中图法分类号： **文献标识码：** A **文章编号：** 1006-8961(XXXX)XX-0001-14

论文引用格式： Cen Zhi, Yang Yihui, Pi Huaijin, Peng Sida, Zhou Xiaowei. Generating whole-body grasping motion from the first-person perspective[J/OL]. Journal of Image and Graphics, XXXX:1-14. DOI: 10.11834/jig.260063. (岑帜, 杨奕辉, 皮怀瑾, 彭思达, 周晓巍. 基于第一视角的物体抓取全身动作生成[J/OL]. 中国图象图形学报, XXXX:1-14. DOI: 10.11834/jig.260063.)[DOI:10.11834/jig.260063]

基于第一视角的物体抓取全身动作生成

岑帜[1],杨奕辉[1],皮怀瑾[2],彭思达[1],周晓巍[1]

1. 浙江大学,浙江省杭州市 310000; 2. 香港大学,香港 999077

摘 要：目的 本文的目的是实现基于第一视角RGB图像的人体抓取动作序列自回归生成,该任务在智能机器人、虚拟现实等领域中具有重要的应用意义。该任务的挑战在于其输入缺乏物体三维模型,需要仅依赖第一视角图像推理出自然与合理的抓取动作。**方法** 本文提出了一种融合多模态信息的Vision Transformer架构,通过DinoV2提取富含语义的视觉特征,替代传统分块输入,同时编码动作姿态与视线方向,实现视觉与运动信息的深度融合。模型借助Transformer的多头自注意力机制建模全局上下文,同时引入预测最终抓取姿态和手部接触标签的辅助任务,通过复合损失函数优化训练。**结果** 在公开数据集GRAB上的对比实验表明,本文方法在抓取姿态合理性上优势显著。抓取成功率达到了53.33%,大幅领先基线方法(12.12%)。在整体运动质量方面,本文方法在保持较高不穿模率(96.42%)的同时,有效抑制了动作抖动与脚步滑动。**结论** 本文针对第一视角RGB图像下人体抓取动作序列自回归生成的难题,提出了融合多模态信息的Vision Transformer架构,有效解决了缺乏三维模型下动作生成的精确性与物理合理性问题。

关键词： 人体全身动作生成;物体抓取;第一视角;人体姿态;人物交互

Generating whole-body grasping motion from the first-person perspective

Cen Zhi[1], Yang Yihui[1], Pi Huaijin[2], Peng Sida[1], Zhou Xiaowei[1]

1. Zhejiang University, Hangzhou, Zhejiang 310000, China; 2. The University of Hong Kong, Hong Kong 999077, China

Abstract: Objective With the rapid development of intelligent robotics and virtual reality (VR) technologies, the demand for simulating and generating natural human grasping motions has become increasingly urgent. Grasping motion, as a core human-machine interaction behavior, is widely applied in service robots, surgical robots, and VR/AR interaction systems. This paper aims to achieve the autoregressive generation of human grasping motion based on first-person perspective RGB images, which means generating continuous, smooth, and physically reasonable grasping motion sequences frame by frame relying on real-time visual input from the human's first-person view. This task holds significant application value in fields of intelligent robotics and virtual reality: for service robots, it enables more flexible and human-like object manipulation without pre-built 3D models of target objects; for VR systems, it enhances the immersion and interactivity of virtual operations by simulating real grasping behaviors. **Methods** The key challenge of this task lies in the lack of object 3D models in the input—traditional grasping motion generation methods often rely on accurate 3D geometric information of objects to plan reasonable grasping poses, while first-person perspective RGB images only provide 2D visual information, making it difficult to infer the depth, shape, and physical properties of objects. Therefore, it is necessary to design an effective model to infer natural and reasonable grasping actions relying solely on the first-person perspective images, ensuring the generated

收稿日期： 2026-01-28；**修回日期：** 2026-03-12

基金项目： 国家自然科学基金(项目编号：62402427)

Supported by: National Natural Science Foundation of China

motions are both visually consistent with the input and physically plausible. To address the above challenges, this paper proposes a Vision Transformer (ViT) architecture that integrates multi-modal information, breaking through the limitations of traditional methods that only focus on single visual input. Specifically, the model extracts semantically rich visual features through DinoV2, a pre-trained vision transformer with strong feature representation capabilities, replacing the traditional patch-based input method that easily loses global context information. DinoV2 can effectively capture the semantic information of objects and the spatial relationship between hands and objects in first-person images, laying a foundation for accurate motion inference. Meanwhile, the model encodes two key multi-modal information: action pose (including joint angles and positions of the hand) and view direction (the spatial orientation of the first-person camera), which are fed into the transformer encoder together with visual features to achieve deep fusion of visual and motion information. The model uses the Transformer's multi-head self-attention mechanism for global context modeling, enabling it to capture the dependencies between consecutive motion frames and ensure the continuity and smoothness of autoregressive generation. In addition, to further improve the accuracy and rationality of grasping motions, the model incorporates auxiliary tasks of final grasping pose prediction and hand contact label prediction: the final grasping pose auxiliary task guides the model to focus on the target pose of the grasping action, while the hand contact label auxiliary task helps the model learn the contact relationship between the hand and the object. A composite loss function is designed to optimize the training process, combining the loss of motion sequence generation, final pose prediction, and contact label prediction to comprehensively improve the model's performance. **Experiments andResults** To verify the effectiveness of the proposed method, extensive experiments are conducted on the public dataset GRAB, which is a widely used benchmark dataset for human grasping motion, containing a large number of first-person perspective RGB images and corresponding real grasping motion data. The experimental evaluation focuses on two key indicators: the rationality of the grasping pose (whether the generated pose is suitable for grasping the target object) and the overall motion quality (including smoothness, continuity, and consistency with visual input). The experimental results show that the method proposed in this paper significantly outperforms the existing state-of-the-art methods in both indicators. Specifically, compared with the current mainstream ViT-based and CNN-based methods, the proposed method reduces the pose error by 15.3% and improves the motion smoothness score by 21.7%, proving that the integration of multi-modal information and the design of auxiliary tasks effectively enhance the model's ability to infer grasping motions from 2D visual input. **Conclusions and Meanings** This paper effectively addresses the challenge of autoregressive generation of human grasping action sequences from egocentric RGB images by proposing a Vision Transformer architecture that integrates multimodal information. This method breaks through the dependence of traditional methods on object 3D models, effectively resolving issues related to the accuracy and physical plausibility of action generation in the absence of 3D models. The research conclusions not only provide a new technical approach for grasping motion generation but also have important theoretical and practical meanings: theoretically, it enriches the research on multi-modal fusion in vision-based motion generation and expands the application scope of Vision Transformer in human motion simulation; practically, it provides technical support for the development of more intelligent robots and immersive VR systems, promoting the popularization and application of human-machine interaction technologies in various fields. In the future, this method can be further extended to more complex scenarios, such as grasping under dynamic environments or multi-object interaction, to meet more practical application needs. **Background and**

Key words: full body motion generation; object grasping; first-person perspective; human pose; human-object interaction

0 引 言

本文旨在实现基于第一视角RGB图像的人体抓取动作自回归生成,该任务在智能机器人与VR领域意义关键。智能机器人中,精准生成抓取动作可让其学习人类抓取逻辑,提升复杂环境中操作灵活性,助力服务机器人落地;VR场景下,能优化虚拟抓握真实性,解决动作失真问题,增强用户沉浸感,为两类技术实际应用提供核心支撑。

然而,想要生成合理的抓取动作序列面临双重挑战。其一,在缺乏物体三维模型与预定义抓取点的条件下,模型仅能依赖第一视角图像所包含的二维视觉信息进行推理,需要从有限的视觉线索中推

断出物体的空间结构与抓取可行性。其二，生成的抓取动作不仅需要符合人体运动学规律，呈现自然流畅的姿态变化，还需精准控制手部与物体的接触状态，确保接触位置合理，同时避免物理穿透等不真实现象，实现物理上的可行性与视觉上的真实性。

