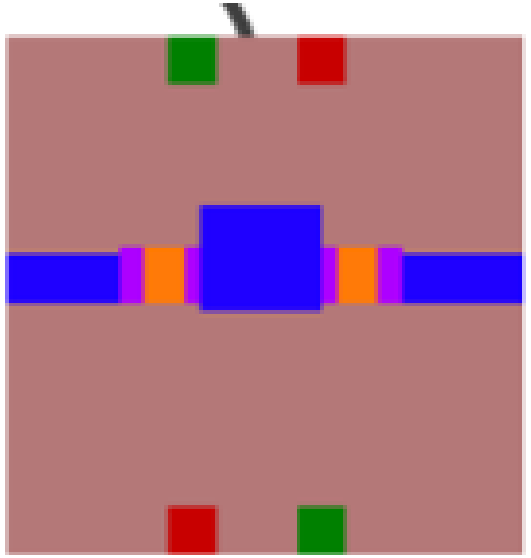
text->多样化动态人群动画；自制dataset

CLIP+2Latent Diffusion+RVO+grouping

each group controlled by 2张图片

Text-Guided Synthesis of Crowd Animation

人群仿真agent-based方法太慢，因此使用图像-based，同时方便生成。



Semantic Map

- 环境信息为one-hot编码，表示为semantic map 9维，[前8维为障碍物类型，最后一维是pixel位置信息]
- LDM的guidance信号为规范化范式后的prompt
- 人群初始化：根据起始区概率分布，泊松sampling去重

Name	Circular Obs.	Cubical Obs.	Triangular Obs.	Zebra Crossing	Narrow Passage	Start	Goal
Figure							
Anchors	TL,TR,BL,BR	TL,TR,BL,BR	T,L,R	T,B,L,R	T,B	C	C
Edges	T,B,L,R	T,B,L,R	B,L,R	V,H	V	N/A	N/A
#Pixel	[70-100]	[70-100]	[70-100]	[170-200]	[170-200]	80	80

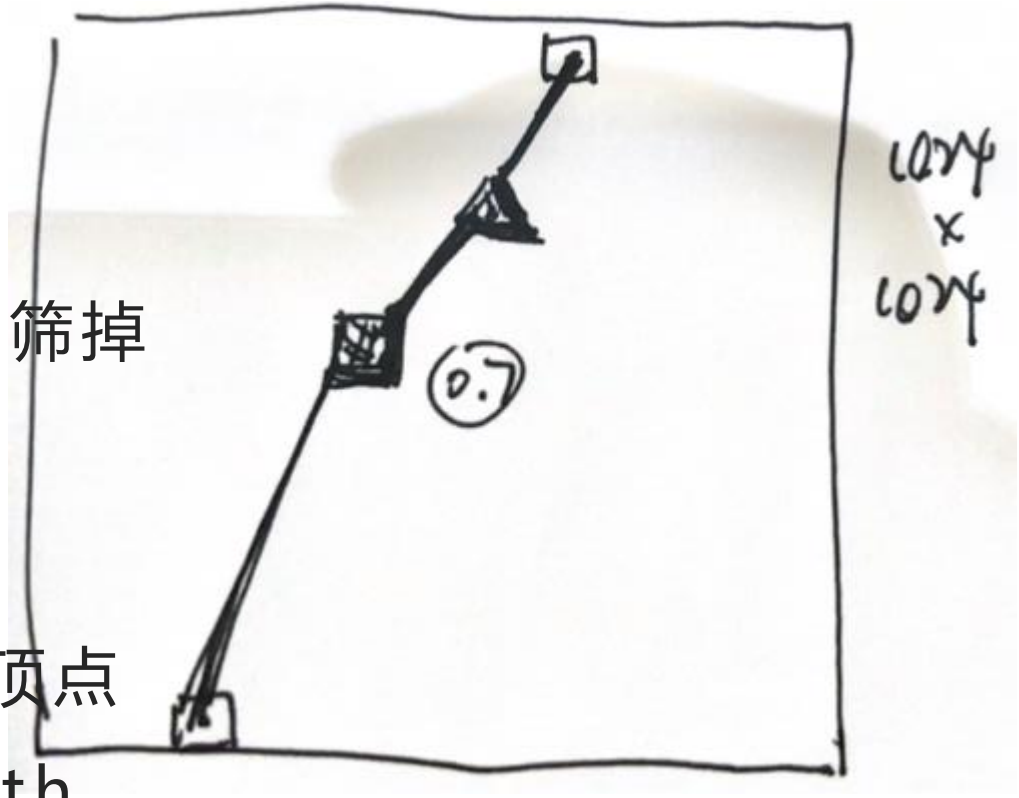
- Agent为圆盘建模
- 训练Loss：DDPM（就是diffusion正常loss）
- 优化器：AdamW

Text-Guided Synthesis of Crowd Animation

自制Dataset：**57600个场景**；每个场景包含元素：**障碍物编码，规范化文本描述，start/goal map图片，速度场图片**

Dataset生成方式：

1. 场景中随机设置1-3组人，每组随机放start/end point
2. 场景中随机设置0-5个5种类型的障碍物，筛掉撞的、出界的等
3. 生成semantic map的one-hot编码
4. 为了避开障碍物，根据起点终点和障碍物顶点生成粗略的navigation path，排除异常path



