




Controllable Human-Object Interaction Synthesis



ECCV 2024: European Conference on Computer Vision
(CCF B会, 计算机视觉三大顶会之一)

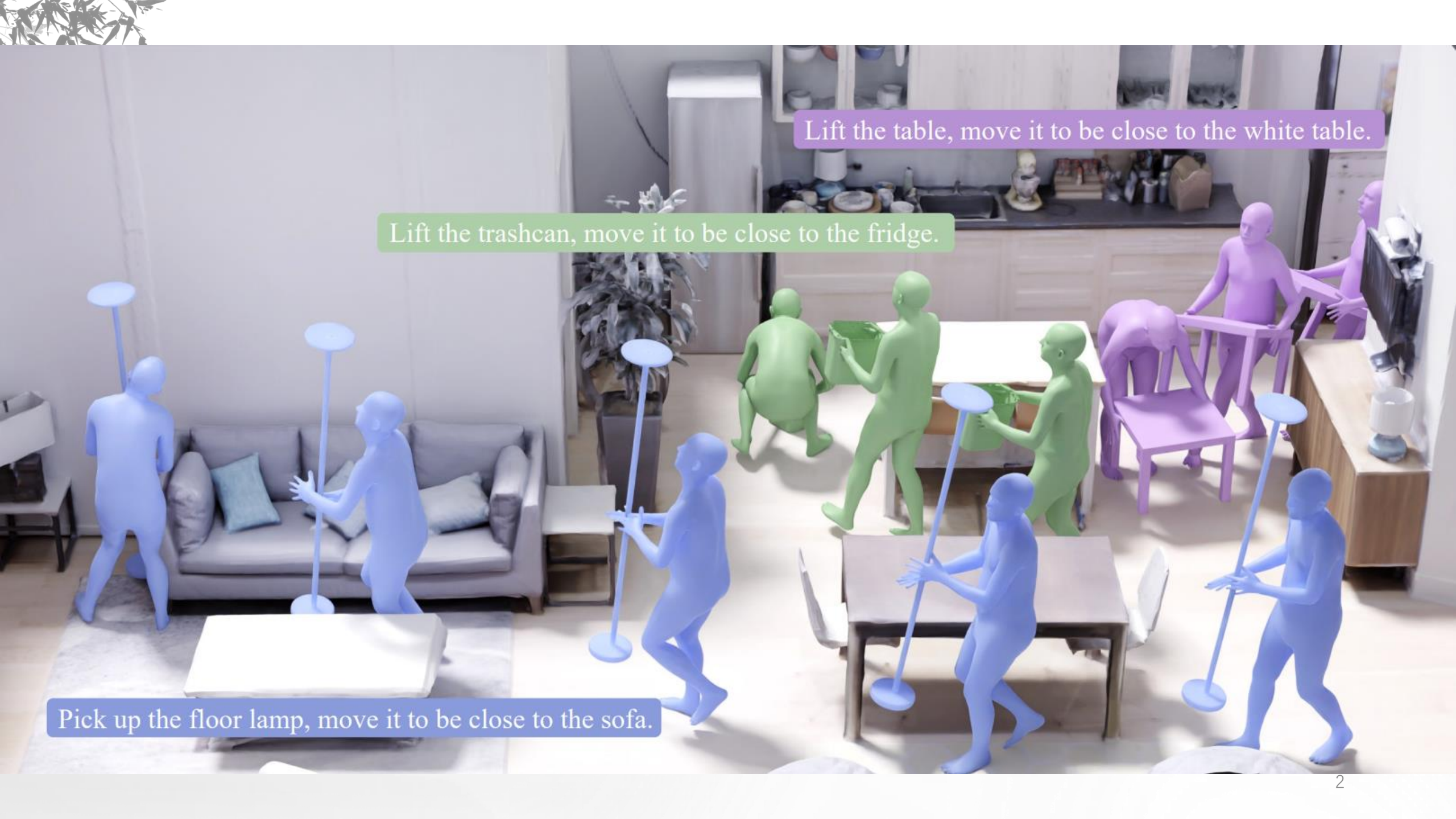


Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, C. Karen Liu
Stanford University, FAIR, Meta

分享者: 陈宇婷

领域: 运动生成

日期: 2024.10.14



Pick up the floor lamp, move it to be close to the sofa.

Lift the trashcan, move it to be close to the fridge.

Lift the table, move it to be close to the white table.

研究团队介绍



李佳蔓
斯坦福大学的博士生
曾于Meta的FAIR实习



Alexander Clegg
研究工程师
FAIR的SIRO项目组



Roozbeh Mottaghi
Senior Scientist Manager
FAIR, 华盛顿大学



吴佳俊
斯坦福大学助理教授
物理场景理解



Xavier Puig
Research Scientist
FAIR



Karen Liu
斯坦福大学教授
The Movement Lab

在3D环境中合成人类行为对于计算机图形学、具身人工智能和机器人技术的各种应用至关重要

- **动态交互的需求**：大部分人-场景交互工作只关注静态物体，忽视了高度动态交互
- **多样物体的操控**：现有动态人-物体交互建模工作仅关注较小的物体或缺乏操控多样物体的能力
- **语言指导的重要性**：操控较大、多样物体的最新工作依赖于过去交互状态的序列或完整的物体运动序列，无法仅从初始状态合成人和物体的运动