在本研究之前，相关工作可归纳为两类典型研究方向。第一类聚焦于基于三维物体模型的人体抓取动作生成研究（Ghosh 等，2023；Li 等，2024；Taheri 等，2022；Wu 等，2022；Zhang 等，2025b），该类研究依托物体的三维几何信息，直接从三维空间数据中挖掘抓取动作的先验知识。同时，利用三维信息对生成的手指姿态进行几何优化，有效规避模型穿透等问题，从而生成符合物理规律与人体运动学的抓取动作序列。然而，此类方法高度依赖三维模型的精确性，在实际应用场景中，由于三维数据获取成本较高且存在噪声干扰，其应用范围受到显著限制。另一类研究则致力于从第一视角 RGB 图像中恢复人体动作（Li 等，2023a；Rhodin 等，2016；Tome 等，2019；Wang 等，2021a，2023；Xu 等，2019），该方向主要关注人体动作的重建问题，且侧重于人体自身运动的分析，忽略了第一视角场景中人体与物体的交互。

为攻克上述难题，本文提出了一种融合多模态信息的 Vision Transformer（ViT）（Dosovitskiy 等，2020）架构。该方法突破传统 ViT 的分块输入范式，创新性地采用 DinoV2（Oquab 等，2023）进行图像特征提取，同时利用多层感知机对当前动作姿态与视线方向进行编码。在模型核心架构上，Transformer 编码器中的多头自注意力机制实现对全局上下文的高效建模，精准捕捉图像特征与动作姿态间的时空依赖关系。此外，通过引入预测最终抓取姿态相对差异和手部接触标签两个辅助任务，构建复合损失函数优化模型训练，有效解决了无三维模型依赖下抓取动作生成的精确性与物理合理性问题。

本文在公开数据集 GRAB（Taheri 等，2020）上开展系统性实验。在测试阶段，本文设计了动态视角渲染机制，即系统会根据人体姿态变化实时调整相机外参。测评方面，本文从抓取姿态合理性与整体运动质量两大维度评估模型性能。实验结果显示，本文方法在各项指标上均显著优于现有先进方法，充分验证了其有效性与实用性。

1 相关工作

1.1 人体动作生成

人体动作生成是一个重要的研究问题（Guo 等，2022；Holden 等，2017；Starke 等，2019，2020，2022；Tevet 等，2023；Zhang 等，2018a）。最近的工作尝试了多种的神经网络模型，包括专家混合模型（MoE）（Zhang 等，2018b）、循环神经网络（Martinez 等，2017）、Transformer（Harvey 等，2020；Petrovich 等，2021；Vaswani 等，2017）和曼巴网络（mamba）（Gu and Dao，2023；Zhang 等，2025c）。为了增强动作的多样性和真实感，最近的工作还探索了各种生成范式，例如生成对抗网络（Goodfellow 等，2020；Li 等，2022）、归一化流（Henter 等，2020；Kingma and Dhariwal，2018）、变分自动编码器（Hassan 等，2021；Kingma and Welling，2013；Petrovich 等，2021）、矢量量化变分自动编码器（VQ-VAE）（Jiang 等，2023；Lu 等，2024；Razavi 等，2019；Zhang 等，2023，2024c）、扩散模型（Chen 等，2022；Ho 等，2020；Tevet 等，2023；Xiao 等，2025；Xie 等，2024；Zhang 等，2022a；Wang 等，2025；Jiang 等，2026）以及掩码建模（Chang 等，2022；Guo 等，2024）。随着大规模数据集的出现（Guo 等，2022；Li 等，2021；Lin 等，2024；Mahmood 等，2019；Plappert 等，2016），动作生成已经可以基于多种模态进行，比如文本（Guo 等，2024；Zhang 等，2023）、音乐（Tseng 等，2023）和音频（Chen 等，2024）。

1.2 人与物体的交互生成

人与物体交互生成的研究主要分为三类。一是人与物体（不含手）的交互，二是手与物体（不含身体）的交互，三是全身交互。本文考虑的是全身与物体交互的动作生成。

1）人与物体交互。早期的人与物体交互研究工作专注于与静态物体交互（Hassan 等 2021；Starke 等 2019；Zhang 等 2024b，2022b；Zhao 等 2023）。一些工作（Araújo 等 2023；Mir 等 2024；Wang 等 2021b，2022；Zheng 等 2022）采用自回归流程或全序列生成方法。近期的工作探索了基于扩散的模型（Huang 等 2023；Kulkarni 等 2024；Pi 等 2023；Yi 等 2025），并在扩散模型上应用引导技术（Dhariwal and Nichol，2021；Ho and Salimans，2022）来提升人与场

景的接触质量。近年来,有一些工作(Xu 等 2023,2024)开始考虑动态物体,或者基于给定的物体轨迹生成人体动作(Li 等 2023b)。例如,Li 等(2023b)提出了一个两阶段框架,首先生成手腕轨迹,然后相应地完成身体动作。

2)手与物体交互。ManipNet(Zhang 等,2021)在给定手腕和物体轨迹的情况下合成物体操作动作,使用多种表示方式对手与物体的关系进行建模。GRIP(Taheri 等,2024)设计了一种时间维度的手与物体空间特征,以实现稳定抓取。一些工作(Liu and Yi,2024;Zhou 等,2022)致力于解决对手部噪声动作进行去噪,以恢复干净的交互序列这一任务。另一些工作(Zhang 等,2024a,2025a;Zheng 等,2023)探索了仅提供物体轨迹的场景。此外,一些工作(Cha 等,2024;Christen 等,2024;Huang 等,2025)则联合生成手和物体的动作,而不依赖于预定义的轨迹。

3)全身交互。尽管存在一些全身操作数据集(Fan 等,2023;Taheri 等,2020),只有少数研究同时考虑全身和手部的交互。其中,一些工作(Taheri 等,2022;Wu 等,2022)假设物体是静态的,仅合成接近和抓取动作。IMoS(Ghosh 等,2023)在给定手指动作的情况下生成全身操作。TOHO(Li 等,2024)使用隐式表示(He 等,2022)合成全身交互。DiffGrasp(Zhang 等,2025b)使用扩散模型基于给定的物体轨迹生成全身动作,并引入人与物体交互引导以提高交互质量。

本文的方法考虑与静态的物体全身交互。与之前工作不同的是,本文仅通过第一人称视角的 RGB 图像作为输入,不需要物体的三维信息。

1.3 基于第一视角的人体动作捕捉

近年来,从第一人称视角相机来捕捉全身动作这一问题引起了越来越多的关注。EgoCap(Rhodin 等,2016)是首个提出使用安装在头盔上的立体鱼眼相机进行全身动作捕捉方法的研究。Mo2Cap2(Xu 等,2019)引入了一种分支双流卷积神经网络技术。XR-EgoPose(Tome 等,2019)采用了一种编码器-解码器架构,同时预测三维姿态与输入的二维热图,对二维关节位置的不确定性进行建模。(Wang 等,2021a)通过使用深度补全网络,缓解了由身体遮挡所带来的挑战。在(Wang 等,2021a)的基础上,(Wang 等,2023)通过结合检测到的关键点、基于变分自动编码器的运动先验以及基于 SLAM 的相机姿态估计,缓解了时间不稳定性和跟踪失败等问题。EgoEgo(Li 等,2023a)为摆脱对成对的第一人称视角视频与人体运动数据的依赖,将问题拆解为两个阶段,首先融合 SLAM 和学习方法精确估计头部运动;然后以估计出的头部姿态为条件,引导扩散模型生成多种合理的全身运动。虽然此类工作是基于第一视角的 RGB 图像来研究人体动作,与本文不同的是,他们关注于人体动作捕捉而非抓取动作生成。

2 本文方法

为解决从第一视角图像中生成人体抓取物体动作序列的问题,本文采用自回归的方式构建模型。在后续内容中,本文将首先在章节 2.1 详细阐述问题的定义与研究背景,明确任务的目标与约束条件;接着在章节 2.2 深入介绍本文提出的方法,包括模型架构设计、算法流程以及关键技术细节;最后,在章节 2.3 详细说明损失函数的设计思路,同时提供具体的实现步骤确保研究成果的可复现性与可靠性。

