Understanding Human-Scene Interaction through Perception and Generation

Understanding Human-Scene Interaction through Perception and Generation

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Hongwei Yi

aus Jiangxi, China

Tübingen

2025

Tag der mündlichen Qualifikation:	7th April, 2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Michael J. Black
2. Berichterstatter:	Prof. Dr. Gerard Pons-Moll

For my wife, dad, and mom

Abstract

Humans are in constant contact with the world as they move through it and interact with it. Understanding Human-Scene Interactions (HSIs) is key to enhancing our perception and manipulation of three-dimensional (3D) environments, which is crucial for various applications such as gaming, architecture, and synthetic data creation. However, creating realistic 3D scenes populated by moving humans is a challenging and labor-intensive task. Existing human-scene interaction datasets are scarce and captured motion datasets often lack scene information.

This thesis addresses these challenges by leveraging three specific types of HSI constraints: (1) depth ordering constraint: humans that move in a scene are occluded or occlude objects, thus, defining the relative depth ordering of the objects, (2) collision constraint: humans move through free space and do not interpenetrate objects, (3) interaction constraint: when humans and objects are in contact, the contact surfaces occupy the same place in space. Building on these constraints, we propose three distinct methodologies: capturing HSI from a monocular RGB video, generating HSI by generating scenes from input human motions (scenes from humans) and generating human motion from scenes (humans from scenes).

Firstly, we introduce MOVER, which jointly reconstructs 3D human motion and the interactive scenes from a RGB video. This optimization-based approach leverages these three aforementioned constraints to enhance the consistency and plausibility of reconstructed scene layouts and to refine the initial 3D human pose and shape estimations.

Secondly, we present MIME, which takes 3D humans and a floor map as input to create realistic and interactive 3D environments. This method applies collision and interaction constraints, and employs an auto-regressive transformer architecture that integrates objects into the scene based on existing human motion. The training data is enriched by populating the *3D FRONT* scene dataset with 3D humans. By treating human movement as a "scanner" of the environment, this method results in furniture layouts that reflect true human activities, increasing the diversity and authenticity of the environments.

Lastly, we introduce TeSMo, which generates 3D human motion from given 3D scenes and text descriptions, adhering to the collision and interaction constraints. It utilizes a text-controlled scene-aware motion generation framework based on denoising diffusion models. Annotated navigation and interaction motions are embedded within scenes to support the model's training, allowing for the generation of diverse and realistic human-scene interactions tailored to specific settings and object arrangements.

In conclusion, these methodologies significantly advance our understanding and synthesis of human-scene interactions, offering realistic modeling of 3D environments.

Kurzfassung

Menschen interagieren ständig mit ihrer Umgebung, was ein tiefes Verständnis für Mensch-Umgebungs-Interaktionen (Human-Scene Interactions, HSIs) essenziell macht, insbesondere zur Verbesserung unserer Wahrnehmung und Manipulation von 3D-Umgebungen. Dies ist für zahlreiche Anwendungen wie Spiele, Architektur und die Erstellung synthetischer Daten von großer Bedeutung. Die Erstellung realistischer 3D-Szenen mit Menschen ist jedoch anspruchsvoll und arbeitsintensiv. Vorhandene Datensätze zu Mensch-Umgebungs-Interaktionen sind rar, und Bewegungsdatensätze enthalten oft keine Informationen über die Umgebung.

Diese Dissertation adressiert diese Herausforderungen durch drei spezifische Einschränkungen bei HSI: (1) Tiefenordnung, bei der Objekte teilweise verdeckt oder verdecken und so eine relative Tiefenstruktur der Szene bilden, (2) Kollision, die verhindert, dass Menschen Objekte durchdringen, und (3) Interaktion, bei der Kontaktflächen denselben Raum einnehmen.

Basierend auf diesen Einschränkungen stellen wir drei Methodologien vor: die Erfassung von HSI aus monokularen RGB-Videos, die Erzeugung von Szenen durch menschliche Bewegungen und das Generieren menschlicher Bewegungen in bestehenden Szenen. Zunächst präsentieren wir MOVER, das 3D-Menschenbewegungen und interaktive Szenen aus einem RGB-Video rekonstruiert. Dieser optimierungsbasierte Ansatz nutzt die genannten Einschränkungen, um Konsistenz und Plausibilität der rekonstruierten Szenenlayouts zu verbessern.

Zweitens präsentieren wir MIME, das auf Basis von 3D-Menschen und Grundrissdaten interaktive 3D-Umgebungen erstellt. Durch eine autoregressive Transformer-Architektur werden Objekte in die Szene integriert, sodass realistische Möbelanordnungen menschliche Aktivitäten widerspiegeln.

Schließlich stellen wir TeSMO vor, das 3D-Menschenbewegungen für gegebene 3D-Szenen und Textbeschreibungen generiert und dabei Kollisionen und Interaktionen berücksichtigt. Mit einem textgesteuerten Framework auf Basis von Denoising-Diffusionsmodellen ermöglicht TeSMO eine vielfältige und realistische Mensch-Umgebungs-Interaktion.

Insgesamt tragen diese Methoden zur realistischen Modellierung von 3D-Umgebungen und einem vertieften Verständnis von Mensch-Umgebungs-Interaktionen bei.

Preface

The following works Yi *et al.* (2022, 2023b, 2024) form the part of this thesis.