Text-Guided Synthesis of Crowd Animation

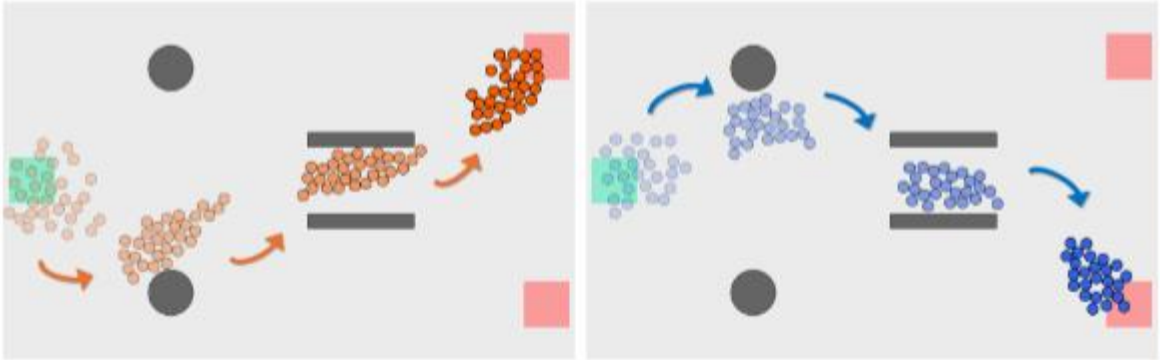
Dataset生成方式：

5. 根据规则生成规范化文本描述

6. 基于[Rezende et al. 2021]生成速度场，用RVO迭代15次调整速度场以生成异质性结果

Constructive time-varying vector fields for robot navigation

创建一个稳定的、误差校正的速度场，这样，当agent偏离路径时，它们将被引导回去跟随路径



(a) *Two different crowd simulations generated with the same prompt "A big group enters from the left entrance, moves past the circle, walks through the passage, exits through the exit.". The drastically different crowd distributions and behaviors demonstrate the controllable diversity of our model.*

(b) *Compared with the model trained without field adjustment (left), our model (right) generates more scattered and realistic agent motions.*

Text-Guided Synthesis of Crowd Animation

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

Evaluation

三个scene，没有设置清每组的人数（train时每组<=50人）

效率不错：less than 6 seconds to infer a set of three maps, and the runtime simulation cost is less than 1ms per timestep

定量分析：指标4个

1. Dynamic Time Warping(speed changes due to inter-agent collisions)
2. Average Agent-to-Trajectory Distance
3. Strict Success Rate(within trajectory margin)
4. Relax Success Rate

用户研究：12人参与，2场景，对比GAScrowd [Kim12]

消融实验：

2个针对prompt的+一个针对RVO调整速度场的

- 仿真平台：**CARLA**

开源模拟器，自动驾驶领域【魔改版UE4.26or5.3+python】

DressCode: Autoregressively Sewing and Generating Garments from Text Guidance【上科大】

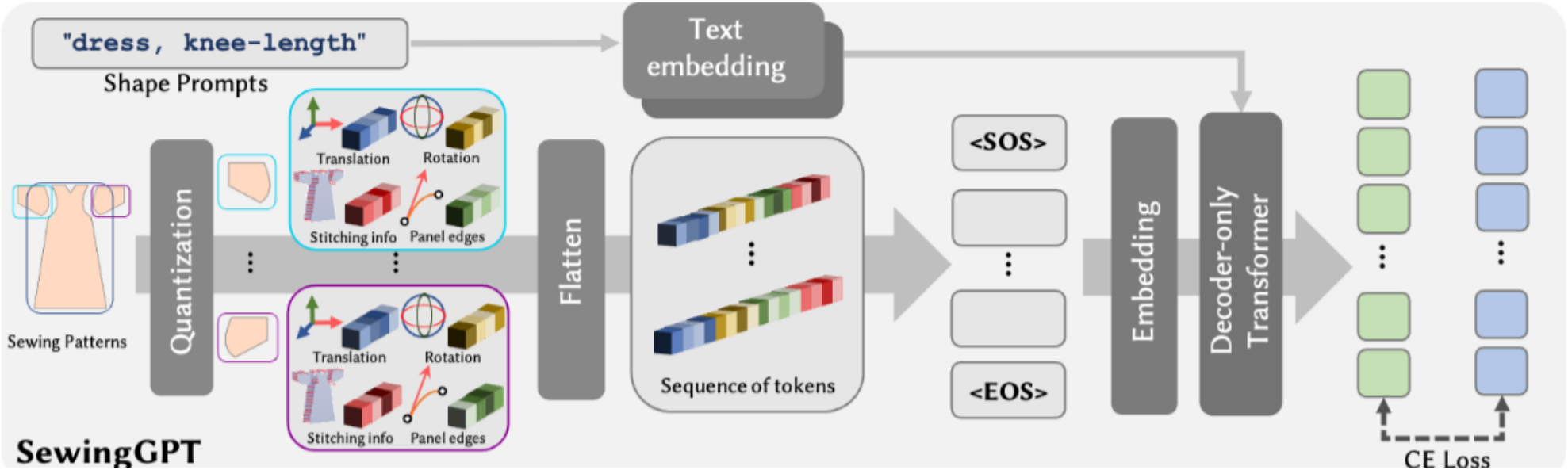
DressCode: Autoregressively Sewing and Generating Garments from Text Guidance