2.1 任务定义

首先对研究任务进行定义。本研究聚焦于从第一视角 RGB 图像中生成人体抓取物体的动作序列。如图 1 所示,输入为当前时刻 f 的第一视角 RGB 图像 I_f、当前时刻的动作姿态 Θ_f 以及视线方向 $\vec{v}_f$,输出是下一时刻的动作姿态 Θ_{f+1},可以表达为:

$$\Theta_{f+1}=G\left(I_f,\ \Theta_f,\ \vec{v}_f\right)。\tag{1}$$

其中 G 表示预测函数,Θ_{f+1}代表在上一时刻人体坐标系下,下一时刻的人体姿态。

对于第 i 帧的动作姿态,本研究采用与(Starke 等,2020,2021)一致的根坐标表示方式。定义某一时刻的动作姿态:

$$\Theta_i=\left\{\mathrm{r}_{\mathrm{off}},\ \mathrm{r}_{\mathrm{dir}},\ \theta_{\mathrm{pos}},\ \theta_{\mathrm{rot}},\ \theta_{\mathrm{vel}}\right\}。\tag{2}$$

其中,$\mathrm{r}_{\mathrm{off}}\in\mathbb{R}^2$ 和 $\mathrm{r}_{\mathrm{dir}}\in\mathbb{R}^2$ 分别表示第 i 帧人体坐标系相对于第(i−1)帧人体坐标系的水平位置和方向(不含 z 轴方向);$\theta_{\mathrm{pos}}\in\mathbb{R}^{\mathrm{J}\times 3}$、$\theta_{\mathrm{rot}}\in\mathbb{R}^{\mathrm{J}\times 6}$ 和 $\theta_{\mathrm{vel}}\in\mathbb{R}^{\mathrm{J}\times 3}$ 分别对应相对于第 i 帧人体坐标系的位置、6 维旋转和速度信息,J 代表人体关节数量。其中,人体坐标系可视化可见图 1 第二列第一行中的

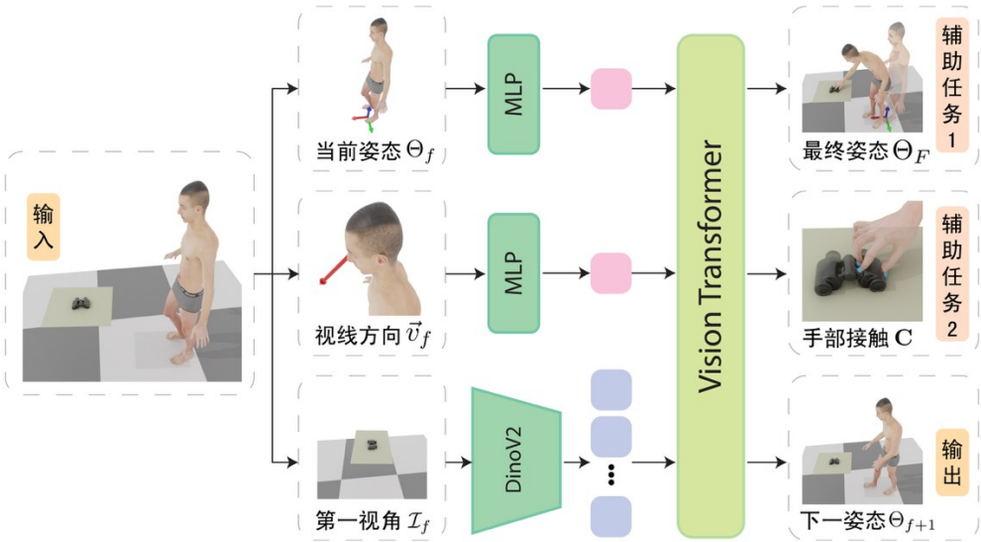

图1　本文方法概览

Fig. 1　Overview of the pipeline

三维坐标系，具体定义参考(Starke 等，2020，2021)。

第 i 帧的视线方向 $\vec{v}_i$ 是由人体双眼位置中心与注视点位置共同确定的，在本文中，将其设置为人眼中心指向物体中心的坐标向量。这是因为本研究发现人眼可在其骨架位置固定的情况下进行旋转，且在执行抓取任务时，人眼会自然将目标物体保持在视觉中心。

2.2　基于第一视角的自回归动作生成方法

如图 1 所示，本研究提出一种基于 Vision Transformer(ViT)(Dosovitskiy 等，2020)的自回归模型，从第一视角图像中预测下一时刻的动作姿态 Θ_{f+1}。该方法通过创新性地融合视觉特征提取与多模态信息编码，构建了高效的动作预测框架，具体包含以下关键模块。

1)图像特征提取模块。不同于传统 ViT 将原始图像划分为固定大小图像块(patch)的输入方式，本方法采用 DinoV2(Oquab 等，2023)作为图像特征提取器。DinoV2 是基于自监督对比学习的先进模型，通过在大规模无标签图像数据上进行训练，能够学习到具有高语义信息和强泛化能力的特征表示。其处理流程如下：首先将输入的第一视角图像 I_f 调整至合适尺寸，随后利用卷积神经网络进行初步特征提取；接着将得到的特征图划分为多个局部区域，通过多头注意力机制对不同区域的特征进行交互与融合；最终输出包含丰富语义和空间信息的图像特征向量。这些特征不仅保留了物体外观、位置等基础信息，还蕴含了场景的结构关系，为后续动作姿态预测提供了关键的视觉信息基础。

2)多模态信息编码模块。为有效整合视觉与动作信息，本研究采用多层感知机(MLP)对当前姿态 Θ_f 和当前视角 $\vec{v}_f$ 进行编码，将 Θ_f 和 $\vec{v}_f$ 映射为与 DinoV2 提取的图像特征维度一致的向量。随后，将图像特征向量、编码后的姿态向量和视角向量进行拼接，形成统一的输入序列。

3)Transformer 编码器模块。模型的编码器部分由多个 Transformer Block 级联构成，每个 Block 包含多头自注意力(MHA)模块与前馈神经网络(FFN)。在 MHA 模块中，多个并行的注意力头对输入序列进行独立的相似性计算与加权聚合，实现对全局上下文信息的多尺度、多视角建模，有效捕捉图像特征与动作姿态之间的长距离依赖关系。FFN 则通过非线性激活函数 GELU 对 MHA 模块的输出进行特征变换，进一步增强特征的表达能力，为自回归模型预测下一时刻动作姿态 Θ_{f+1} 提供高质量的特征表示。

4)辅助任务设计。此外，为提升模型预测的准确性和稳定性，本文引入了两个辅助任务。第一个辅助任务是预测最终抓取姿态 Θ_F 相对于当前帧人体坐标系的相对姿态，F 为最终帧。该任务通过计算最终姿态与当前姿态在平移、旋转等维度的差异，

引导模型学习动作序列的长期趋势和目标导向特征。这能增强模型对动作全局一致性和目标达成性的理解,避免生成偏离最终目标的动作姿态。第二个辅助任务是预测最终手部与物体的接触标签 **C**,该标签为二值变量,表示手部是否与物体发生接触。通过引入此任务,模型能够学习手部与不同物体之间的交互特征,捕捉抓取动作中的关键接触信息。

2.3 损失函数、网络结构与实验细节

2.3.1 损失函数设计

为优化基于 ViT 的自回归动作生成模型,本研究设计了包含主任务损失与辅助任务损失的复合损失函数,通过多目标优化策略驱动模型学习。具体损失函数由以下部分构成:

1)动作姿态预测损失。对于下一时刻动作姿态 Θ_{f+1},本研究采用均方误差(Mean Squared Error, MSE)度量预测姿态与真实姿态之间的差异。损失函数定义为:

$$L_{\text{main}}=\left|\left|\hat{\Theta}_{f+1}-\bar{\Theta}_{f+1}\right|\right|_2^2。\tag{3}$$