- **Human-Aware Object Placement for Visual Environment Reconstruction**
  Authors: Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, Michael J. Black
  Conference: Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR), 2022

- **MIME: Human-Aware 3D Scene Generation**
  Authors: Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, Michael J. Black
  Conference: Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR), 2023

- **Generating Human Interaction Motions in Scenes with Text Control**
  Authors: Hongwei Yi, Justus Thies, Michael J. Black, Xue Bin Peng, Davis Rempe
  Conference: European Conference on Computer Vision (ECCV), 2024

These following works Yi *et al.* (2023a); Dai *et al.* (2023); Huang *et al.* (2023b); Tripathi *et al.* (2023); Liao *et al.* (2024); Huang *et al.* (2024); Dwivedi *et al.* (2024); Zhang *et al.* (2024c) were completed during my PhD but are not part of the thesis:

- **Generating Holistic 3D Human Motion from Speech**
  Authors: Hongwei Yi*, Hualin Liang*, Yifei Liu*, Qiong Cao†, Yandong Wen, Timo Bolkart, Dacheng Tao, Michael J. Black†
  Conference: Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR), 2023

- **SLOPER4D: A Scene-Aware Dataset For Global 4D Human Pose Estimation In Urban Environments**
  Authors: Yudi Dai, Yitai Lin, Xiping Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yuexin Ma, Cheng Wang
  Conference: Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR), 2023

*Equal Contribution.
†Joint Corresponding Authors.

- **DECO: Dense Estimation of 3D Human-Scene Contact in the Wild**
  Authors: Shashank Tripathi*, Agniv Chatterjee*, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, Michael J. Black
  Conference: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023

- **TeCH: Text-guided Reconstruction of Lifelike Clothed Humans**
  Authors: Yangyi Huang*, Hongwei Yi*, Yuliang Xiu*, Tingting Liao, Jiaxiang Tang, Deng Cai, Justus Thies
  Conference: Proceedings of the IEEE/CVF International Conference on 3D Vision (3DV), 2024

- **TADA! Text to Animatable Digital Avatars**
  Authors: Tingting Liao*, Hongwei Yi*, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, Michael J. Black
  Conference: Proceedings of the IEEE/CVF International Conference on 3D Vision (3DV), 2024

- **POCO: 3D Pose and Shape Estimation using Confidence**
  Authors: Sai Kumar Dwivedi, Cordelia Schmid, Hongwei Yi, Michael J. Black, Dimitris Tzoinas
  Conference: Proceedings of the IEEE/CVF International Conference on 3D Vision (3DV), 2024

- **Real-time Monocular Full-body Capture in World Space via Sequential Proxy-to-Motion Learning**
  Authors: Yuxiang Zhang, Hongwen Zhang, Liangxiao Hu, Hongwei Yi, Shengping Zhang, Yebin Liu
  Conference: Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR), 2024

Acknowledgments

The past four years have been transformative, and I am amazed at how much my world has changed. My initial motivations for pursuing this PhD were to immerse myself in this unique journey and rediscover who I am. Now, as I reflect on this period, I see both myself and the world through a refreshed lens.

I owe an immense debt of gratitude to my advisor, Michael J. Black. His extraordinary ability to transform complex ideas into elegant research papers is nothing short of magical. Beyond his academic prowess, Michael's life philosophies, subtly interwoven into our conversations, have offered profound inspiration and guidance. His precision in every detail has taught me how to be a responsible person and, subsequently, a dedicated research scientist. Michael built a collaborative, open research lab filled with wonderful people. I will particularly miss Nicole and Melanie, whose support made our PhD life easier, especially amidst the bureaucratic challenges. During the coronavirus pandemic, they helped me integrate into research life in Tübingen and immerse myself in the Perceptive Systems group.

During my time in Perceptive Systems, I collaborated extensively with Justus Thies. His hands-on experience in computer graphics and fruitful research insights greatly aided my first published paper and enriched my overall research in human-scene interaction.

I am grateful to my post-doc mentors, Dimitrios Tzionas and Chunhao Paul Huang. Dimitrios guided me in shaping research projects and emphasized the importance of identifying the key problem to solve. Paul supported my research from the beginning, often responding promptly to my queries. Especially during the pandemic, Paul was a constant presence in the office, helping me organize and prioritize tasks, which was invaluable as deadlines approached.

A special thanks to my TAC member, Andreas Geiger. His passion for research science and his perspective on academic life have been inspiring. His guidance to always seek fun in both research and life resonates deeply with the same philosophies of Michael.

I would like to thank my lab mates and collaborators, Yuliang Xiu, Shashank Tripathy, Muhammad Kocabas, Mohamed Hassan, Yandong Wen, Zhen Liu and Weiyang Liu, for creating a vibrant and collaborative environment. Your camaraderie has been a source of motivation.

I am grateful to Lea Hering, an intern in Perceiving Systems. Her enthusiasm and dedication have made my research life more inspiring, and her contributions to rendering set a high bar for my future work.

During my Ph.D. journey, I was fortunate to work with external collaborators, Yudi Dai, and Yuxiang Zhang, on various projects. They enriched my research experience in

human-scene interaction, social avatars, in-the-wild motion capturing, and more.

Collaborating with younger master's and junior PhD students has also been rewarding. Thanks to Yifei Liu, Hualin Liang, Tingting Liao, Yangyi Huang, and Liang Pan for their self-motivation and pursuit of excellence, which continually inspired me.

During my time in Tuebingen, my labmates – Haiwen Feng, Yao Feng, Soubik Sanyal, Lea Muller, Yinghao Huang, Qianli Ma, Peter Kulits, Omid Taheri, Nikos Athanasiou, Mathias Petrovich, Yufeng Zheng, and Xu Chen – made this journey enjoyable and intellectually stimulating. Your support and feedback have been invaluable.

I am fortunate to have friends like Huanbo Sun, Haolong Li, Zhenzhong (Tim) Xiao, Zeju Qiu, Zehao Li and Yuxuan Xue. Your companionship has been a source of joy and strength.