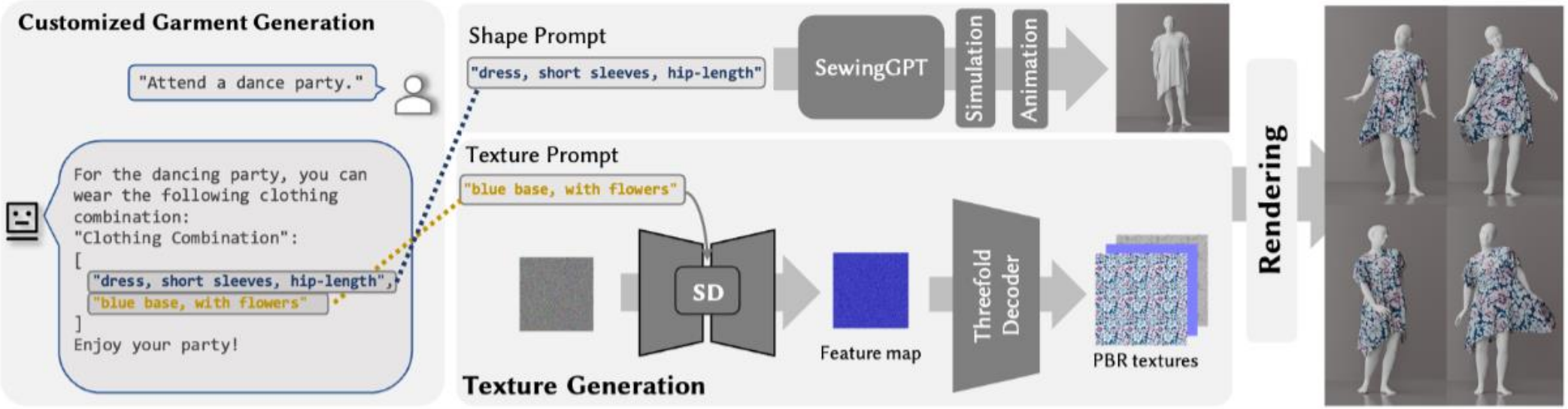
Limitation:缝纫图案数据集限制

- 服装不够复杂
- 类别不够全面

--->从预训练的基础模型SDS中提取知识以提高泛化能力

DressCode: Autoregressively Sewing and Generating Garments from Text Guidance





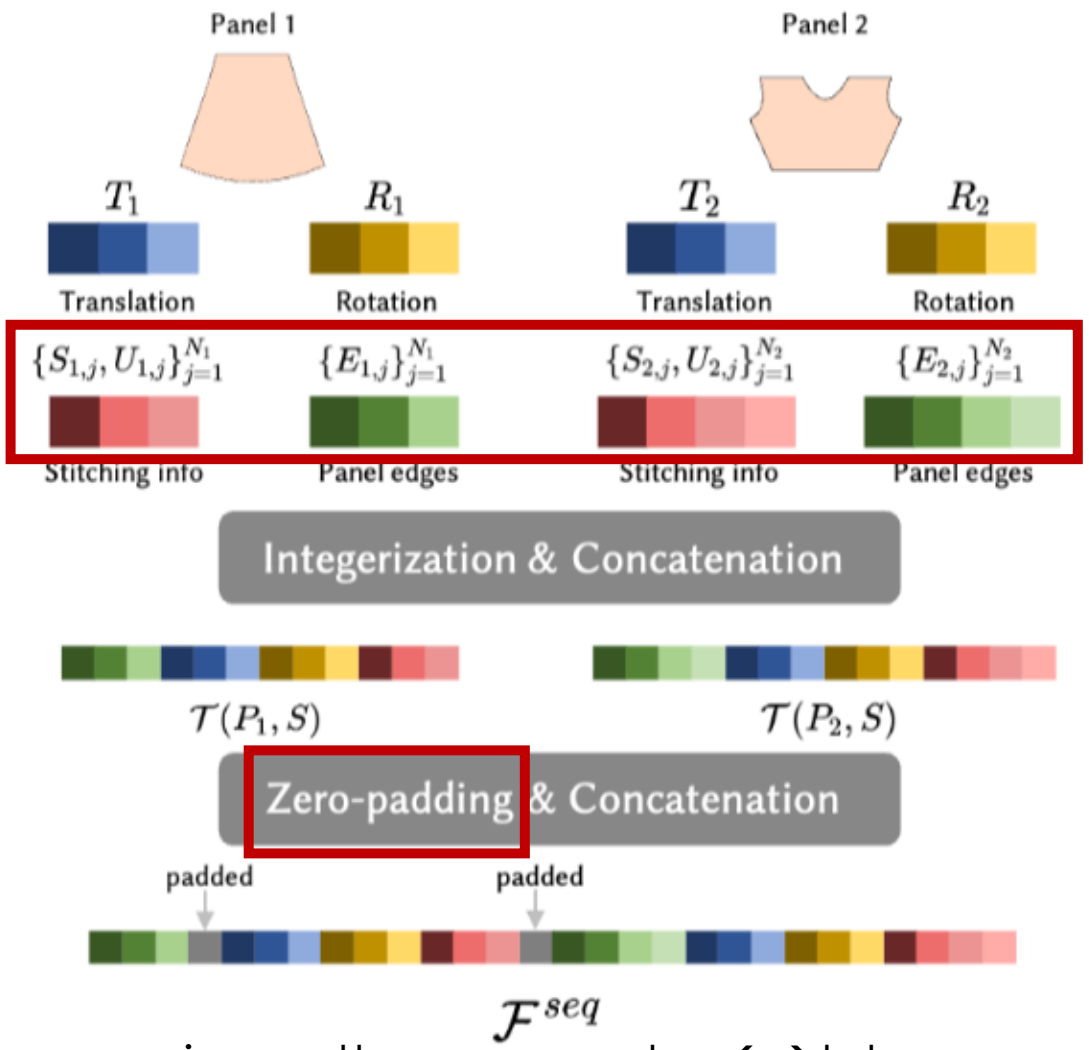
text->3D服装；在已有dataset基础上处理数据

CLIP+LDM+mask Transformer

量化sewing pattern parameters->token（预测）

Diffusion生成PBR纹理

DressCode: Autoregressively Sewing and Generating Garments from Text Guidance



sewing pattern parameters←→token

- 现有dataset：

[Korosteleva and Lee 2022] **Neuraltailor**