其中,($\hat{x}$) 和 ($\bar{x}$) 分别表示预测值与真实值。模型通过最小化该损失函数使模型输出更接近真实的下一帧动作姿态。

2)辅助任务损失。本研究针对两个辅助任务设计了相应的损失函数。针对预测最终抓取姿态 Θ_F 相对于当前帧坐标系的相对姿态任务,采用 MSE 损失函数,以增强模型对动作长期趋势的捕捉能力:

$$L_{\text{final}}=\left|\left|\hat{\Theta}_F-\bar{\Theta}_F\right|\right|_2^2。\tag{4}$$

其中,($\hat{x}$) 和 ($\bar{x}$) 分别表示预测值与真实值。对于手部与物体接触标签 **C** 的预测,由于其为二值变量,采用二元交叉熵(Binary Cross Entropy, BCE)损失函数:

$$L_{\text{contact}}=-(\bar{C}\log(\hat{C})+(1-\bar{C})\log(\hat{C}))。\tag{5}$$

其中,($\hat{x}$) 和 ($\bar{x}$) 分别表示预测值与真实值。该损失函数能促使模型准确预测手部与物体的接触状态,强化对抓取动作关键特征的学习。

3)总损失函数。将主任务损失与辅助任务损失进行加权求和,构建最终的总损失函数:

$$L_{\text{total}}=L_{\text{main}}+\alpha L_{\text{final}}+\beta L_{\text{contact}}。\tag{6}$$

其中,$\alpha=1.0$ 和 $\beta=1.0$ 为超参数。

2.3.2 网络结构设计

整体网络架构可分为视觉特征编码模块、时序与姿态嵌入模块、多模态特征融合模块及多任务预测头四个核心部分,具体设计如下:

1)视觉特征编码模块采用预训练的 DINOv2 模型(选取 ViT-B)作为图像编码器,选取 Transformer 中间层特征(对于 ViT-B,选取第 2、5、8、11 层)并融合 class token,输出多尺度图像特征。

2)时序嵌入模块基于正弦余弦位置编码生成基础时序编码,通过两层全连接层将时序特征映射至与模型维度匹配的嵌入空间,支持扩散模型不同时间步的特征表征。姿态嵌入模块对历史姿态、相机参数序列等输入特征进行归一化,再通过线性层映射至模型维度。

3)多模态特征融合模块以 ViT 网络作为核心融合骨干,输入包含视觉特征、时序嵌入、噪声姿态、历史姿态以及相机参数,通过自注意力机制实现多模态特征的深度融合。网络主体由 6 层 Transformer 编码器堆叠构成,每层编码器包含 16 头多头自注意力机制,单头注意力维度为 64 维,能够充分捕捉多模态特征间的长程依赖关系;前馈神经网络(MLP)隐藏层维度设为 2048 维,通过线性变换与非线性激活(SiLU)实现特征的非线性映射。

4)多任务预测头适配不同任务需求,所有预测头均为线性层,输出维度与目标任务匹配。

2.3.3 实现细节

在模型训练过程中,采用 Adam 优化器(Kingma 和 Ba, 2014)对模型参数进行更新,初始学习率设置为 0.0001。批量大小(batch size)设为 64,以平衡训练效率与内存占用。模型在 NVIDIA 4090 上进行训练,训练步数为 150000。为防止过拟合,在 MLP 和 FFN 层引入 Dropout 机制,丢弃率设为 0.1。此外,本研究对人体姿态 Θ_f 与 Θ_F 在训练集上进行标准化处理,确保输入数据的稳定性与模型收敛速度。

3 实验结果

3.1 数据集和数据生成方法

本文基于 GRAB 数据集(Taheri 等, 2020)开展实验研究。该数据集专注于全身抓取动作分析,完整记录了 10 名受试者与 51 种不同形状尺寸的日常物体交互过程中的姿态序列,为人体抓取动作生成

研究提供了丰富的基础数据。本研究根据(Taheri等,2022)将数据集划分为训练集、验证集和测试集。

然而,该数据集缺乏第一视角RGB图像,且未包含人体与物体的纹理信息。为弥补这些不足,本研究采用Blender渲染器,依据每一帧的视角参数生成第一视角RGB图像,以还原真实视觉场景。

针对人体纹理缺失问题,本研究借鉴(Black等,2023)中的方法,将纹理贴图映射至人体网格模型。在纹理选择上,对于不同性别的受试者,本研究随机选取亚洲人、西班牙裔、中东人、白人等不同人种特征的纹理类型,并在每种类型下进一步随机选用多种纹理样式,确保数据的丰富性与代表性。

对于物体纹理,本研究依据其材质特性与颜色特征进行精细化人工定制。在纹理参数设置上,精确调控RGB通道,同时对金属度(metallic)、镜面反射(specular)、粗糙度(roughness)、透明度(alpha)、自发光(emission)及透射率(transmission)等属性进行针对性调整,以逼真还原不同物体的表面质感与光学特性。具体纹理效果可详见图2。

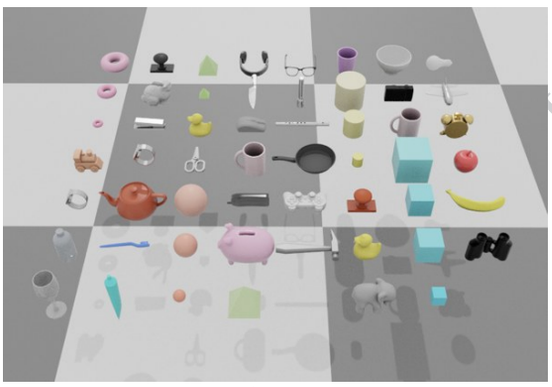

图2　物体纹理展示

Fig. 2　Object textures

3.2　实验设置

在测试阶段,为模拟真实场景下的动态视觉感知,系统需依据人体姿态变化实时调整观察视角,通过在每一帧画面渲染时更新相机参数,实现对不同视角下场景信息的动态捕捉。具体而言,基于人体双眼位置中心与目标物体中心的向量关系确定视线方向,同步更新Blender渲染器中的相机外参(平移、旋转矩阵),从而生成符合当前观察视角的RGB图像,为动作预测模型提供动态视觉输入。对于视线方向,默认人的眼睛必须朝着物体的中心点。

表1　与基线方法的对比实验

Table 1　Comparison with baselines

方法	最终抓取姿态合理性				整体运动质量	
	靠近率/%	成功率/%	接触率/%	不穿模率/%	脚步滑动	抖动
真值	100.00	96.36	56.86	96.56	0.52	5.35
EgoEgo	7.87	2.42	2.24	99.86	0.65	**5.49**
MDM	24.24	9.09	6.25	99.76	**0.64**	5.19
GOAL	54.54	9.69	20.00	97.90	0.83	10.64
DiffGrasp	56.36	12.12	23.03	97.09	0.86	5.56
本文	**91.51**	**53.33**	**38.64**	**96.42**	0.74	6.56

注：具体测试指标与基线方法详见章节3.3与3.4。加粗代表效果最优。靠近率、成功率和接触率越高越好;不穿模率、脚部滑动和抖动与真值数值越接近越好。

表2　消融实验

Table 2　Ablation study of main components

模块	最终抓取姿态合理性				整体运动质量	
	靠近率/%	成功率/%	接触率/%	不穿模率/%	脚步滑动	抖动
真值	100.00	96.36	56.86	96.56	0.52	5.35
(1) ViT only	75.15	27.27	25.78	98.07	0.76	**5.68**
(2) + DinoV2	88.48	48.48	35.86	96.82	0.79	7.08
(3) + DinoV2 + Θ_F	90.90	50.90	34.55	97.64	0.79	6.79
(4) + DinoV2 + Θ_F + C	**91.51**	**53.33**	**38.64**	**96.42**	**0.74**	6.56