从语言和初始状态出发，在3D环境中合成面向多样物体的逼真的人-物体交互

为物体和人类生成真实且同步的运动

- 手与物体保持适当接触
- 物体运动与人类动作保持因果关系

适应杂乱的真实环境

- 真实场景中充满大量物体，限制了可行的运动轨迹空间

文本生成运动

- AMASS 大规模高质量动捕数据集
- BABEL和Human-ML3D进一步引入动作标签和语言描述丰富动捕数据集
- VAE方法有效
- 扩散模型有效

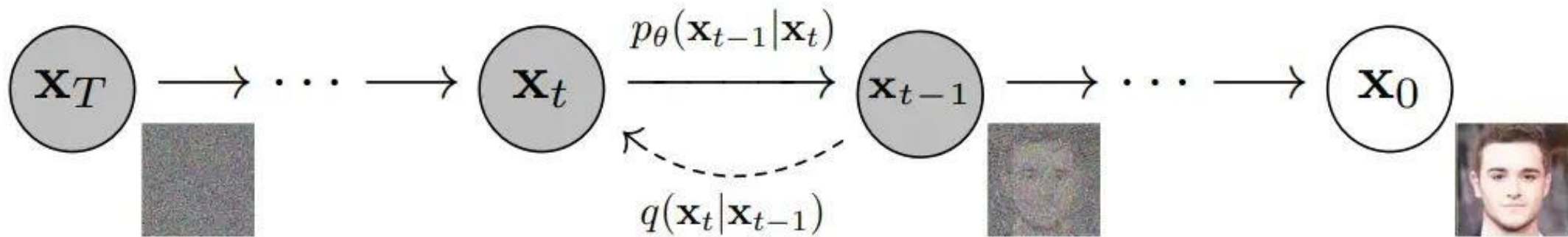
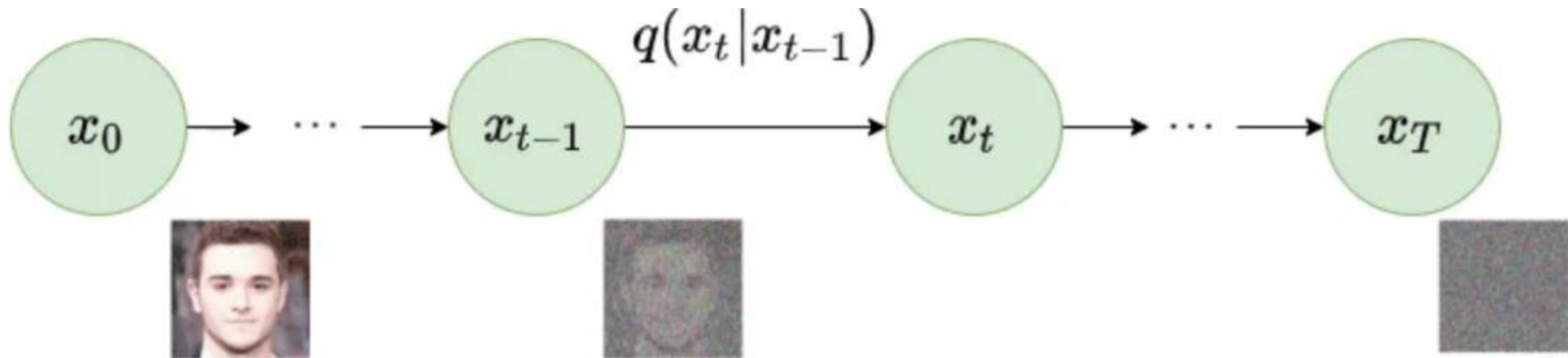
3D场景运动合成

- 配对场景-运动数据和配对物体-运动数据的出现
- 路径规划算法引导人类运动生成
- 强化学习框架训练场景感知策略

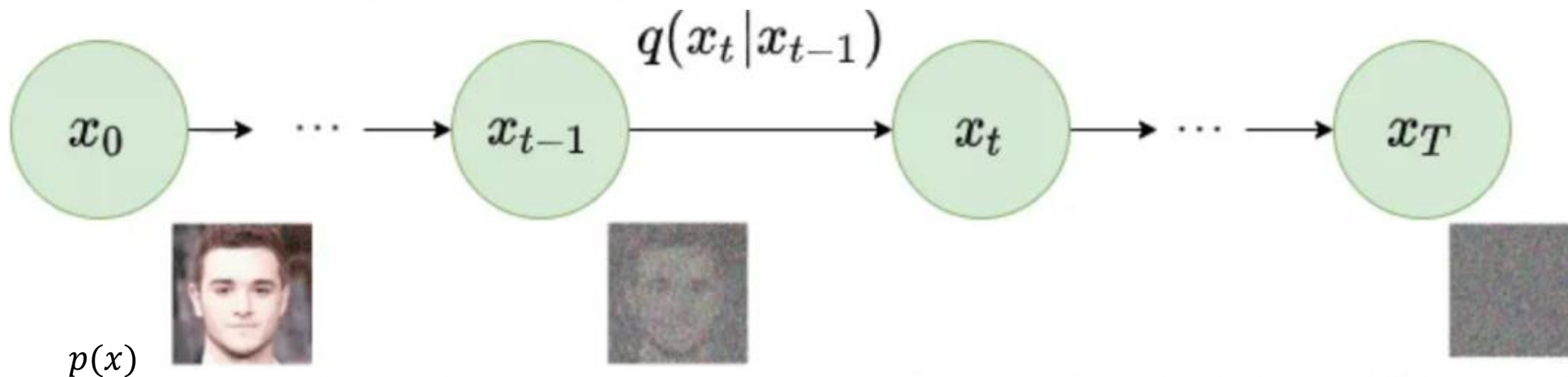
交互合成

- 具有手-物体交互的全身运动数据集
- 关注于手部运动为重点的小型物体
- 训练强化学习策略操控较大物体
- 从过去交互状态序列或物体运动序列预测交互行为

扩散模型



扩散模型 – 正向扩散过程



$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I)$$

概率分布

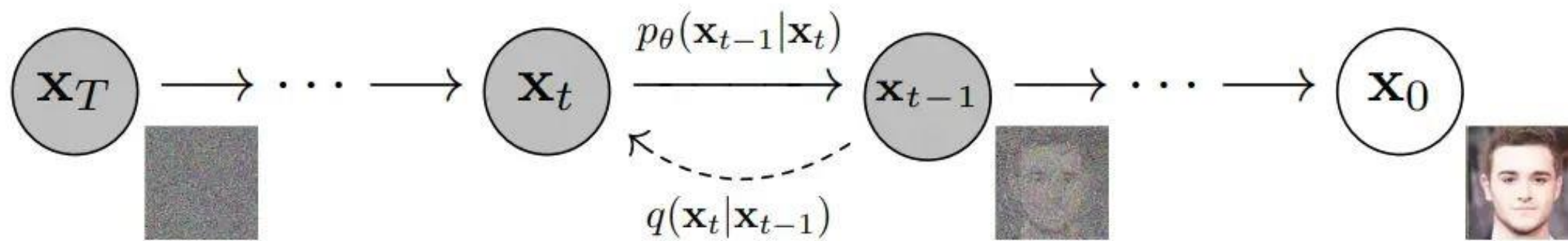
从第t-1步到第t步的状态转移

正态分布

具有均值和协方差矩阵

噪声强度

扩散模型 – 反向扩散过程

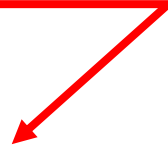


$$\underline{p_\theta(x_{t-1}|x_t, c)} = \underline{N(x_{t-1}; \mu_\theta(x_t, t, c), \Sigma_t)}$$