- 数据处理：

将所有边缘向量和控制点标准化为标准正态分布

对于每个token，triple embedding

Pos embedding（which panel）

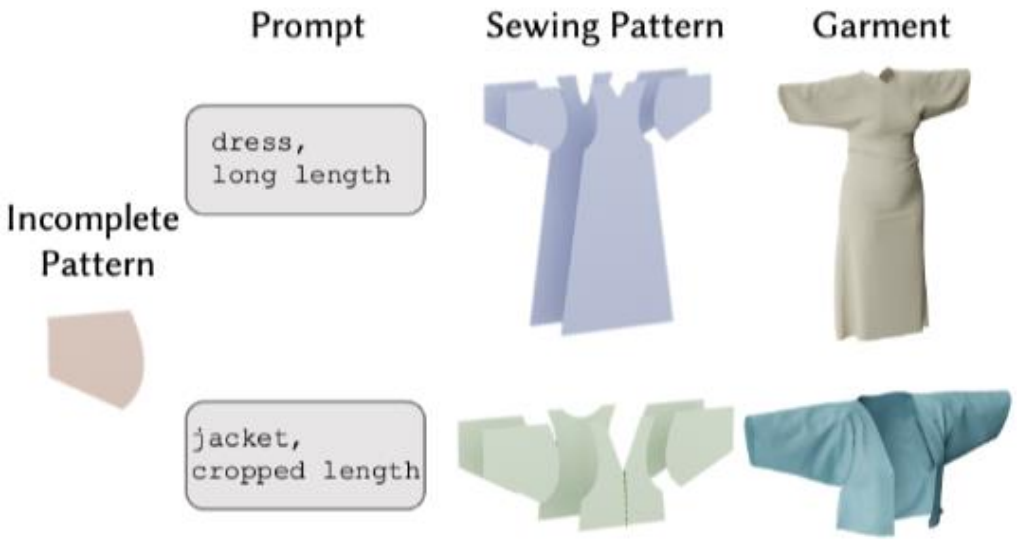
Param embedding（token type classify）

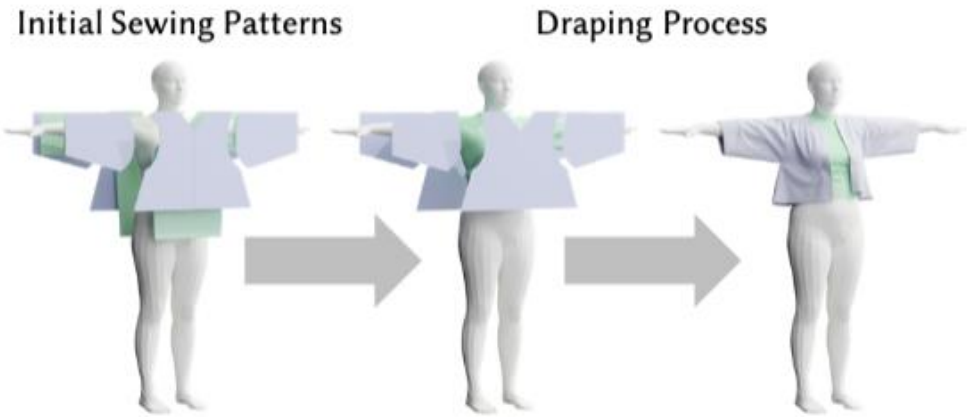
Value embedding（量化后的数值）

- 对于dataset中服装，用**GPT - 4V**生成文本描述，然后**CLIP** embedding
- token预测：训练masked Transformer decoder with **cross attention and text-conditioned** embeddings
- 优化器：Adam

DressCode: Autoregressively Sewing and Generating Garments from Text Guidance

SewingGPT result:





- 仿真平台：

Qualoth in Maya、Marvelous Designer

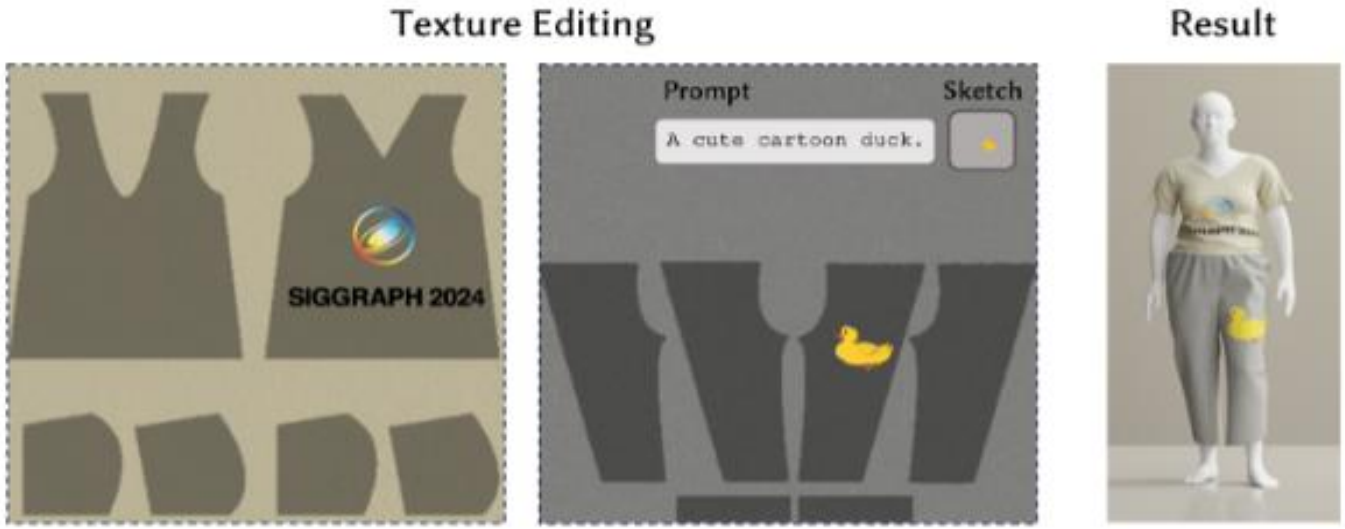
- PBR Diffusion：

微调U-Net denoiser，一个encoder，三个decoder

生成text→漫反射、粗糙度、法线贴图

- 材质编辑：

本方法可以创建结构化UV映射，因此可以加入自定义元素



DressCode: Autoregressively Sewing and Generating Garments from Text Guidance

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

Evaluation

①定性分析：通用3D物品生成baseline

1、NeuralTailor+Surf-D实现NeuralTailor的3D点云输入

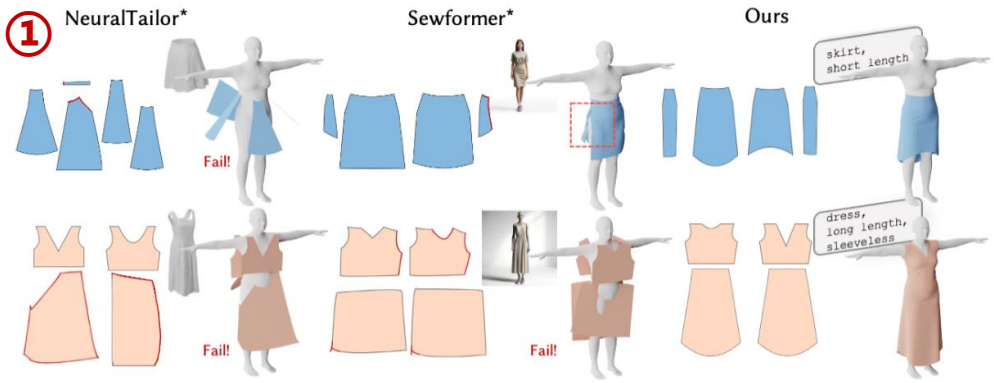
2、Sewformer+DALLE-3实现Sewformer的图片输入