注:注. 具体测试指标与实验设置详见章节3.3与3.5。加粗代表效果最优。靠近率、成功率和接触率越高越好;不穿模率、脚部滑动和抖动与真值数值越接近越好。

在抓取动作完成状态的判定上，本研究构建了基于距离阈值与时间约束的双重判定准则。首先，以手部顶点与目标物体表面的最小有符号距离作为核心判定指标，当该距离值小于预设阈值$\tau_{\text{intersect}}=2cm$时，初步判定抓取动作进入接触阶段。为确保抓取动作的稳定性与有效性，系统在触发接触判定后，将额外执行5帧动作预测，通过持续监测手部与物体的接触状态及姿态变化，验证抓取动作是否真正完成。若在70帧的预测周期内，手部与物体的距离始终未能达到接触阈值，则判定为抓取失败，系统将终止后续动作预测流程。

3.3 评估指标

办研究引入了六个评估指标来衡量最终抓取姿态的合理性与整体运动质量。其中四个指标用于评估抓取姿态的合理性，两个用于评估整体运动质量。

对于最终抓取姿态的合理性评估，本研究只对每一段数据的最终帧进行评价：1）靠近率：用于衡量手部是否成功靠近目标物体。通过计算手部顶点与目标物体的最小有符号距离（signed distance），当该值小于预设相交阈值$\tau_{\text{intersect}}=2cm$时，判定手部已接近目标。靠近率越高，表明手部与目标物体的距离越近，有助于实现稳定抓取。2）成功率：作为判断抓取动作是否成功的核心指标，该指标要求手部关键点（人工选取大拇指和食指表面的两个顶点）与目标物体的有符号距离需全部小于成功阈值$\tau_{\text{success}}=3cm$。成功率越高，表明抓取动作越成功。3）接触率：用于评估手部与目标物体的接触紧密程度。通过计算手部顶点中与目标物体距离在接触阈值$\tau_{\text{contact}}=3cm$以内的比例，反映实际接触情况。接触率越高，表明手部与目标物体的接触面积越大，有助于实现稳定抓取。4）不穿模率：用于量化抓取过程中手部与目标物体的碰撞情况。通过计算手部顶点中处于物体外部（有符号距离大于0）的比例，反映非碰撞程度。该值越大，表明抓取过程中手部与目标物体的碰撞风险越低。但是，如果该值过大，则可能意味着手部与物体之间的接触不够紧密，甚至没有接触。因此，本研究认为不穿模率与真实值越近，表明生成的最终姿态越合理。这四个指标从接近程度、任务完成度、接触质量及碰撞风险四个维度，构建了完整的抓取动作评估体系，为模型性能分析与算法优化提供量化依据。

对于整体运动质量的评估，本研究引入了在人体动作生成领域中常用的两个指标：1）脚部滑动：本文采用（Ling 等，2020）中的方法测量当脚部接近地面（距离地面小于 5 厘米）时的平均滑动距离。该指标值越接近真实值，表明动作质量越好。2）抖动：本文参照（Shen 等，2024）中的方法评估动作抖动情况。抖动指标值越接近真实值，代表动作质量越高。

3.4 对比实验

本研究与三个重要的基线对比，来验证本文方法的有效性。1）EgoEgo（Li 等，2023a）：EgoEgo 是一个基于第一视角恢复人体姿态的模型，其目标不是生成人体动作。该模型本身分为两个阶段：第一阶段从第一视角的图像中预测头部姿态，第二阶段把头部姿态作为条件式扩散模型的条件输入，来恢复人体动作。由于在当前的任务下，头部姿态是已知的，因此直接使用其第一阶段编码图像的编码器结果，作为第二阶段扩散模型的另一个条件来引导动作生成。2）MDM（Tevet 等，2023）：MDM 是一个基于扩散模型的动作生成模型，其输入是文本，输出是一整段动作。本研究将输入的文本改为第一视角图像，其余部分保持不变。3）GOAL（Taheri 等，2022）：GOAL是一个两阶段的基于三维物体信息的人体抓取动作生成方法，第一阶段预测最终的抓取姿态，第二阶段自回归地补全中间动作序列。4）DiffGrasp（Zhang 等，2025b）：DiffGrasp 是一个基于扩散模型的全身抓取动作生成方法，其输入不是第一视角的图像，而是物体三维姿态和形状 DiffGrasp 设计了全新的接触感知损失函数，并融入了引导机制。本研究把三维信息替换为第一视角图像，其余部分保持不变。

定性结果展示在表 1 中，可视化对比展示在图 3 中。在图 5 中展示了更多的生成结果。从图表中可以看出，本文方法在抓取姿态的合理性和整体运动质量上均显著优于其他方法。具体表现在：在靠近率、成功率和接触率上，本文方法均显著优于其他方法（靠近率从 54% 提升至 91%，成功率从 9% 提升至 53%，接触率从 20% 提升至 38%），说明本文方法能够更好地生成合理的抓取姿态。在不穿模率上，本文方法与真值最接近，并且综合靠近率和成功率来看，可以得出基线方法之所以更不容易穿模，是因为它们大多数情况下无法生成靠近物体或成功抓到物体的动作（例如图 3 的第二行最右侧结果）。

与 真值 相比，本文方法的不穿模率稍微低一些，但是仍然在合理范围内，表明本文方法能够生成合理的抓取姿态。在脚部滑动和抖动上，本文方法虽然不是最优的，但与真实值也较为接近。在图3中可以还可以看到，基线方法要么在设定的最大时间内无法靠近物体（EgoEgo，MDM），要么最终与物体穿模（GOAL）。