While doing my PhD, I also spent time at NVIDIA. There, I had the privilege of working with exceptional mentors. Davis and Xue Bin believed in my ideas and provided the support I needed. I am grateful for the opportunities to learn and grow under their mentorship. Ye Yuan also supported my research, and Sanya Fidler welcomed me warmly, making me feel comfortable and proud to be part of her group at NVIDIA.

Reflecting on my early research years, I want to thank my Master's advisor, Guoping Wang, Sheng Li, and other external supervisors Chen Li and Qiong Cao, for generously supporting my exploration of 3D computer vision. My peers Chaopeng Zhang, Yu Chen, Yao Wang, Zizhuang Wei, and Xu Gao accompanied me not only in research but also enriched my life at Peking University. I would also like to thank my bachelor friends, Chao Chen, Liang Li, Xinyu Li, Longcheng Zhai, Cheng Ren, Jie Wen, and Lige Ding, whose friendship I deeply appreciate.

I would not have embarked on this PhD journey without the encouragement from my wife, Yunyao Xue. She believed in me even before I did and always supported me in pursuing my true passions. My deepest gratitude goes to my wife, mother, and father. Your unwavering love and support have been my constant anchors, providing the foundation for my success.

Thank you all for being an integral part of this incredible journey. Your support, mentorship, and friendship have made my PhD experience truly unforgettable.

Contents

Chapter 1

Introduction

Understanding Human-Scene Interactions (HSIs) is pivotal for deciphering human behavior and enhancing our manipulation of three-dimensional (3D) environments. From the earliest civilizations, humans have been in constant contact with their surroundings, interacting with the world as they move through it. These interactions, whether walking through a room, touching objects, resting on a chair, or sleeping in a bed, contain rich information about scene layout and object placement. Gibson's seminal work in the ecological approach to visual perception Gibson (2014) emphasizes the direct, adaptive relationship between visual perception and action in the environment, where people perceive scenes in terms of the opportunities for interaction they afford. Where and how human interacts with a scene can be used to predict future motions and interactions for human-centered AI and robots, or to synthesize these for AR/VR and other computer-graphics applications.

1.1 What is Human-Scene Interaction?

Human-scene interaction refers to the complex relationship between humans and their surrounding environment, encompassing various behaviors and dynamics. In computer vision and scene understanding, it is essential to model and understand these interactions to accurately reconstruct three-dimensional (3D) scenes and predict human behavior. Here, we delineate three fundamental constraints that govern human-scene interaction, as shown in Figure 1.1:

Depth Ordering Observation and Constraint: In natural scenes, when humans move, they often occlude or are occluded by objects within the environment. This interaction provides essential depth cues, helping to determine the relative positions of objects in the scene. **Observation**: The occlusion patterns observed during human motion reveal the depth ordering of objects. **Constraint**: To maintain accurate scene reconstruction, the depth ordering inferred from these occlusion patterns must be respected, ensuring that objects are positioned correctly relative to one another.