②定量分析：通用3D物品生成baseline

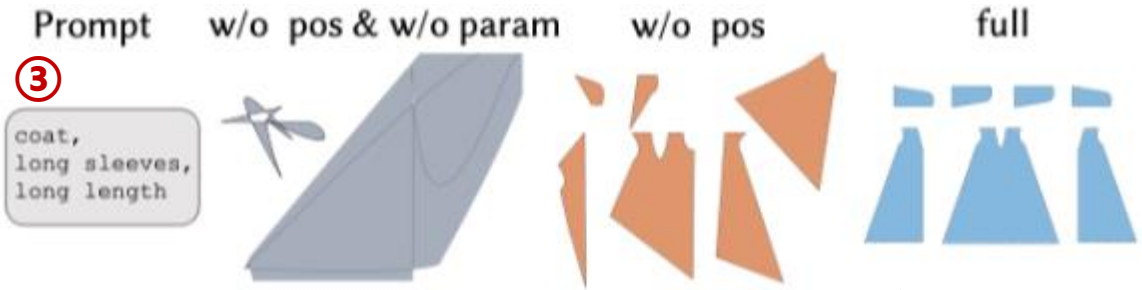
1、Wonder3D+DALLE-3实现Wonder3D的图片输入

2、RichDreamer

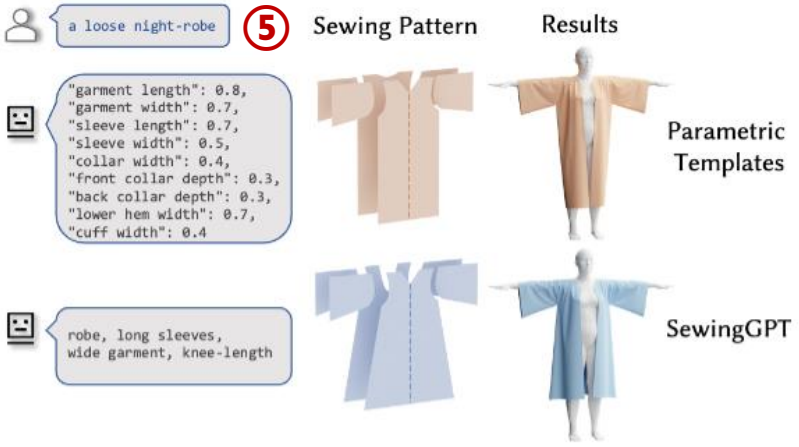
③消融实验：针对embedding方式



②	Wonder3D*	RichDreamer	Ours
CLIP score ↑	0.302	0.324	**0.327**
Runtime ↓	~ 4 mins	~ 4 hours	**~ 3 mins**
PBR Texture	✗	✓	✓
Texture Editing	✗	✗	✓
Draping	✗	✗	✓



④用户研究：定量分析的延申。30人参与，20件服装。



⑤定性分析：

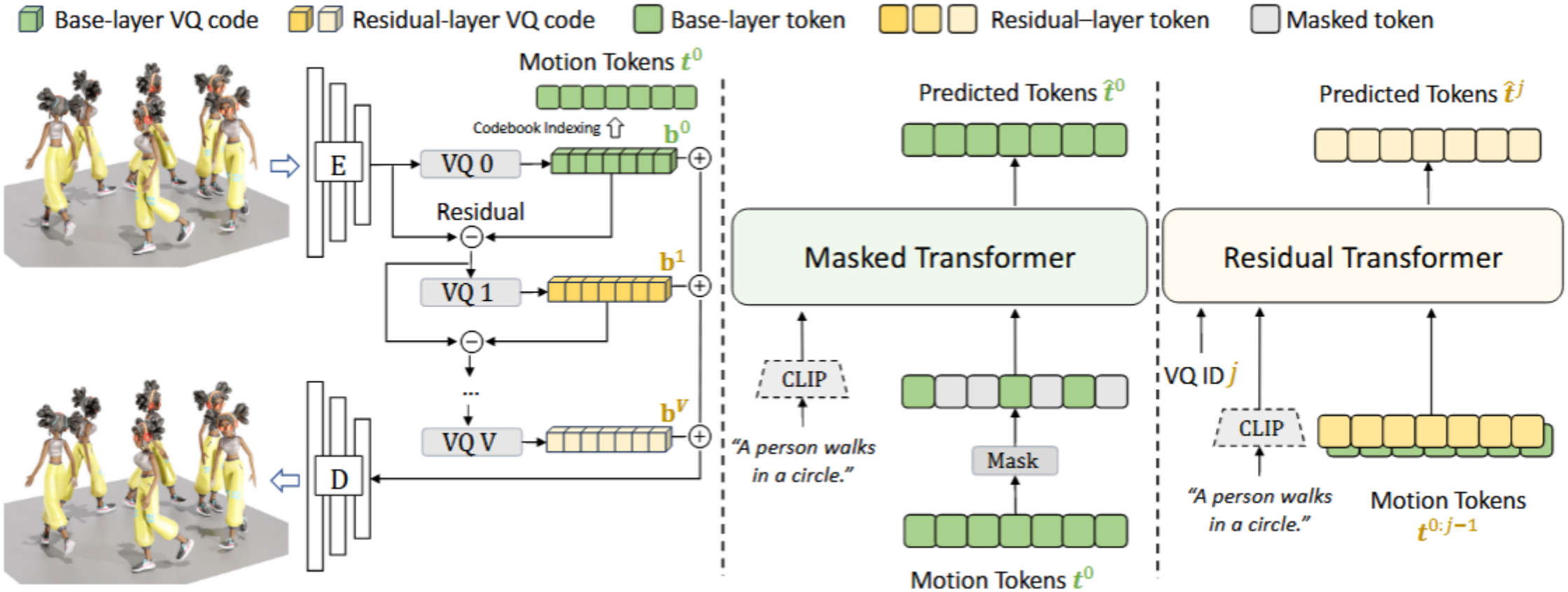
1、未见text prompt

2、与参数化模板比较

What if…?