均值函数



条件概率



给定条件 x_t 和 c

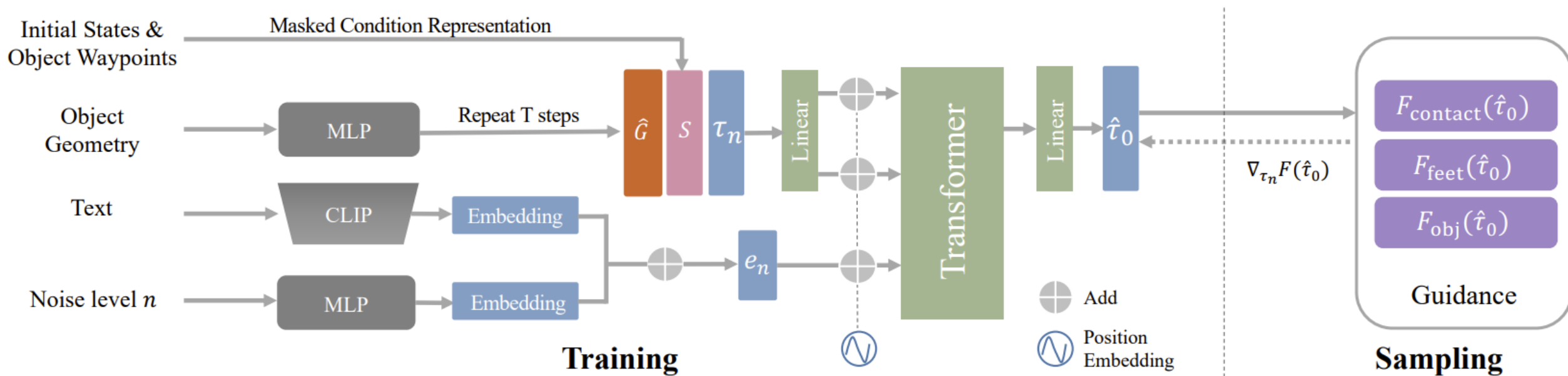
正态分布



具有均值和协方差矩阵

研究方法

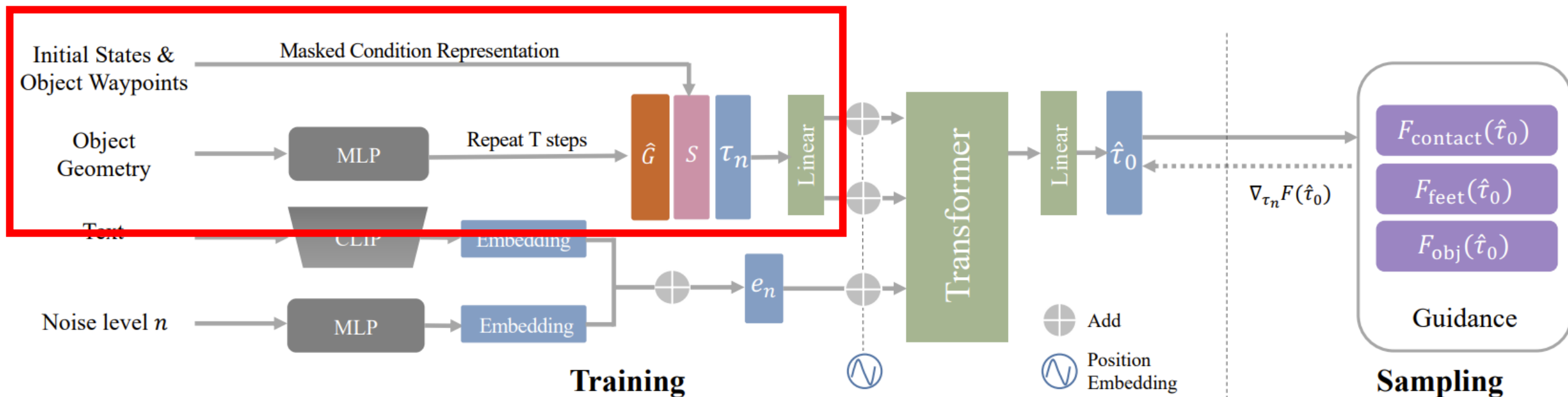
- 目标：基于语言描述、物体几何、初始物体和人体状态、稀疏物体路径点，**生成同步的物体和人类运动**
- 挑战：① 建模同步物体和人体运动的复杂性；② 确保人和物体之间接触的真实性
- 方法：条件扩散模型+引导函数



研究方法 – 数据表示

物体和人体运动表示

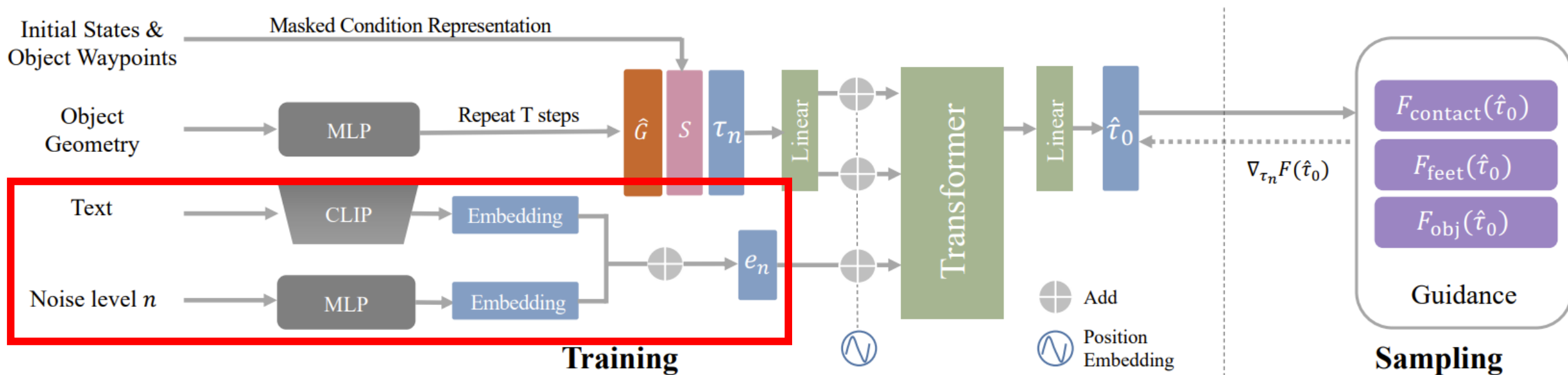
- 物体运动: $O \in \mathbb{R}^{T \times 12}$, 每一行包含一个时间步长下的物体状态信息