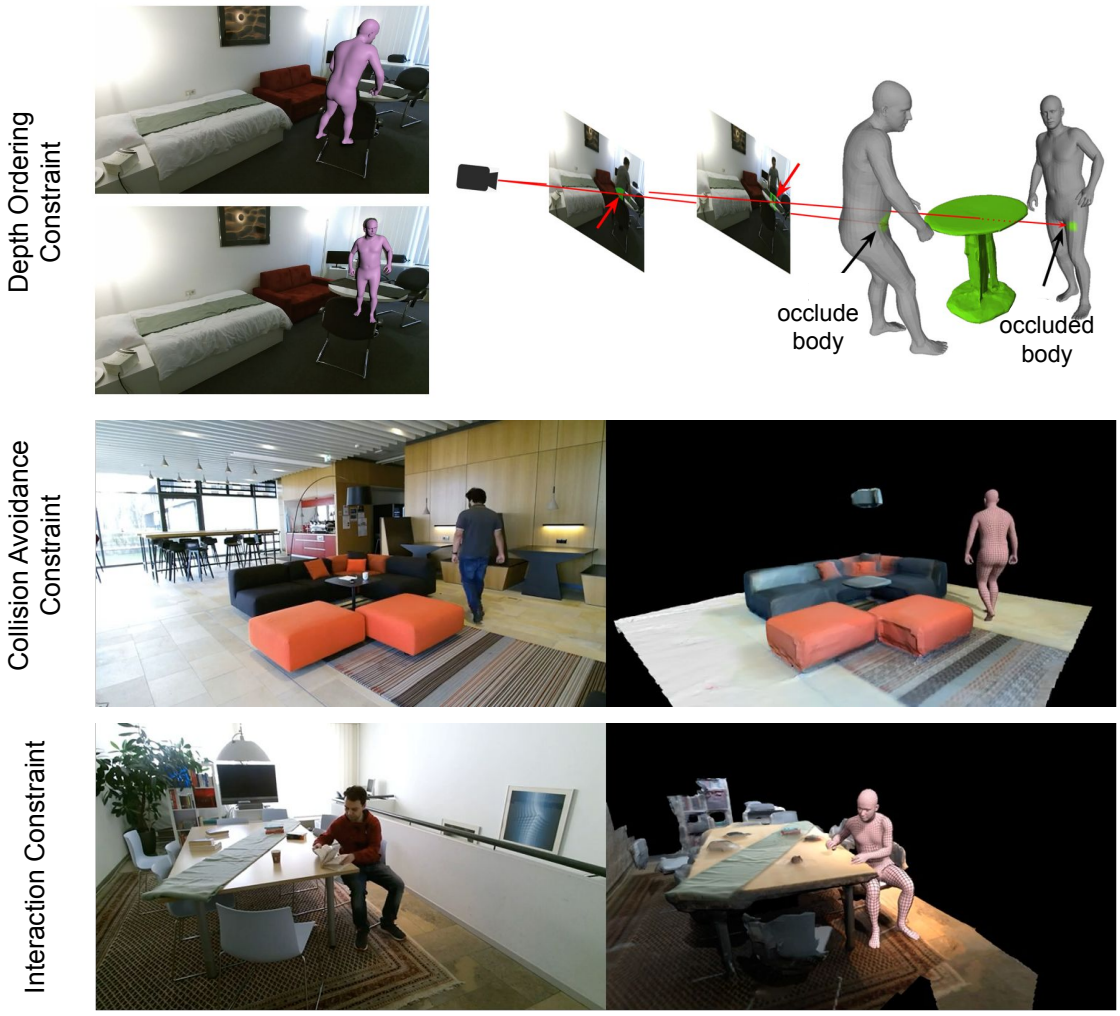


Figure 1.1: The three fundamental human-scene interaction constraints: depth ordering constraint, collision avoidance constraint, and interaction constraint.

Collision Avoidance Observation and Constraint: Humans typically navigate through a scene by moving within free space and avoiding direct collisions with objects. **Observation**: The paths taken by humans naturally avoid interpenetration with objects, highlighting the spatial boundaries within the scene. **Constraint**: To preserve physical realism, scene reconstruction must enforce collision avoidance, ensuring that humans and objects do not overlap or intersect unnaturally, guiding the generation of plausible human trajectories.

Contact Consistency Observation and Constraint: When interacting with objects, such as grasping, leaning, or sitting, humans establish physical contact that aligns the surfaces of their bodies with those of the objects. **Observation**: The contact points

between humans and objects are consistent, occupying the same physical space. **Constraint**: Scene reconstruction must enforce this contact consistency to maintain spatial coherence, accurately modeling the relationships between humans and objects to enhance the realism of the generated scenes.

1.2 Perceiving and Generating Human-Scene Interaction

These human-scene interaction constraints serve as foundational principles to model and understand the interactions between humans and their environment. By integrating these constraints into computational frameworks, we aim to advance the synthesis of realistic 3D scenes and enable more accurate predictions of human behavior in various applications, including robotics, augmented reality, and computer graphics.

1.3 Capturing Human-scene Interactions

Tremendous progress has been achieved in reconstructing three-dimensional (3D) human bodies and scenes from monocular images or videos, as evidenced by a plethora of recent works Kocabas *et al.* (2021); Guler and Kokkinos (2019); Joo *et al.* (2020); Kanazawa *et al.* (2018); Kocabas *et al.* (2020); Kolotouros *et al.* (2019b); Pavlakos *et al.* (2019a,c); Yuan *et al.* (2022, 2021); Luo *et al.* (2021); Huang *et al.* (2018a); Nie *et al.* (2020); Zhang *et al.* (2021a); Dahnert *et al.* (2021); Božič *et al.* (2021). However, these advancements have predominantly focused on reconstructing humans and scenes in isolation, neglecting the inherent interaction between humans and scenes. In reality, humans interact with their environment, causing partial occlusion of both the scene by humans and humans by the scene. This interaction poses challenges for accurate reconstruction, particularly in scenarios involving strong occlusion.

1.3.1 Capture Human-scene Interaction through Multiple Sensors

As depicted in Figure 1.2, PROX Hassan *et al.* (2019) addresses the challenge of capturing human-scene interaction (HSI) in indoor environments using monocular RGB-D input. By pre-scanning the scene and capturing dynamic RGB-D sequences of subjects, PROX generates a dataset of 100K RGB-D frames with reconstructed human motion serving as pseudo Ground Truth. However, this approach faces limitations, as reconstructed HSI from a monocular RGB-D view is often unreliable due to frequent occlusion. Capturing large-scale datasets requires cumbersome offline 3D reconstruction from multiple viewpoints (Zollhöfer *et al.* (2018)). RICH Huang *et al.* (2022) extends this capability to outdoor scenarios, employing eight HDR multiple-view cameras and improving 3D scene reconstruction quality with an industrial laser scanner, Leica RTC360. On

the other hand, SLOPER4D Dai *et al.* (2023) records human activities in urban environments from an egocentric view using a head-mounted device integrated with LiDAR and camera technology. It captures 15 sequences of human motion, each spanning trajectory lengths from 200 to 1,300 meters and covering areas from 200 to 30,000 m^2, including extensive LiDAR, video, and IMU-based motion frames. While SLOPER4D is comprehensive in synchronization across multiple sensors, it requires considerable time to capture data.

1.3.2 Independent Reconstruction of Human Motion and 3D Scenes from Monocular Videos

Instead, there is a pressing need for methods capable of estimating both scene and humans solely from images, as the absence of depth information can lead to inconsistencies in scale and object placement relative to interacting humans. This is challenging, as the lack of depth information causes the scale and placement of objects to be inconsistent with respect to the humans interacting with them. This leads to physically implausible results, like humans penetrating objects, or lacking physical contact when walking, sitting, or lying down, causing bodies to "hover" in the air (see Figure 1.3). Methods that reconstruct 3D humans from single views leverage statistical body models Joo *et al.* (2018); Loper *et al.* (2015); Pavlakos *et al.* (2019a); Xu *et al.* (2020) as priors on the body shape and pose. However, the same tools do not exist for the collective space of 3D scene layouts. This is due to the enormous space of possible object arrangements in indoor 3D scenes, a large number of different object classes, and the huge inter-class (e.g., chairs and desks) and intra-class (e.g., desk chair and club chair) shape variability. Without considering human-scene interaction (HSI) cues, such as depth ordering, collision avoidance, and contact consistency, these methods struggle to produce coherent and realistic reconstructions, leading to issues like misaligned objects and unnatural human poses.