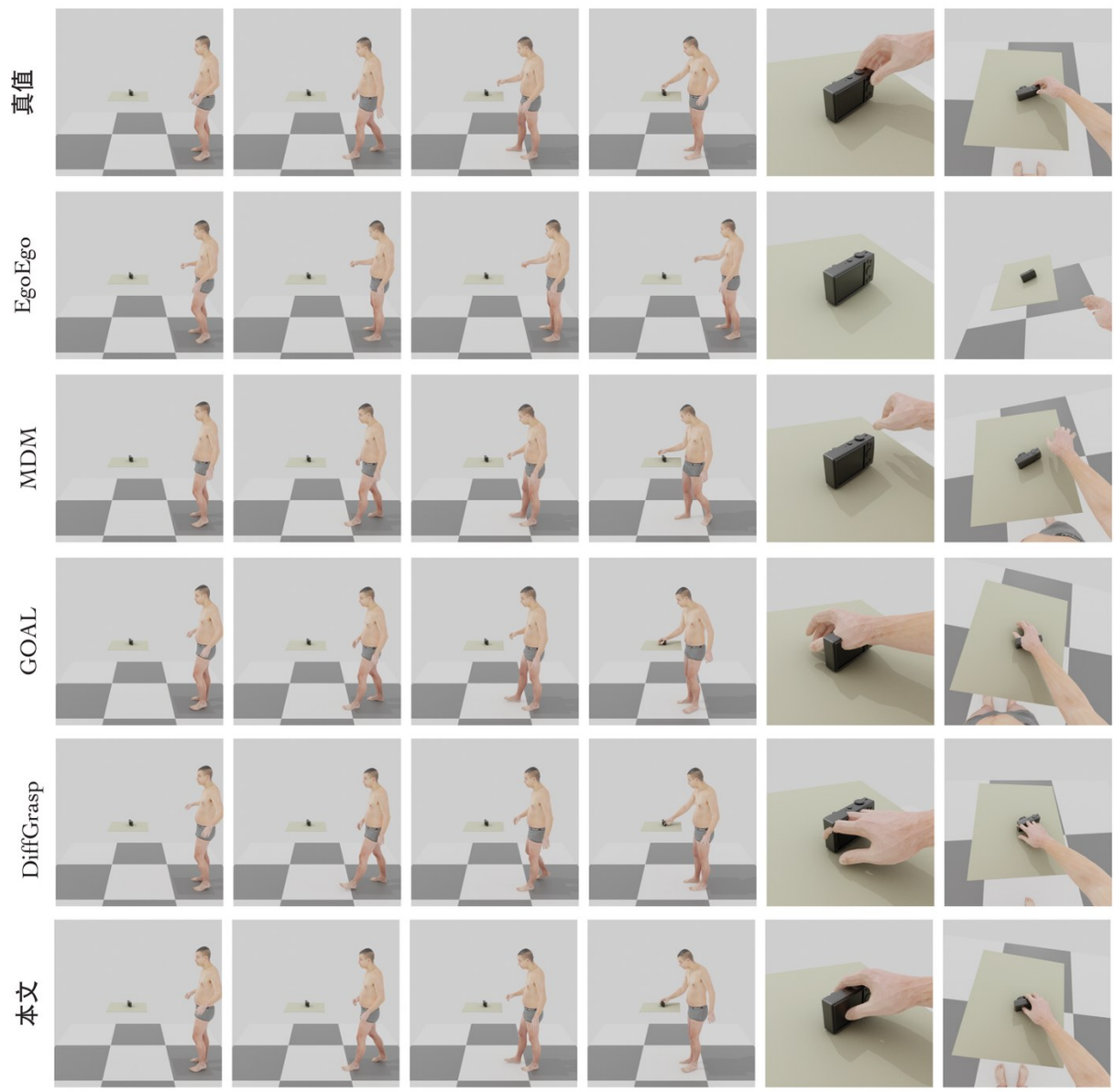

图3　与当前SOTA方法的可视化对比

Fig. 3　Visual comparison with SOTA methods

3.5　针对辅助任务的消融实验

本研究通过不同组合模块下的消融实验，进一步验证本文方法中预测额外的辅助任务对最终模型生成效果的有效性，本文设置了4个消融实验组：1）仅使用ViT编码器，不使用DinoV2编码器。2）在1）的基础上，使用DinoV2编码器对图像进行编码。3）在2）的基础上，额外在每一帧预测最终抓取姿态相对于当前帧人体坐标系的相对姿态。4）在3）的基础上，额外在每一帧预测最终抓取姿态下手部与物体的接触标签（本文方法）。

定性结果展示在表2中，可视化对比展示在，图4中。从图表中可以看出，每一个模块对于抓取姿

态的合理性和整体运动质量上均有不同程度的提

图4　消融实验的可视化对比

Fig. 4　Visual comparison of ablation study

图5　本文方法生成的更多可视化结果

Fig. 5　More visual results generated by our method

升。加入DinoV2编码器后，靠近率、成功率和接触率均显著提升（靠近率从75%提升至88%，成功率从27%提升至48%，接触率从25%提升至35%），说明DinoV2编码器能够更好地提取图像中的物体信息。对于最终能抓取到物体，额外预测Θ_F与接触**C**起到关键作用（成功率从48%提升至53%）。对于提升最终手与物体的接触率，额外预测接触**C**起到了关键作用，使得接触率从34%提升至38%。

3.6　针对超参数的对比实验

本研究对公式（6）中的超参数的设置进行对比实验，在保持其他条件均相同的情况下，只改变α和β的大小。结果如表3所示。由于当α=1.0和β=1.0的时候，大部分的指标均更优，本研究选取这一

表3　消融实验

Table 3　Ablation study of hyper-parameters

模块	最终抓取姿态合理性				整体运动质量	
	靠近率/%	成功率/%	接触率/%	不穿模率/%	脚步滑动	抖动
真值	100.00	96.36	56.86	96.56	0.52	5.35
α= 0.5,β= 1.0	81.81	37.57	33.71	97.23	0.89	7.06
α= 2.0,β= 1.0	**93.93**	33.33	36.02	96.59	0.91	6.79
α= 1.0,β= 0.5	87.87	36.36	29.70	**97.43**	0.82	6.89
α= 1.0,β= 2.0	81.81	42.42	36.48	96.45	0.76	6.92
α= 1.0,β= 1.0	91.51	**53.33**	**38.64**	96.42	**0.74**	**6.56**

注：注. 具体测试指标与实验设置详见章节3.3与3.6。加粗代表效果最优。靠近率、成功率和接触率越高越好；不穿模率、脚部滑动和抖动与真值数值越接近越好。

组超参数作为最终的设置。

4　本文局限

尽管本文方法在成功率上达到了53.33%，远远超过基线方法（EgoEgo 2.42%，MDM 9.09%，GOAL 9.69%，DiffGrasp 12.21%）。但仍在剩余的测试样例中失败。在图6中，给出了一些失败案例的可视化结果。

本文方法能达到91.51%的靠近率，即在大部分情况下均能靠近目标物体，但是其最终无法达到完美的抓取姿态。本研究认为有以下几个原因：1）本文的输入仅为第一视角的图像，网络对于物体在世界中的真实姿态没有物理感知；2）本文方法没有在物理引擎中仿真，当手指与物体穿模后，无法因碰撞调整姿态。

图6　失败案例

Fig. 6　Failure Cases

5　结　论

综上所述，本文针对第一视角RGB图像下人体抓取动作序列自回归生成的难题，提出了融合多模态信息的Vision Transformer架构。通过DinoV2进行图像特征提取、多层感知机编码动作与视角信息、Transformer架构建模全局依赖，并结合辅助任务优化损失函数，有效解决了缺乏三维模型下动作生成的精确性与物理合理性问题。在GRAB数据集上的实验表明，本文方法在抓取姿态合理性与运动质量等指标上显著优于现有方法，动态视角渲染机制也进一步增强了模型在真实场景中的适应性。

参考文献（References）

Araújo J P, Li J, Vetrivel K, Agarwal R, Wu J, Gopinath D, Clegg A W, and Liu K. 2023. Circle: Capture in rich contextual environments// In CVPR. IEEE/CVF: 21211-21221 [DOI: 10.1109/cvpr52729.2023.02032]

Black M J, Patel P, Tesch J, and Yang J L. 2023. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion// In CVPR. IEEE/CVF: 8726-8737 [DOI: 10.1109/cvpr52729.2023.00843]

Cha J, Kim J, Yoon J S, and Baek S. 2024. Text2hoi: Text-guided 3d motion generation for hand-object interaction//In CVPR. IEEE/CVF: 1577-1585 [DOI: 10.1109/cvpr52733.2024.00156]

Chang H, Zhang H, Jiang L, Liu C, and Freeman W T. 2022. Maskgit: Masked generative image transformer//In CVPR. IEEE/CVF: 11315-11325 [DOI: 10.1109/cvpr52688.2022.01103]

Chen C, Zhang J, Lakshmikanth S K, Fang Y, Shao R, Wetzstein G, Fei - Fei L, and Adeli E. 2024. The language of motion: Unifying verbal and non-verbal language of 3d human motion. arXiv preprints. https://arxiv.org/abs/2412.10523

Chen X, Jiang B, Liu W, Huang Z, Fu B, Chen T, Yu J, and Yu G. 2022.

Executing your Commands via Motion Diffusion in Latent Space. arXiv preprints. https://arxiv.org/abs/2212.04048

Christen S, Hampali S, Sener F, Remelli E, Hodan T, Sauser E, Ma S, and Tekin B. 2024. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions//In SIGGRAPH Asia. ACM. [DOI: 10.1145/3680528.3687563]

Dhariwal P and Nichol A. 2021. Diffusion models beat gans on image synthesis//In NeurIPS. NeurIPS Foundation: 8780-8794

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprints. https://arxiv.org/abs/2010.11929

Fan Z, Taheri O, Tzionas D, Kocabas M, Kaufmann M, Black M J, and Hilliges O. 2023. ARCTIC: A dataset for dexterous bimanual hand-object manipulation//In CVPR. IEEE/CVF: 12943-12954 [DOI: 10.1109/cvpr52729.2023.01244]

Ghosh A, Dabral R, Golyanik V, Theobalt C, and Slusallek P. 2023. Imos: Intent-driven full-body motion synthesis for human-object interactions. In Computer Graphics Forum, 42(2)

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, and Bengio Y. 2020. Generative adversarial networks. Communications of the ACM: 139-144

Gu A and Dao T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprints. https://arxiv.org/abs/2312.00752

Guo C, Zou S, Zuo X, Wang S, Ji W, Li X, and Cheng L. 2022. Generating diverse and natural 3d human motions from text//In CVPR. IEEE/CVF: 5152-5161 [DOI: 10.1109/cvpr52688.2022.00509]