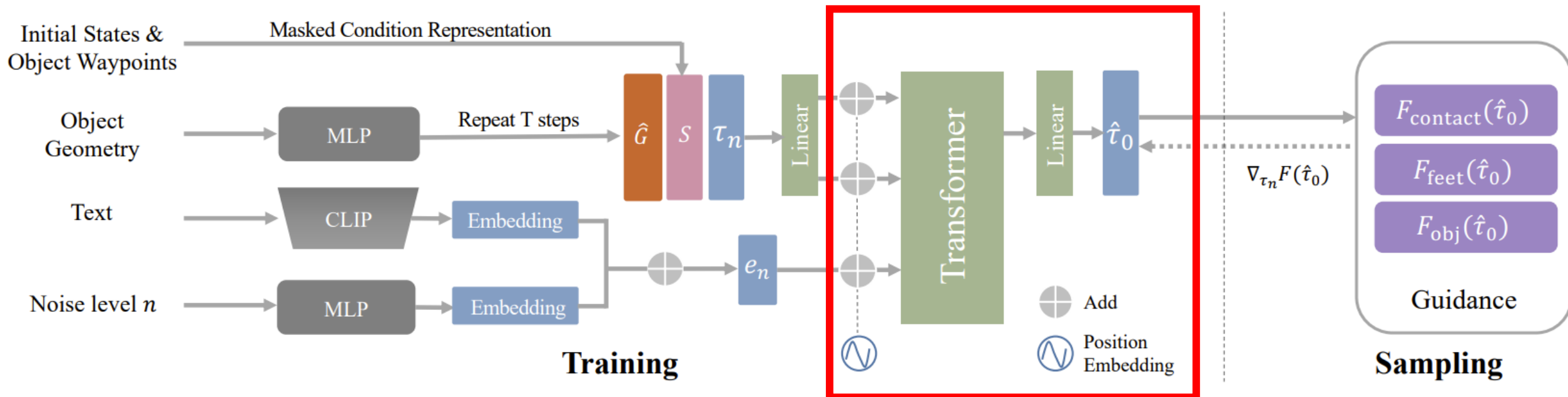
物体表示

- MLP降维: $G \in \mathbb{R}^{1024 \times 3} \rightarrow 256\text{维的向量} \rightarrow \hat{G} \in \mathbb{R}^{T \times 256}$
- 掩码运动数据: $S \in \mathbb{R}^{T \times (12+D)}$, 表示初始状态和路径点条件
- 交互数据表示: $\tau = \{X, O, H\}$

研究方法 – 交互合成模型



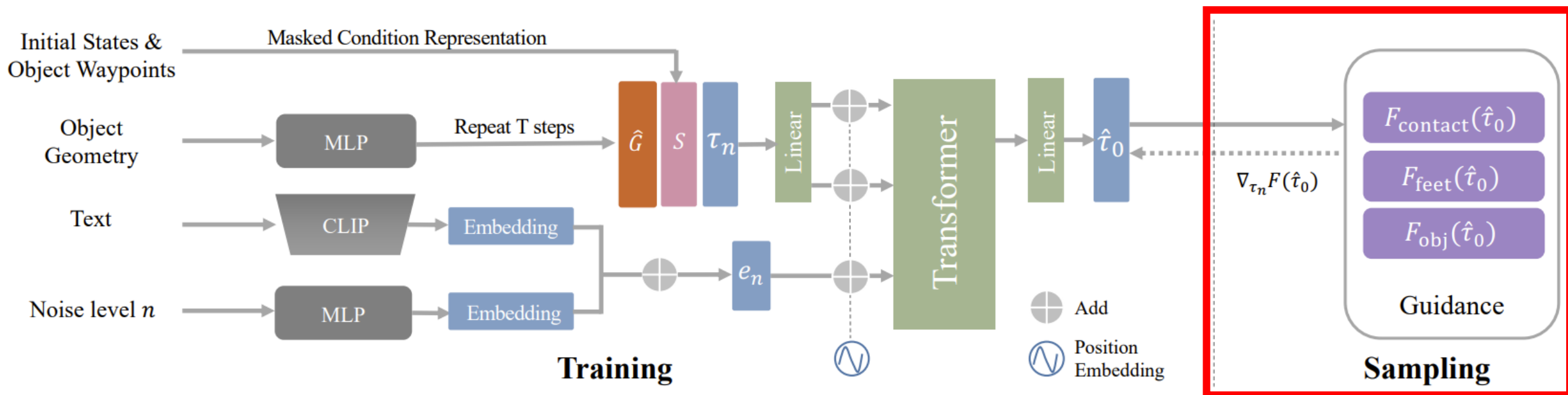
研究方法 – 交互合成模型



$$\mathcal{L} = \mathbb{E}_{\tau_0, n} \|\hat{\tau}_\theta(x_n, n, \mathbf{c}) - \tau_0\|_1$$

$$\mathcal{L}_{\text{obj}} = \sum_{t=1}^T \|\hat{\mathbf{R}}_t \mathbf{K}_{\text{rest}} + \hat{\mathbf{d}}_t - \mathbf{K}_t\|_1 \quad \mathbf{K}_{\text{rest}} \in \mathbb{R}^{100 \times 3}$$

研究方法 – Sampling Guidance



手部-物体接触引导：提高CHOIS模型生成帧的手物接触准确性

$$F_{\text{contact}} = \| \mathbf{M}_l \odot \|\mathbf{J}_l - \mathbf{V}_l\| \|_1 + \| \mathbf{M}_r \odot \|\mathbf{J}_r - \mathbf{V}_r\| \|_1$$