1.3.3 Joint Reconstruction of 3D Humans and Scenes from Monocular Videos

The input to MOVER includes color frames from a static monocular camera, 3D human meshes, and detected object shapes. The output is a refined 3D scene where objects are repositioned to align with human interactions, avoiding unrealistic interpenetration. This process is guided by several key observations about human-scene interactions, which inform specific constraints that ensure the physical plausibility of the reconstructed scenes.

First, when humans move within a scene, they often occlude or are occluded by objects, providing crucial depth cues. MOVER leverages this observation by enforcing a **depth ordering constraint** that ensures objects are positioned correctly relative to one another based on these occlusion patterns. This prevents inconsistencies in the spatial

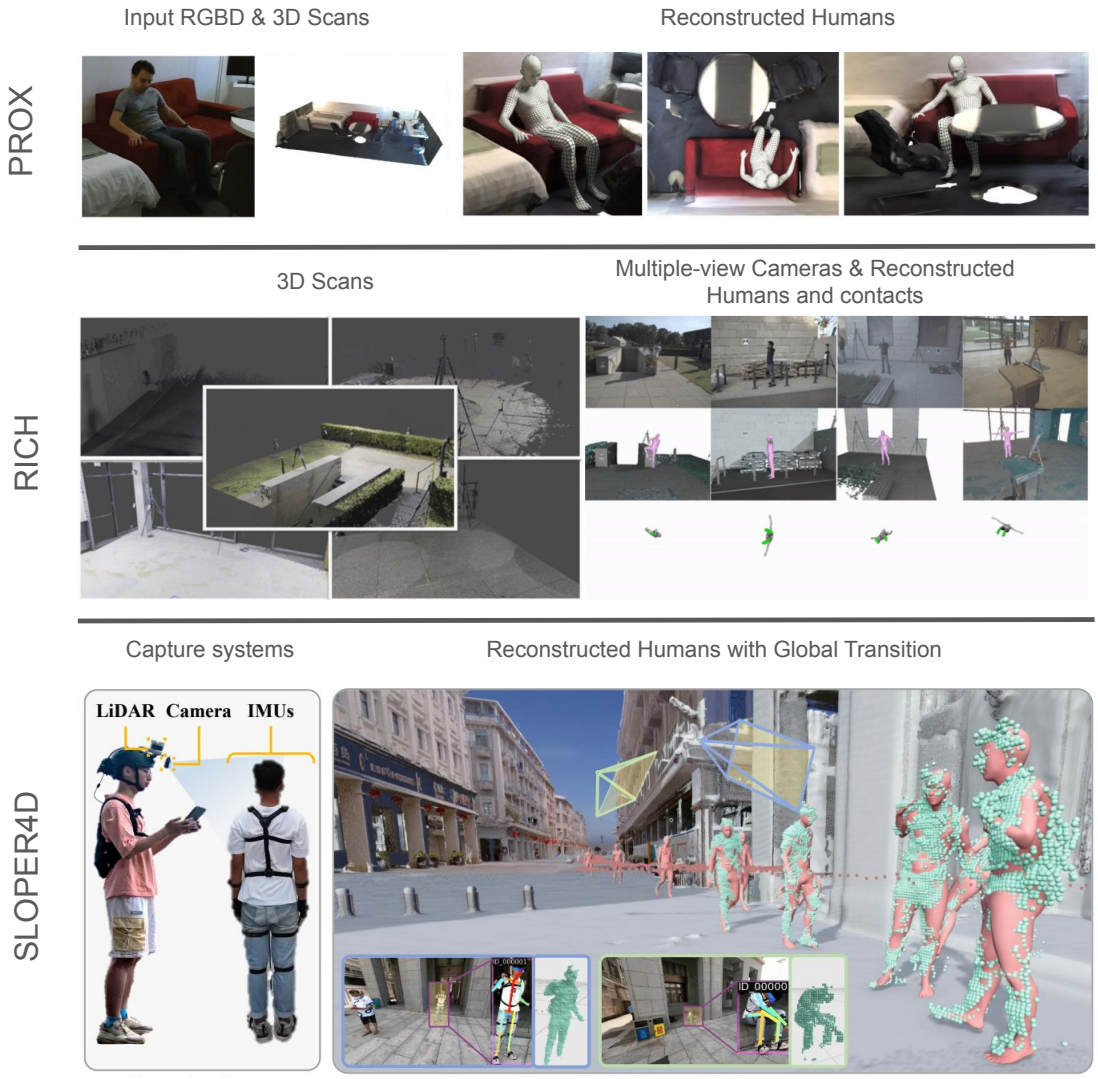
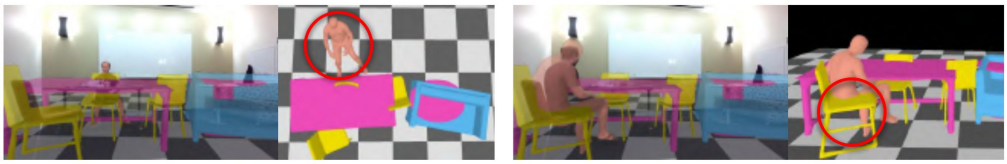