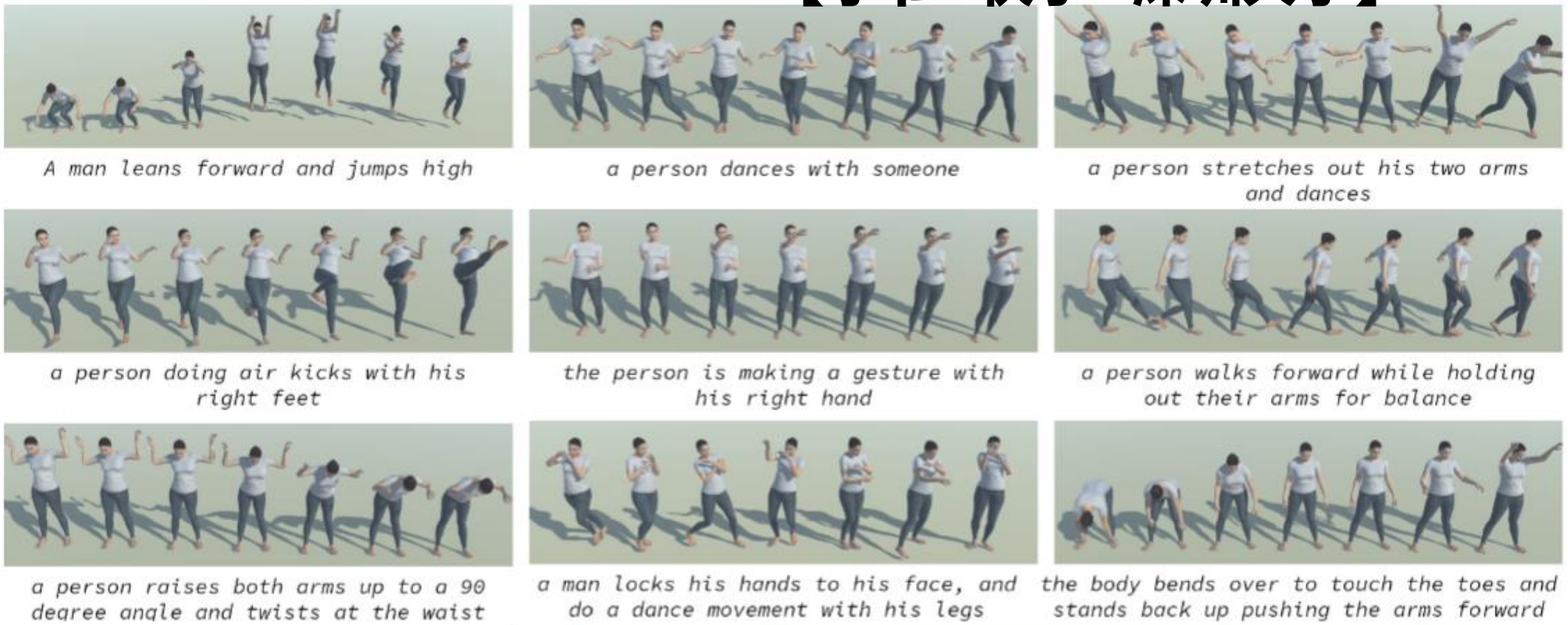
如果某个子领域里已经有人吃过螃蟹了，怎么办?