$\mathbf{M}_l, \mathbf{M}_r = (\mathbf{H} > 0.95)$, 根据预测接触标签H来识别左右手可能发生接触的帧

$\mathbf{J}_l, \mathbf{J}_r \in \mathbb{R}^{T \times 3}$, 左右手关节位置, 表示每个时间步的三维坐标

$\mathbf{V}_l, \mathbf{V}_r \in \mathbb{R}^{T \times 3}$, 物体网格上与左右手最近的点, 评估接近程度

\odot , Hadamard积, 逐元素乘积, 应用于掩码与对应的绝对差异

使模型更关注于接触可能性较高的时刻

足部-地面接触引导：重建人体网格可能导致双脚不触地的情况

$$F_{\text{feet}} = ||\min(\mathbf{J}_l^z, \mathbf{J}_r^z) - h||_2$$

$\mathbf{J}_l^z, \mathbf{J}_r^z$, 分别表示左右脚趾关节位置的z分量

测量并最小化脚趾与地面期望接触高度之间的差异

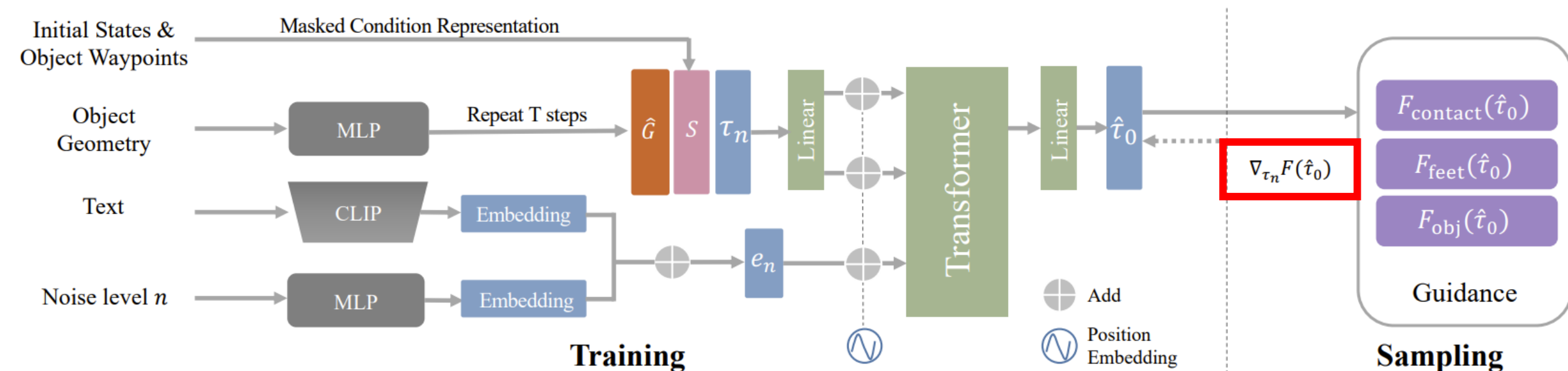
物体地面穿透引导：生成的物体状态可能穿透地面的情况

$$F_{\text{obj}} = ||\min(\mathbf{V}^z, 0)||_1$$

\mathbf{V}^z , 表示物体顶点的 z 坐标集合

$$F_{\text{all}} = \lambda_1 F_{\text{contact}} + \lambda_2 F_{\text{feet}} + \lambda_3 F_{\text{obj}}$$

研究方法 – Sampling Guidance



$$\tilde{\tau}_0 = \hat{\tau}_0 - \alpha \sum_n \nabla_{\tau_n} F(\hat{\tau}_0)$$

调整后的干净数据

扰动强度因子

相对于 τ_n 的梯度

去噪中预测的干净数据

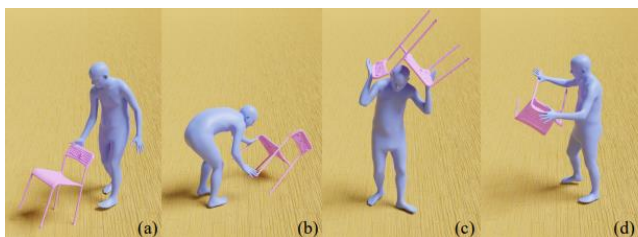
实现效果



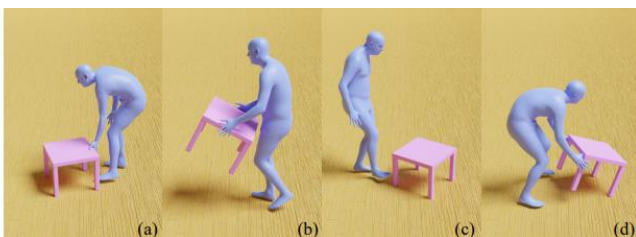
Pick up the floorlamp, move the floorlamp to be close to the sofa.

实验 – 数据集

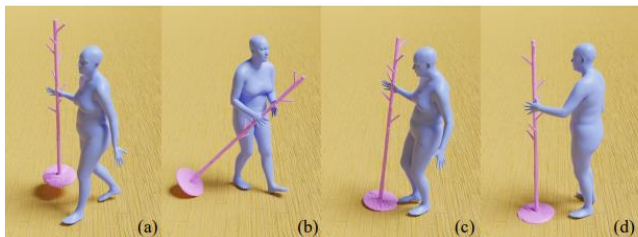
- **FullBodyManipulation数据集** (自产, 2023TOG收录) : 时长10h, 涉及15种不同物体的交互, 包含配对的物体与人体运动数据。
- **3D-FUTURE数据集**: 包含各种家具的3D模型, 选择了其中17种代表不同类型的物体。



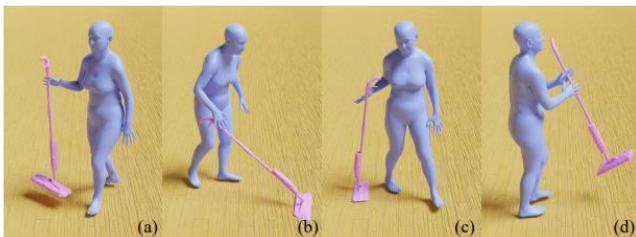
(a) Pull the chair to move. (b) Grab one of the chair's legs and tilt it. (c) Lift the chair over your head, walk and place the chair onto the floor. (d) Lift the chair, flip it upside down and place it on top of the table.



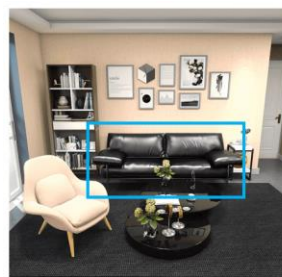
(a) Pull the table to the desired location. (b) Lift and move the table. (c) Kick the table to move across the room. (d) Lift two legs, slide your feet and rotate the table, and lower the table.



(a) Lift and move the clothes stand. (b) Pull and move the clothes stand. (c) Kick the base of the clothes stand to move. (d) Lift and adjust the clothes stand to a different orientation.



(a) Pick up and move it to the desired location. (b) Push and pull to clean the floor. (c) Drag it to the desired location. (d) Lift and swing it, walk.