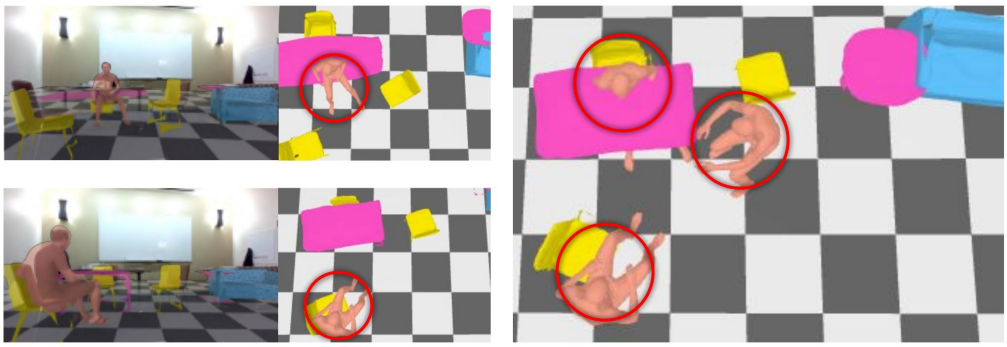
Figure 1.2: The illustration of capturing humans and scenes using multiple sensors. PROX Hassan *et al.* (2019) uses RGB-D and pre-scanned scenes from Kinect to reconstruct 3D humans within rooms. RICH Huang *et al.* (2022) utilizes multi-view cameras and laser scanners to enhance motion and scene reconstruction quality. SLOPER4D Dai *et al.* (2023) records human activities in urban settings from an egocentric perspective with a head-mounted device that combines LiDAR and camera technology.

relationships within the scene.

Second, humans naturally navigate through a scene by avoiding collisions with objects, indicating where free space exists. This observation informs a **collision avoidance constraint** in MOVER, which prevents human and object geometries from overlapping or interpenetrating, maintaining the realism of the reconstructed scene.



(a) 3D scene reconstruction and HPS in isolation



(b) HolisticMesh. L: single-image results. R: multiple-images result.

Figure 1.3: Where existing methods struggle: (a) humans in estimated scenes penetrate objects or lack contact with objects and "hover" in the air when estimated in isolation Nie *et al.* (2020); Pavlakos *et al.* (2019a) (b) humans interpenetrate objects, even, when the 3D scenes and humans are jointly optimized with single (left) or sequential images (right) Weng and Yeung (2020).

Finally, when humans interact with objects, such as sitting or leaning, the contact points between the human and object surfaces should align correctly. This leads to the **contact consistency constraint**, where MOVER ensures that these interactions are spatially coherent, preventing issues like floating or misaligned body parts.

These integrated HSI principles—depth ordering, collision avoidance, and contact consistency—guide the optimization process in MOVER. The approach leverages both explicit cues, like the alignment of contact points, and implicit cues, such as maintaining free space and preventing penetrations, to achieve highly realistic scene reconstructions.

In contrast, existing methods often rely on oversimplified object shapes Chen *et al.* (2019) or focus solely on static human poses interacting with a single object Zhang *et al.* (2020a), without considering HSI across multiple interaction frames Weng and Yeung (2020); Zhang *et al.* (2020a); Chen *et al.* (2019). This leads to less accurate and physically inconsistent results.

MOVER outperforms these existing methods on datasets such as PROX Hassan *et al.* (2019) and PiGraphs Savva *et al.* (2016), producing more accurate 3D scene layouts while minimizing penetrations with moving humans. Additionally, MOVER can refine human poses using the estimated 3D scenes, similar to PROX-like methods Hassan *et al.* (2019). This highlights the synergy between accurately estimating 3D scenes and human poses from monocular camera data, showcasing the effectiveness of integrating HSI principles in the reconstruction process.

1.4 Generating Human-scene Interactions through Scenes from Humans

Capturing human-scene interactions is labor-intensive, requiring extensive manual effort to record and annotate data. This raises the question: can we generate human-scene interactions instead? Humans constantly interact with their environment—walking through rooms, touching objects, resting on chairs, or sleeping in beds. These interactions provide valuable information about the layout of scenes and the placement of objects within them. For instance, mimes use our understanding of such interactions to convey a rich, imaginary 3D world using only their body motions. Inspired by this, we explore whether a computer can be trained to interpret human motion and similarly conjure the 3D scene in which it belongs.

Generating plausible 3D scenes from human motion presents several challenges. Human motion alone does not provide explicit visual information about the surrounding environment, making it difficult to accurately infer object types, positions, and scales. Additionally, without clear cues, there is a risk of generating physically implausible scenes, such as floating objects or overlapping geometries. Despite these challenges, the potential applications of such a method are vast, including synthetic data generation, architecture, gaming, and virtual reality. Existing large datasets of 3D human motion,

like AMASS Mahmood *et al.* (2019), rarely include information about the 3D scenes in which the motions were captured. If we can generate plausible 3D scenes for the motions in AMASS, we could produce training data that contains realistic human-scene interactions.

1.4.1 Generate Scenes in Isolation

Generative scene synthesis is a rapidly evolving field within computer graphics and artificial intelligence, focusing on the automatic creation of 3D environments. However, most prior work in this area has predominantly ignored the human element, leading to scenes that may not accommodate human presence or interaction. These techniques include procedural modeling with grammars Müller *et al.* (2006); Parish and Müller (2001); Prakash *et al.* (2019); Talton *et al.* (2011); Kar *et al.* (2019); Devaranjan *et al.* (2020); Purkait *et al.* (2020), graph neural networks Li *et al.* (2019a); Wang *et al.* (2019a); Zhou *et al.* (2019,?); Luo *et al.* (2020); Purkait *et al.* (2020); Zhang *et al.* (2020e,c); Keshavarzi *et al.* (2020); Di *et al.* (2020), auto-regressive neural networks Ritchie *et al.* (2019); Wang *et al.* (2018a), and transformers Wang *et al.* (2021c); Paschalidou *et al.* (2021); Para *et al.* (2023). These methods often generate 3D scenes based on various modeling techniques but fail to incorporate human motion as a guiding factor.