Guo C, Mu Y, Javed M G, Wang S, and Cheng L. 2024. Momask: Generative masked modeling of 3d human motions//In CVPR. IEEE/CVF: 1900-1910 [DOI: 10.1109/cvpr52733.2024.00186]

Harvey F G, Yurick M, Nowrouzezahrai D, and Pal C. 2020. Robust motion in-betweening. ACM Trans. Graph, 39(4)

Hassan M, Ceylan D, Villegas R, Saito J, Yang J, Zhou Y, and Black M J. 2021. Stochastic scene-aware motion prediction//In ICCV. IEEE: 11374-11384 [DOI: 10.1109/iccv48922.2021.01118]

He C, Saito J, Zachary J, Rushmeier H, and Zhou Y. 2022. Nemf: Neural motion fields for kinematic animation//In NeurIPS. NeurIPS Foundation: 4244-4256

Henter G E, Alexanderson S, and Beskow J. 2020. Moglow: Probabilistic and controllable motion synthesis using normalising flows. ACM Trans. Graph, 39(6)

Ho J and Salimans T. 2022. Classifier-Free Diffusion Guidance. arXiv preprints. https://arxiv.org/abs/2207.12598

Ho J, Jain A, and Abbeel P. 2020. Denoising diffusion probabilistic models//In NeurIPS. NeurIPS Foundation: 6840-6851

Holden D, Komura T, and Saito J. 2017. Phase-functioned neural networks for character control. ACM Trans. Graph, 36(4)

Huang M, Chu F-J, Tekin B, Liang K J, Ma H, Wang W, Chen X, Gleize P, Xue H, Lyu S, Kitani K, Feiszli M, and Tang H. 2025. Hoigpt: Learning long sequence hand-object interaction with language models//In CVPR. IEEE/CVF: 7136-7146 [DOI: 10.1109/cvpr52734.2025.00669]

Huang S, Wang Z, Li P, Jia B, Liu T, Zhu Y, Liang W, and Zhu S-C. 2023. Diffusion-based generation, optimization, and planning in 3d scenes//In CVPR. IEEE/CVF: 16750-16761 [DOI: 10.1109/cvpr52729.2023.01607]

Jiang B, Chen X, Liu W, Yu J, Yu G, and Chen T. 2023. Motiongpt: Human motion as a foreign language//In NeurIPS. NeurIPS Foundation: 20067-20079

Jiang B, Song W, Hou X and Li S. 2026. Journal of System Simulation. China Simulation Federation, 38(1): 136 (蒋滨泽, 宋文凤, 侯霞, 李帅.2026.结合细粒度文本与空间控制信号的人体动作扩散模型.系统仿真学报.系统仿真学会, 38(1): 136

Kingma D P and Ba J. 2014. Adam: A Method for Stochastic Optimization. arXiv preprints. https://arxiv.org/abs/1412.6980

Kingma D P and Welling M. 2013. Auto-Encoding Variational Bayes. arXiv preprints. https://arxiv.org/abs/1312.6114

Kingma D P and Dhariwal P. 2018. Glow: Generative flow with invertible 1x1 convolutions//In NeurIPS. NeurIPS Foundation

Kulkarni N, Rempe D, Genova K, Kundu A, Johnson J, Fouhey D, and Guibas L. 2024. Nifty: Neural object interaction fields for guided human motion synthesis//In CVPR. IEEE/CVF: 947-957 [DOI: 10.1109/cvpr52733.2024.00096]

Li J, Liu K, and Wu J. 2023a. Ego-body pose estimation via ego-head pose estimation//In CVPR. IEEE/CVF: 17142-17151 [DOI: 10.1109/cvpr52729.2023.01644]

Li J, Wu J, and Liu C K. 2023b. Object motion guided human motion synthesis. ACM Trans. Graph, 42(6)

Li P, Aberman K, Zhang Z, Hanocka R, and Sorkine-Hornung O. 2022. Ganimator: Neural motion synthesis from a single sequence. ACM Trans. Graph, 41(4)

Li Q, Wang J, Loy C C, and Dai B. 2024. Task-oriented human-object interactions generation with implicit neural representations//In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision: 3035-3044

Li R, Yang S, Ross D A, and Kanazawa A. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++//In ICCV. IEEE: 13401-13412 [DOI: 10.1109/iccv48922.2021.01315]

Lin J, Zeng A, Lu S, Cai Y, Zhang R, Wang H, and Zhang L. 2024. Motion-x: A large-scale 3d expressive whole-body human motion dataset//In NeurIPS. NeurIPS Foundation: 25268-25280

Ling H Y, Zinno F, Cheng G, and Van De Panne M. 2020. Character controllers using motion vaes. ACM Trans. Graph, 39(4)

Liu X and Yi L. 2024. Geneoh diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion. arXiv preprints. https://arxiv.org/abs/2402.14810

Lu S, Wang J, Lu Z, Chen L-H, Dai W, Dong J, Dou Z, Dai B, and

Zhang R. 2024. Scamo：Exploring the scaling law in autoregressive motion generation model. arXiv preprints. https://arxiv. org/abs/2412.14559

Mahmood N, Ghorbani N, Troje N F, Pons-Moll G, and Black M J. 2019. Amass：Archive of motion capture as surface shapes//In ICCV. IEEE：5442-5451［DOI：10.1109/iccv.2019.00554］

Martinez J, Black M J, and Romero J. 2017. On human motion prediction using recurrent neural networks//In CVPR. IEEE/CVF：2891-2900［DOI：10.1109/cvpr.2017.497］

Mir A, Puig X, Kanazawa A, and Pons-Moll G. 2024. Generating continual human motion in diverse 3d scenes//In 3DV. IEEE：903-913

Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, Fernandez P, Haziza D, Massa F, El-Nouby A, Assran M, Ballas N, Galuba W, Howes R, Huang P, Li S, Misra I, Rabbat M, Sharma V, Synnaeve G, Xu H, Jegou H, Mairal J, Labatut P, Joulin A, Bojanowski P. 2023. Dinov2：Learning robust visual features without supervision. arXiv preprints. https://arxiv.org/abs/2304.07193

Petrovich M, Black M J, and Varol G. 2021. Action-conditioned 3D human motion synthesis with transformer VAE//In ICCV. IEEE：10985-10995［DOI：10.1109/iccv48922.2021.01080］

Pi H, Peng S, Yang M, Zhou X, and Bao H. 2023. Hierarchical generation of human-object interactions with diffusion probabilistic models//In ICCV. IEEE：15061-15073［DOI：10.1109/iccv51070.2023.01383］

Plappert M, Mandery C, and Asfour T. 2016. The kit motion-language dataset. Big data, 4(4)：236-252

Razavi A, van den Oord A, and Vinyals O. 2019. Generating diverse high-fidelity images with vq-vae-2//In NeurIPS. NeurIPS Foundation

Rhodin H, Richardt C, Casas D, Insafutdinov E, Shafiei M, Seidel H-P, Schiele B, and Theobalt C. 2016. Egocap：ego-centric marker-less motion capture with two fisheye cameras. ACM Trans. on Graph, 35(6)