Realistic Synthetic Scene



Instance Segmentation



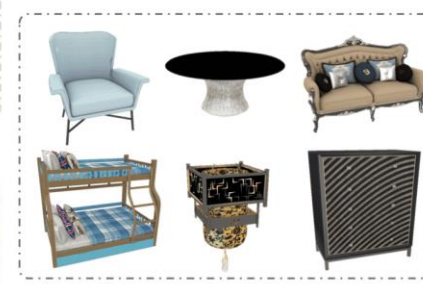
Fine-grained Mesh and Texture

Category: Two-seat Sofa **Theme:** Smooth Net
Material: Smooth Leather **Style:** Modern

Attributes



3D-FUTURE Realistic Rendering



3D-FUTURE renderings with texture

实验 – 评价指标

- **条件匹配指标：** 计算预测和输入物体路径点之间的欧氏距离
- **人体运动质量指标：** 包括足部滑动得分 (FS)、足部高度 (H_{feet})、Fréchet Inception Distance (FID) 和R精度 (R_{prec})
- **交互质量指标：** 评估手物交互的准确性，包括接触和穿透
 - 接触准确性：精度 (C_{prec})、召回率 (C_{rec}) 和F1分数 (C_{F_1}) 指标
 - 穿透得分： $P_{\text{hand}} = \frac{1}{n} \sum_{i=1}^n |\min(d_i, 0)|$
- **真实值 (GT) 差异指标：**
 - 欧氏距离计算：每个关节位置的平均误差 (MPJPE)、根部关节的平移误差 (T_{root}) 和物体位置误差 (T_{obj})
 - Frobenius范数计算：根部关节旋转误差 (O_{root}) 和物体旋转误差 (O_{obj})

实验 - 设置

● Baseline:

- InterDiff (ICCV2023) : 使用前10帧预测人-物交互 (原) ; **接受文本和稀疏物体路径点输入 (调)**
- MDM (ICLR2023) : 基于语言描述生成人体动作 (原) ; **整合物体几何表示和稀疏物体路径点 (调)**
- OMOMO (TOG2023) : 根据提供的物体运动轨迹合成人体动作 (原) ; **线性插值 (调)**

● 消融实验:

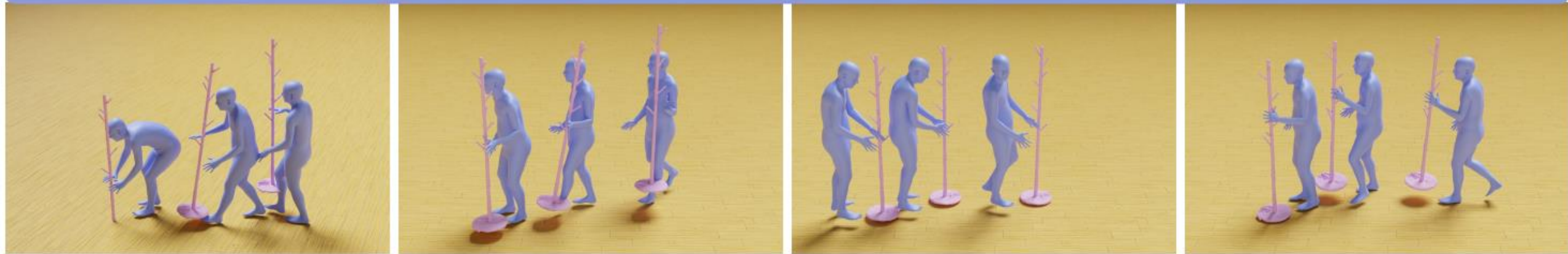
- Pred-OMOMO: 文本生成运动模块与OMOMO结合
- GT-OMOMO: 利用真实物体运动作为OMOMO输入
- CHOIS w/o L_{obj} : 作为条件扩散模型进行训练, 但不包括额外的物体几何损失
- CHOIS w/o F_{all} : 整合了物体几何损失, 但在推理时不使用引导函数

Table 1: Interaction synthesis on the FullBodyManipulation dataset [28].

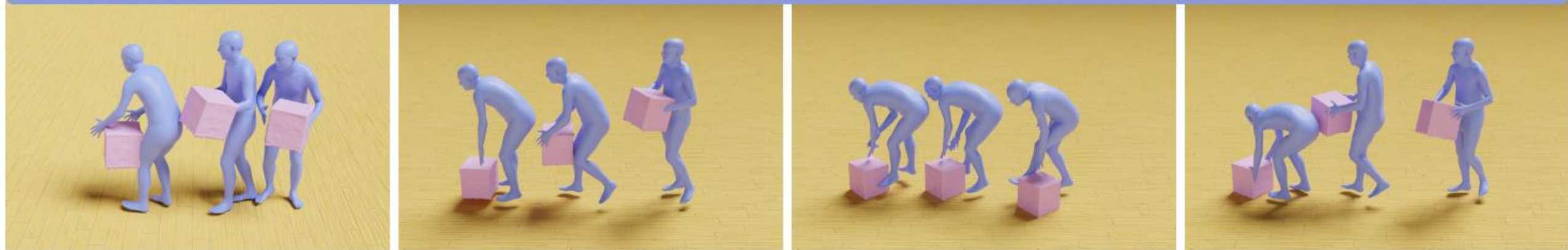
| Method | Condition Matching | | | Human Motion | | | | Interaction | | | | | GT Difference | | | |
|---------------------|--------------------|------------------|---------------------|-----------------------|-----------------|---------------------|------------------|---------------------|--------------------|--------------------|-------|-----------------------|--------------------|-----------------------|----------------------|----------------------|
| | $T_s \downarrow$ | $T_e \downarrow$ | $T_{xy} \downarrow$ | $H_{feet} \downarrow$ | FS \downarrow | $R_{prec} \uparrow$ | $FID \downarrow$ | $C_{prec} \uparrow$ | $C_{rec} \uparrow$ | $C_{F_1} \uparrow$ | $C\%$ | $P_{hand} \downarrow$ | MPJPE \downarrow | $T_{root} \downarrow$ | $T_{obj} \downarrow$ | $O_{obj} \downarrow$ |
| Interdiff [61] | 0.00 | 158.84 | 72.72 | 0.90 | 0.42 | 0.08 | 208.0 | 0.63 | 0.28 | 0.33 | 0.27 | 0.55 | 25.91 | 63.44 | 88.35 | 1.65 |
| MDM [51] | 5.18 | 33.07 | 19.42 | 6.72 | 0.48 | 0.51 | 6.16 | 0.72 | 0.47 | 0.53 | 0.43 | 0.66 | 17.86 | 34.16 | 24.46 | 1.85 |
| Lin-OMOMO [28] | 0.00 | 0.00 | 0.00 | 7.21 | 0.41 | 0.29 | 15.33 | 0.68 | 0.56 | 0.57 | 0.54 | 0.51 | 21.73 | 36.62 | 17.12 | 1.21 |
| Pred-OMOMO [28] | 2.39 | 8.03 | 4.15 | 7.08 | 0.40 | 0.54 | 4.19 | 0.73 | 0.66 | 0.66 | 0.62 | 0.58 | 18.66 | 28.39 | 16.36 | 1.05 |
| GT-OMOMO [28] | 0.00 | 0.00 | 0.00 | 7.10 | 0.41 | 0.48 | 5.69 | 0.77 | 0.66 | 0.67 | 0.59 | 0.55 | 15.82 | 24.75 | 0.00 | 0.00 |
| CHOIS w/o L_{obj} | 5.76 | 14.16 | 8.44 | 6.55 | 0.40 | 0.65 | 3.26 | 0.75 | 0.50 | 0.55 | 0.43 | 0.66 | 14.34 | 21.97 | 15.53 | 0.98 |
| CHOIS w/o F_{all} | 1.75 | 6.61 | 2.69 | 6.64 | 0.38 | 0.65 | 3.58 | 0.78 | 0.49 | 0.55 | 0.41 | 0.65 | 15.23 | 24.13 | 11.51 | 0.99 |
| CHOIS (ours) | 1.71 | 6.31 | 2.87 | 4.20 | 0.35 | 0.64 | 0.69 | 0.80 | 0.64 | 0.67 | 0.54 | 0.59 | 15.30 | 24.43 | 12.53 | 0.99 |