Some approaches utilize lexical text Chang *et al.* (2015) or sentences Chang *et al.* (2017b) as input to guide the 3D scene synthesis. Fisher et al. Fisher *et al.* (2012) take 3D scans as input and synthesize the corresponding 3D object arrangements, extending to include functionality aspects in the reconstruction Fisher *et al.* (2015). Recently, ATISS Paschalidou *et al.* (2021) has advanced this field by performing scene synthesis using a transformer-based architecture. ATISS takes a floorplan as input and auto-regressively generates a 3D scene represented as an unordered set of objects. However, these methods do not consider the dynamic element of human motion, which is critical for generating scenes that are not only visually coherent but also functional for human interaction. The challenge here lies in integrating human motion data to guide the 3D scene synthesis in a way that maintains physical plausibility and realism.

1.4.2 Generate Scenes from Humans

To address the aforementioned challenges, we develop a method called MIME (Mining Interaction and Movement to infer 3D Environments), which generates plausible indoor 3D scenes based on 3D human motion.

The feasibility of this approach rests on two key intuitions:

1. **Human Motion Defines Free Space**: A human's motion through free space indicates the absence of objects, carving out regions free of furniture.

2. **Human Interaction Constrains Object Placement**: When humans interact with the scene, their contact points constrain the type and placement of 3D objects. For example, a sitting human might be on a chair, sofa, or bed.

Given an empty floor plan and a human motion sequence, MIME predicts furniture in contact with the human and plausible objects that fit with other objects while respecting the free-space constraints induced by human motion.

To condition the 3D scene generation, we estimate possible contact poses using POSA Hassan *et al.* (2021a) and divide the motion into contact and non-contact snippets. Non-contact poses define free space, encoded as 2D floor maps by projecting foot vertices onto the ground plane. Contact poses and corresponding 3D human body models are represented by 3D bounding boxes of the contact vertices predicted by POSA.

To train MIME, we built a new dataset called *3D-FRONT HUMAN*, extending the large-scale synthetic scene dataset 3D-FRONT Fu *et al.* (2021a). We populated 3D scenes with humans, including non-contact humans (walking or standing) and contact humans (sitting, touching, and lying). This process leverages motion sequences from AMASS Mahmood *et al.* (2019) and static contact poses from RenderPeople Patel *et al.* (2021) scans.

In summary, our contributions are:

- A novel motion-conditioned generative model for 3D room scenes that generates objects in contact with the human auto-regressively and respects free-space constraints defined by motion.
- A new 3D scene dataset, *3D-FRONT HUMAN*, with interacting humans and free-space humans, constructed by populating 3D FRONT with static contact/standing poses from RenderPeople and motion data from AMASS.

1.5 Generating Human-scene Interaction through Humans from Scenes

Creating realistic human movements that interact with 3D scenes is essential for various applications, such as gaming and embodied AI. For example, animators in games and films need to craft motions that navigate through complex scenes and interact realistically with objects while maintaining control over the movement's style. One intuitive way to control style is through text, e.g., "skip happily to the chair and sit down".

Recent advancements in diffusion models have shown impressive capabilities in generating human motion from user inputs. Text prompts Tevet *et al.* (2023); Zhang *et al.* (2022a) allow users to control the style, while spatial constraints enable detailed control, such as specifying joint positions and trajectories Xie *et al.* (2024); Shafir *et al.* (2023); Karunratanakul *et al.* (2023). However, these methods often focus on characters in isolation, without considering the surrounding environment or object interactions.

What we want is to generate human motions in 3D scenes while the motion style can be controlled with text input.

1.5.1 Generate Human Motions: Text Isolation and Scene Isolation

Text-Driven Motion Generation. Recent advancements in diffusion models have significantly increased their capacity to generate high-quality human motions, particularly when conditioned on textual prompts. Models like those documented in the literature Tevet *et al.* (2023); Zhang *et al.* (2022a); Petrovich *et al.* (2024) excel in rendering realistic motions that are tightly aligned with textual descriptions. This represents a notable improvement over previous methods that aimed to synchronize text with motion Petrovich *et al.* (2022); Ahn *et al.* (2018), as well as those focused on spatial composition Athanasiou *et al.* (2022). However, these models primarily generate motions in isolation, without considering interactions with the environment, limiting their usefulness in real-world applications where such interactions are essential.

Environment-Oriented Motion Generation. Traditionally, motion generation technologies have focused mainly on creating animations for characters in isolation, often ignoring the important effects of environmental interactions. The AMASS dataset Mahmood *et al.* (2019) is a good example of this trend. It offers a comprehensive collection of motion data but does not include any environmental interactions, which limits the diversity and realism of the generated motions in various scenes. Efforts to integrate more complex environmental interactions have utilized specialized datasets, which are often small and lack text annotations. These datasets have been used to train various models, including VAEs Hassan *et al.* (2021b); Zhang *et al.* (2022b); Starke *et al.* (2019), diffusion models like SceneDiffuser Huang *et al.* (2023a), and projects focused on specific object interactions Pi *et al.* (2023). These methods, though innovative, still struggle to accurately capture the nuanced dynamics of human-environment interactions due to the limited and sometimes noisy data used. In contrast, our methodology seeks to overcome these challenges by employing a pre-trained text-to-motion diffusion model Tevet *et al.* (2023), augmented with a fine-tuned scene-aware component.