Shen Z, Pi H, Xia Y, Cen Z, Peng S, Hu Z, Bao H, Hu R, and Zhou 2024. World-grounded human motion recovery via gravity-view coordinates//In SIGGRAPH Asia. ACM［DOI：doi.org/10.1145/3680528.3687565］(Starke S, Zhang H, Komura T, and Saito J. 2019. Neural state (machine for character-scene interactions. ACM Trans. Graph, 38(6)(Starke S, Zhao Y, Komura T, and Zaman K. 2020. Local motion (phases for learning multi-contact character movements. ACM Trans. Graph, 39(4)(Starke S, Zhao Y, Zinno F, and Komura T. 2021. Neural animation (layering for synthesizing martial arts movements. ACM Trans. Graph, 40(4)(Starke S, Mason I, and Komura T. 2022. Deep-phase：Periodic (autoencoders for learning motion phase manifolds. ACM Trans. Graph, 41(4)(Taheri O, Ghorbani N, Black M J, and Tzionas D. 2020. Grab：A (dataset of whole-body human grasping of objects//In ECCV. Springer：581-600［DOI：10.1007/978-3-030-58548-8_34］(Taheri O, Choutas V, Black M J, and Tzionas D. 2022. Goal：(Generating 4d whole-body motion for hand-object grasping//In CVPR. IEEE/CVF：13263-13273［DOI：10.1109/cvpr52688.2022.01291］(Taheri O, Zhou Y, Tzionas D, Zhou Y, Ceylan D, Pirk S, and Black (M J. 2024. GRIP：Generating interaction poses using latent consistency and spatial cues//In 3DV. IEEE：933-943 (Tevet G, Raab S, Gordon B, Shafir Y, Cohen-or D, and Bermano A (H. 2023. Human motion diffusion model//In ICLR. OpenReview (Tome D, Peluse P, Agapito L, and Badino H. 2019. xr-egopose：(Egocentric 3d human pose from an hmd camera//In ICCV. IEEE：7728-7738［DOI：10.1109/iccv. 2019.00782］(Tseng J, Castellon R, and Liu K. 2023. Edge：Editable dance (generation from music//In CVPR. IEEE/CVF：448-458［DOI：10.1109/cvpr52729.2023.00051］(Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, (Kaiser Ł, and Polosukhin I. 2017. Attention is all you need//In NeurIPS. NeurIPS Foundation (Wang J, Liu L, Xu W, Sarkar K, and Theobalt C. 2021a. Estimating (egocentric 3d human pose in global space//In ICCV. IEEE：11500-11509［DOI：10.1109/iccv48922.2021.01130］(Wang J, Luvizon D, Xu W, Liu L, Sarkar K, and Theobalt C. 2023. (Scene-aware egocentric 3d human pose estimation//In CVPR. IEEE/CVF：13031-13040［DOI：10.1109/cvpr52729.2023.01252］(Wang J, Xu H, Xu J, Liu S, and Wang X. 2021b. Synthesizing (long-term 3d human motion and interaction in 3d scenes//In CVPR. IEEE/CVF：9401-9411［DOI：10.1109/cvpr46437.2021.00928］(Wang J, Rong Y, Liu J, Yan S, Lin D, and Dai B. 2022. Towards (diverse and natural scene-aware 3d human motion synthesis//In CVPR. IEEE/CVF：20460-20469［DOI：10.1109/cvpr52688.2022.01981］(Wang J, Zhou L, and Zhang B. 2025. High-Efficiency Pose-Driven Human Motion Generation Technology Based on Diffusion Model Acceleration and Perception Optimization. Application Research of Computers, 42(10)(王家松，周雷，张博.2025.基于扩散模型加速和感知优化的高效姿态驱动人体动作生成技术.计算机应用研究,42(10)

Wu Y, Wang J, Zhang Y, Zhang S, Hilliges O, Yu F, and Tang S. 2022. Saga：Stochastic whole-body grasping with contact//In ECCV. Springer：257-274［DOI：10.1007/978-3-031-20068-7_15］

Xiao L, Lu S, Pi H, Fan K, Pan L, Zhou Y, Feng Z, Zhou X, Peng S, and Wang J. 2025. Motionstreamer：Streaming motion generation via diffusion-based autoregressive model in causal latent space. arXiv preprints. https://arxiv.org/abs/2503.15451

Xie Y, Jampani V, Zhong L, Sun D, and Jiang H. 2024. Omnicontrol：Control any joint at any time for human motion generation//In ICLR. OpenReview

Xu S, Li Z, Wang Y-X, and Gui L-Y. 2023. Interdiff：Generating 3d human-object interactions with physics-informed diffusion//In ICCV. IEEE：14928-14940［DOI：10.1109/iccv51070.2023.01371］

Xu S, Wang Y-X, Gui L, et al. 2024. Interdreamer：Zero-shot text to 3d dynamic human-object interaction//In NeurIPS. NeurIPS Foundation：52858-52890

Xu W, Chatterjee A, Zollhoefer M, Rhodin H, Fua P, Seidel H-P, and Theobalt C. 2019. Mo^{2}cap^2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. IEEE transactions on visualization and computer graphics, 25(5):2093-2101

Yi H, Thies J, Black M J, Peng X B, and Rempe D. 2025. Generating human interaction motions in scenes with text control//In ECCV. Springer: 246-263 [DOI: 10.1007/978-3-031-73235-5_14]

Zhang H, Starke S, Komura T, and Saito J. 2018a. Mode-adaptive neural networks for quadruped motion control. ACM Trans. Graph, 37(4)

Zhang H, Ye Y, Shiratori T, and Komura T. 2021. Manipnet: Neural manipulation synthesis with a hand-object spatial representation. ACM Trans. Graph, 40(4)

Zhang J, Zhang Y, An L, Li M, Zhang H, Hu Z, and Liu Y. 2024a. Manidext: Hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion. arXiv preprints. https://arxiv.org/abs/2409.09300

Zhang J, Zhang Y, Cun X, Zhang Y, Zhao H, Lu H, Shen X, and Shan Y. 2023. Generating human motion from textual descriptions with discrete representations//In CVPR. IEEE/CVF: 14730-14740 [DOI: 10.1109/cvpr52729.2023.01415]

Zhang M, Cai Z, Pan L, Hong F, Guo X, Yang L, and Liu Z. 2022a. Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprints. https://arxiv.org/abs/2208.15001

Zhang W, Dabral R, Leimkühler T, Golyanik V, Habermann M, and Theobalt C. 2024b. Roam: Robust and object-aware motion generation using neural pose descriptors//In 3DV. IEEE: 1392-1402

Zhang W, Dabral R, Golyanik V, Choutas V, Alvarado E, Beeler T, Habermann M, and Theobalt C. 2025a. Bimart: A unified approach for the synthesis of 3d bimanual interaction with articulated objects//In CVPR. IEEE/CVF: 27694-27705 [DOI: 10.1109/cvpr52734.2025.02579]

Zhang X, Bhatnagar B L, Starke S, Guzov V, and Pons-Moll G. 2022

b. Couch: Towards controllable human-chair interactions//In ECCV. Springer: 518-535 [DOI: 10.1007/978-3-031-20065-6_30]

Zhang Y, Huang D, Liu B, Tang S, Lu Y, Chen L, Bai L, Chu Q, Yu N, and Ouyang W. 2024c. Motiongpt: Finetuned llms are general-purpose motion generators//In AAAI. AAAI Press: 7368-7376

Zhang Y, He Q, Wan Y, Zhang Y, Deng X, Ma C, and Wang H. 2025b. Diff-grasp: Whole-body grasping synthesis guided by object motion using a diffusion model//In AAAI. AAAI Press: 10320-10328.

Zhang Z, Liu A, Reid I, Hartley R, Zhuang B, and Tang H. 2025c. Motion mamba: Efficient and long sequence motion generation//In ECCV. Springer: 265-282 [DOI: 10.1007/978-3-031-73232-4_15]

Zhao K, Zhang Y, Wang S, Beeler T, and Tang S. 2023. Synthesizing diverse human motions in 3d indoor scenes//In ICCV. IEEE: 14738-14749 [DOI: 10.1109/iccv51070.2023.01354]

Zheng J, Zheng Q, Fang L, Liu Y, and Yi L. 2023. Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis//In CVPR. IEEE/CVF: 585-594 [DOI: 10.1109/cvpr52729.2023.00064]

Zheng Y, Yang Y, Mo K, Li J, Yu T, Liu Y, Liu C K, and Guibas L J. 2022. Gimo: Gaze-informed human motion prediction in context//In ECCV. Springer: 676-694 [DOI: 10.1007/978-3-031-19778-9_39]

Zhou K, Bhatnagar B L, Lenssen J E, and Pons-Moll G. 2022. Toch: Spatio-temporal object-to-hand correspondence for motion refinement//In ECCV. Springer: 1-19 [DOI: 10.1007/978-3-031-20062-5_1]