MoMask: Generative Masked Modeling of 3D Human Motions 【Alberta】

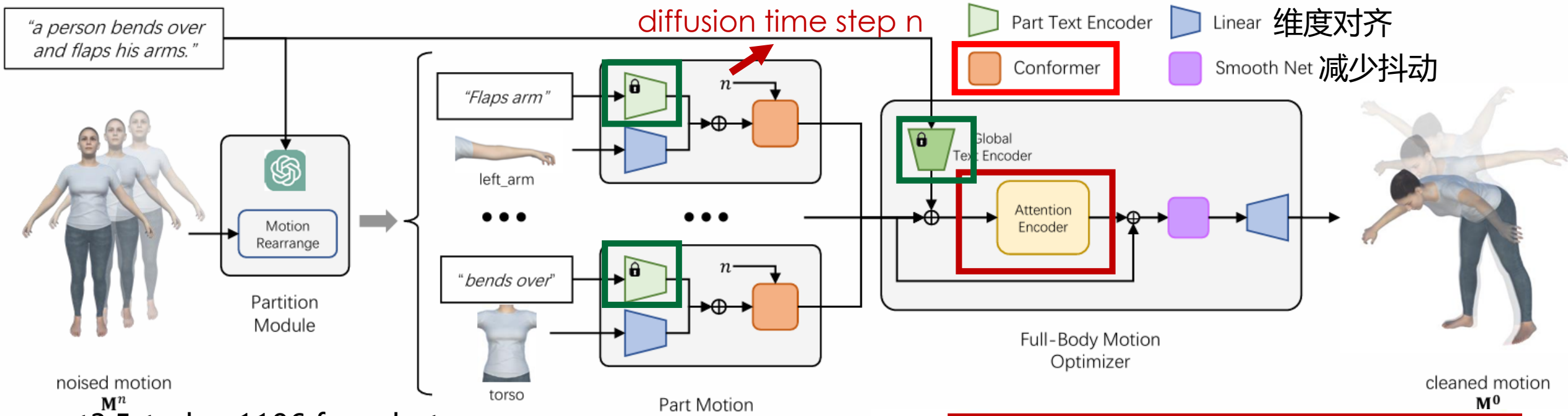


用**CLIP** embedding text作**mask Transformer**隐含条件，**预测**token encoder-decoder模式

LGTM: Local-to-Global Text-Driven Human Motion Diffusion Model 【字节+快手+深圳大学】



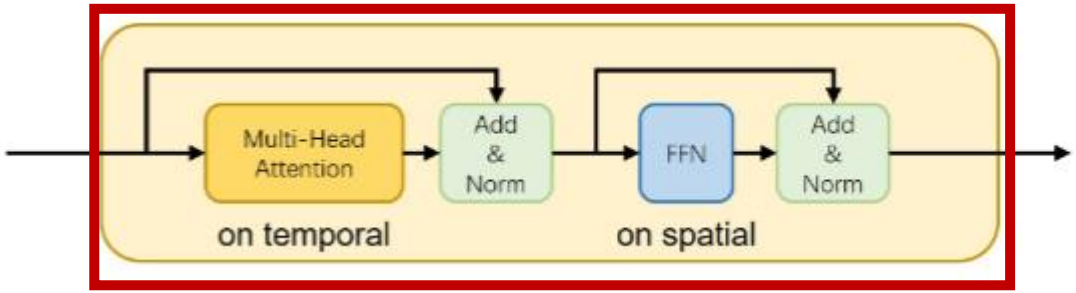
LGTM: Local-to-Global Text-Driven Human Motion Diffusion Model



gpt3.5-turbo-1106 few-shot

Part name	Part description
head	dose nothing
left arm	dose nothing
right arm	waves hand
torso	slightly bends down
left leg	takes a few steps forward
right leg	takes a few steps forward

$$\mathbf{M}_{\text{head}} = [\mathbf{p}_{\text{head}}, \mathbf{r}_{\text{head}}, \mathbf{v}_{\text{head}}] \in \mathbb{R}^{F\times 24}$$
$$\mathbf{M}_{\text{left\_arm}} = [\mathbf{p}_{\text{left\_arm}}, \mathbf{r}_{\text{left\_arm}}, \mathbf{v}_{\text{left\_arm}}] \in \mathbb{R}^{F\times 48}$$
$$\mathbf{M}_{\text{right\_arm}} = [\mathbf{p}_{\text{right\_arm}}, \mathbf{r}_{\text{right\_arm}}, \mathbf{v}_{\text{right\_arm}}] \in \mathbb{R}^{F\times 48}$$
$$\mathbf{M}_{\text{torso}} = [\mathbf{p}_{\text{torso}}, \mathbf{r}_{\text{torso}}, \mathbf{v}_{\text{torso}}, \dot{\mathbf{r}}_{\text{root}}, \mathbf{v}_{\text{root}}, h] \in \mathbb{R}^{F\times 43}$$
$$\mathbf{M}_{\text{left\_leg}} = [\mathbf{p}_{\text{left\_leg}}, \mathbf{r}_{\text{left\_leg}}, \mathbf{v}_{\text{left\_leg}}, \mathbf{c}_{\text{left\_leg}}] \in \mathbb{R}^{F\times 50}$$
$$\mathbf{M}_{\text{right\_leg}} = [\mathbf{p}_{\text{right\_leg}}, \mathbf{r}_{\text{right\_leg}}, \mathbf{v}_{\text{right\_leg}}, \mathbf{c}_{\text{right\_leg}}] \in \mathbb{R}^{F\times 50},$$



TMR[仅在运动描述和运动数据上训练的CLIP]+DDIM[加快LDM sampling]+分区[2-stage，分区生成+整体优化]

LGTM: Local-to-Global Text-Driven Human Motion Diffusion Model

- 运动数据表示：HumanML3D（SMPL变种）
- 训练损失：DDIM L2 loss
- 优化器：AdamW

Evaluation

定性分析：对比MDM和MLD

定量分析：指标8个。编码上和生成结果上来说，潜在空间中的运动样本更加分散，从而更容易区分不相似的运动；滑动、渗透和浮动更少

消融实验：1、用CLIP还是TMR；2、transformer还是conformer；3、有没有全身运动优化器

货比三家，启动！

Text to	人群动画	3D服装	Motion-CVPR	Motion-Sig
Introduction/Related Work	分析本领域后，提出之前不使用text-to的方法局限性or已有text-to方法的局限性			
技术路线	DDPM生成	预测+DDPM生成	预测	DDIM生成
Dataset	自制	已有，自己预处理	已有	已有
Text Embedding	CLIP	CLIP	CLIP	TMR
Loss	DDPM loss	CE loss+DDPM loss	自定义L2loss	DDIM loss（L2）
优化器	AdamW	Adam	-	AdamW
实验设置【定性分析，定量指标，消融实验】	√	√	√	√
实验设置【用户研究】	√	√	√	×
训练时长	4 RTX 4090 GPU for 192 hours	a single A6000 GPU for 30 hours	-	8 hours on 3 NVIDIA RTX 4090 GPUs

积木搭建法则

Text to	人群动画	3D服装	Motion-CVPR	Motion-Sig
Introduction/Related Work	分析本领域后，提出之前不使用text-to的方法局限性or已有text-to方法的局限性			
技术路线				
Dataset				
Text Embedding				
Loss	DD			.2)
优化器	AdamW			AdamW
实验设置【定性分析，定量指标，消融实验】	√			
实验设置【用户研究】	√	√	√	×
训练时长	4 RTX 4090 GPU for 192 hours	a single A600 GPU for 30 hou	-	8 hours on 3 NVIDIA RTX 4090 GPUs

两类：预测用mask transformer；生成用DDPM【及它们的变种】

没有dataset可自制然后成为contribution之一

Text encoder用CLIP或本领域调过的专用encoder

Loss和优化器用基础模型原有的

定性分析，定量指标，消融实验，用户研究

4090能训

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

谢谢观看
敬请各位批评指正