实验 - 结果

Pick up the clothes stand, move it, and put it down.



Lift the box, move the box, and put down the box.



(a) InterDiff

(b) MDM

(c) Lin-OMOMO

(d) CHOIS (ours)

Fig. 3: Qualitative results of the FullBodyManipulation dataset [28].

Table 2: Interaction synthesis on the 3D-FUTURE dataset [12]

| | Condition Matching | | | Human Motion | | | | Interaction | |
|---------------------|--------------------|------------------|---------------------|-----------------------|-----------------|---------------------|------------------|-------------|-----------------------|
| | $T_s \downarrow$ | $T_e \downarrow$ | $T_{xy} \downarrow$ | $H_{feet} \downarrow$ | FS \downarrow | $R_{prec} \uparrow$ | $FID \downarrow$ | $C\%$ | $P_{hand} \downarrow$ |
| InterDiff [61] | 0 | 161.26 | 72.77 | -0.26 | 0.42 | 0.09 | 207.3 | 0.24 | 0.11 |
| MDM [51] | 12.58 | 40.55 | 28.72 | 7.02 | 0.49 | 0.53 | 8.50 | 0.34 | 0.26 |
| Lin-OMOMO [28] | 0 | 0 | 0 | 6.32 | 0.42 | 0.23 | 23.17 | 0.44 | 0.11 |
| Pred-OMOMO [28] | 4.15 | 9.03 | 3.89 | 6.08 | 0.40 | 0.46 | 3.74 | 0.50 | 0.18 |
| CHOIS w/o L_{obj} | 6.70 | 13.73 | 7.99 | 5.68 | 0.41 | 0.66 | 3.26 | 0.36 | 0.30 |
| CHOIS w/o F_{all} | 5.75 | 7.96 | 2.68 | 5.84 | 0.39 | 0.62 | 4.78 | 0.33 | 0.26 |
| CHOIS (ours) | 4.12 | 7.35 | 2.92 | 3.75 | 0.38 | 0.62 | 1.60 | 0.48 | 0.15 |

Table 3: Ablation study on the FullBodyManipulation dataset [28]. We measure the effect of different guidance terms in the human and object motion generation.

| Method | Condition Matching | | | Human Motion | | | | Interaction | | | | | GT Difference | | | |
|-------------------------|--------------------|------------------|---------------------|-----------------------|-----------------|---------------------|------------------|---------------------|--------------------|-------------------|-------------|-----------------------|--------------------|-----------------------|----------------------|----------------------|
| | $T_s \downarrow$ | $T_e \downarrow$ | $T_{xy} \downarrow$ | $H_{feet} \downarrow$ | FS \downarrow | $R_{prec} \uparrow$ | $FID \downarrow$ | $C_{prec} \uparrow$ | $C_{rec} \uparrow$ | $C_{F1} \uparrow$ | $C\%$ | $P_{hand} \downarrow$ | MPJPE \downarrow | $T_{root} \downarrow$ | $T_{obj} \downarrow$ | $O_{obj} \downarrow$ |
| CHOIS w/o $F_{contact}$ | 1.70 | 6.42 | 2.70 | 3.93 | 0.32 | 0.66 | 0.74 | 0.78 | 0.49 | 0.55 | 0.41 | 0.65 | 15.41 | 23.63 | 11.44 | 0.99 |
| CHOIS w/o F_{feet} | 1.72 | 6.34 | 2.90 | 6.65 | 0.39 | 0.63 | 3.76 | 0.81 | 0.64 | 0.66 | 0.54 | 0.58 | 15.44 | 25.09 | 13.31 | 0.99 |
| CHOIS w/o F_{all} | 1.75 | 6.61 | 2.69 | 6.64 | 0.38 | 0.65 | 3.58 | 0.78 | 0.49 | 0.55 | 0.41 | 0.65 | 15.23 | 24.13 | 11.51 | 0.99 |
| CHOIS (ours) | 1.71 | 6.31 | 2.87 | 4.20 | 0.35 | 0.64 | 0.69 | 0.80 | 0.64 | 0.67 | 0.54 | 0.59 | 15.30 | 24.43 | 12.53 | 0.99 |

实验 - 结果

Hold and turn the clothes stand around to a different orientation.

OMOMO



CHOIS w/o L



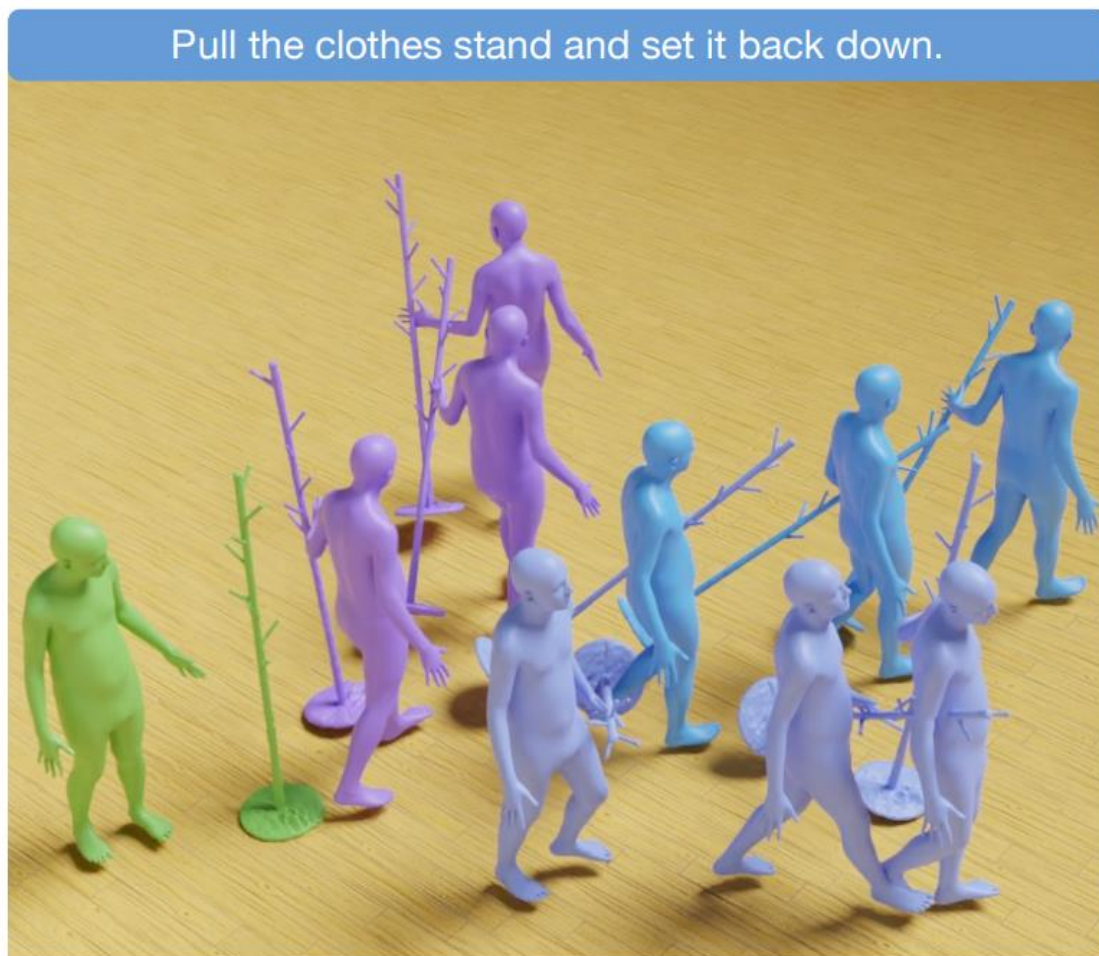
CHOIS w/o F



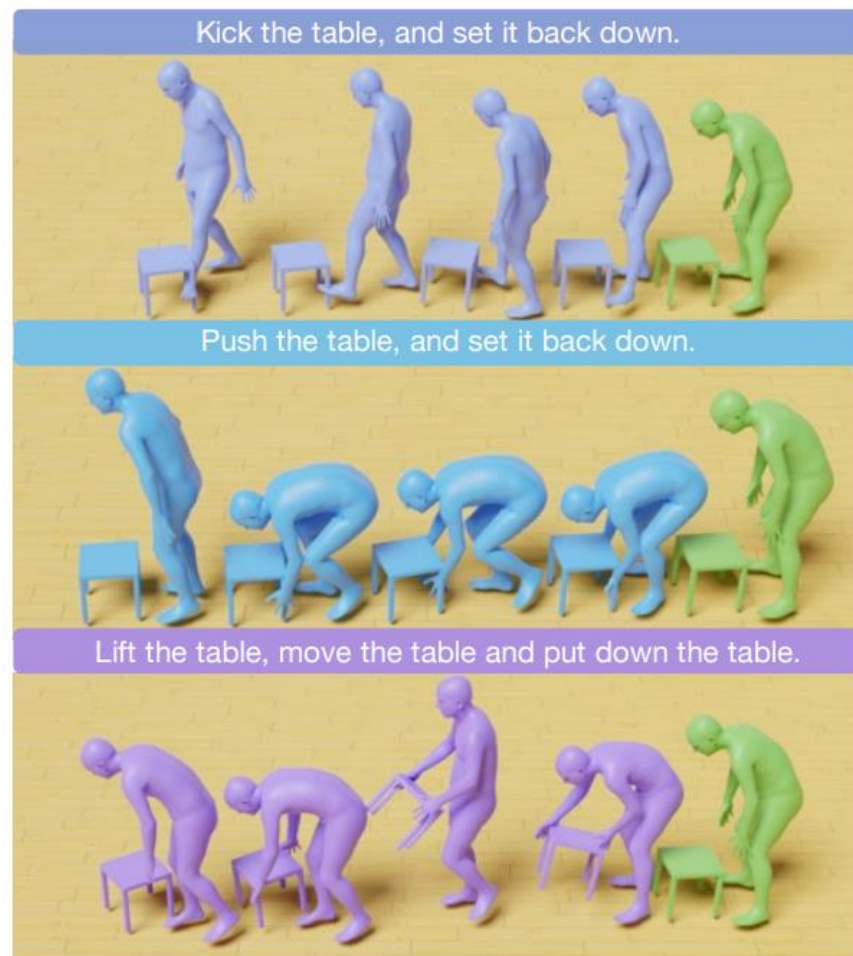
CHOIS



实验 - 结果



(a)



(b)

Fig. 6: Results of interaction synthesis using the same text but different waypoints (a) and using the same waypoints but different text (b). The initial state is in green.

- **CHOIS 使用条件扩散模型**，当给定人 / 物位置的初始状态以及所需操作的语言描述之后，CHOIS 就会据此生成一系列动作，在训练时引入物体几何损失作为**额外监督**，以及设计在采样过程中**加强接触约束**的指导项，最终完成任务要求的**交互效果**
- **贡献：**
 - ① 发现**语言和物体路径点的结合**为人-物交互合成提供了丰富信息
 - ② 通过语言和物体的稀疏路径点引导人-物体交互合成
 - ③ 可以在现有数据集上合成逼真的交互，并推广到新物体的数据集应用上



欢迎探讨

分享者：陈宇婷

领域：运动生成

日期：2024.10.14