1.5.2 Generate Human Motions: Scenes and Text Integration

Our key idea combines general, scene-agnostic text-to-motion diffusion models with paired human-scene data for realistic interactions. First, we pre-train a text-conditioned diffusion model Tevet *et al.* (2023) on a diverse motion dataset without objects (e.g., HumanML3D Guo *et al.* (2022)) to learn realistic motion patterns and their correlation with text. We then fine-tune this model with an additional scene-aware component that incorporates scene information, refining the motion outputs to align with the environment.

Given a target object and a text prompt describing the desired motion, we break down the task into two parts: *navigation* (approaching the object while avoiding obstacles) and *interaction* (interacting with the object). Both stages use diffusion models pre-trained on scene-agnostic data and fine-tuned with a scene-aware branch.

In summary, our contributions are:

- A novel approach for scene-aware and text-conditioned motion generation by fine-tuning an augmented model on top of a pre-trained text-to-motion diffusion model.
- A method, TeSMo, that leverages this approach for navigation and interaction components to generate high-quality motions in a scene from text.
- Data augmentation strategies for realistically placing navigation and interaction motions with text annotations in scenes to enable scene-aware fine-tuning.

1.6 Thesis organization

The organization of this thesis is structured as follows:

Chapter 2 provides the background needed to explore human-scene interactions and highlights recent advances in reconstructing and generating these interactions. It begins by reviewing existing datasets for human-scene interactions, then covers methods for reconstructing these interactions from videos and other sources. The chapter also discusses progress in creating 3D scenes and generating 3D human motions from different inputs.

Chapter 3 presents joint reconstruction of human-scene interaction from a monocular RGB video. It details an optimization-based approach to jointly reconstruct 3D human motion and scenes from RGB videos. We conduct a comprehensive evaluation using the PROX Hassan *et al.* (2019) and PiGraphs Savva *et al.* (2016) datasets to demonstrate the method's effectiveness.

Chapter 4 explores human-aware 3D scene generation, introducing a novel generative model that incorporates human motion data to create interactive 3D environments. This chapter shifts focus from reconstructing to generating human-scene interactions by synthesizing 3D scenes based on input human motions.

Chapter 5 focuses on scene-aware motion generation via text control. This chapter describes an innovative approach that integrates text-driven commands into the motion generation process. Annotated navigation and interaction motions are embedded within scenes to facilitate training. We evaluate the performance on our generated dataset and demonstrate that our approach exceeds previous techniques in terms of the plausibility, realism, and variety of the generated human-scene interactions.

Chapter 6 wraps up the thesis by summarizing the key insights from the three constraints used in reconstructing and generating human-scene interaction datasets. It also outlines future work focused on generating realistic human videos that include not only human-scene interactions but also social behaviors, such as conversations and assisting actions.

Chapter 2

Related Work

2.1 Human-Scene Interaction Datasets.

To provide a comprehensive overview of datasets for modeling human-scene interactions, this discussion is divided into two main sections: first, the shortcomings of existing datasets in capturing detailed human-scene interactions; and second, the development of new synthetic datasets designed to overcome these limitations.

2.1.1 Capturing Human-Scene Interaction Datasets

Current datasets primarily focus on either human movements or static scenes, using technologies such as optical markers Sigal *et al.* (2010); Ionescu *et al.* (2014); CMU Graphics Lab (2000), IMU sensors von Marcard *et al.* (2018); Huang *et al.* (2018c), and multiple RGB cameras Joo *et al.* (2017); Yu *et al.* (2020); Mehta *et al.* (2018). These often exclude essential 3D environments like floors, walls, and furniture, which are crucial for understanding human interactions within a space. Conversely, datasets like Matterport3D Chang *et al.* (2017a), ScanNet Dai *et al.* (2017), and Replica Straub *et al.* (2019) capture detailed static scenes using time-of-flight sensors but lack dynamic human elements, limiting their effectiveness in modeling human-scene interactions. While recent datasets Huang *et al.* (2022); Guzov *et al.* (2021b); Hassan *et al.* (2019); Monszpart *et al.* (2019); Bhatnagar *et al.* (2022); Wang *et al.* (2019b) combine humans and environments, their static nature and absence of dynamic scene elements reduce their utility. Efforts like those by Hassan et al. Hassan *et al.* (2021b) to enhance datasets by adjusting object sizes and human poses still fail to capture full scene dynamics.

As noted in the bottom part of Table 2.1, many existing datasets related to contact focus on self-contact Fieraru *et al.* (2021); Müller *et al.* (2021) or person-to-person interactions Fieraru *et al.* (2020), but inadequately address human-scene contact (HSC). Relevant datasets for HSC, such as Guzov *et al.* (2021a) and PROX Hassan *et al.* (2019), have limitations. The former provides egocentric images for localization, unsuitable for HSC detection from third-party perspectives. While potentially useful, the PROX dataset is limited to indoor scenes and suffers from low-quality and occlusion issues in its RGB-D data-derived ground-truth bodies. This restricts the variety of captured human-scene

Datasets	Contact Label	Scene
Captured Dataset		
MTP Müller *et al.* (2021)	self-contact	N/A
GRAB Taheri *et al.* (2020)	hand-object	N/A
ContactHands Narasimhaswamy *et al.* (2020)	hand-X‡	N/A
Fieraru et al. Fieraru *et al.* (2020)	person-person	N/A
Fieraru et al. Fieraru *et al.* (2021)	self-contact	N/A
PiGraph Savva *et al.* (2016)	joint-scene	RGB-D scans
i3DB Monszpart *et al.* (2019)	N/A	CAD
GPA Wang *et al.* (2019b)	N/A	Cubes
Guzov et al. Guzov *et al.* (2021a)	foot-ground	laser scans
PROX Hassan *et al.* (2019)	body-scene	RGB-D scans
RICH Huang *et al.* (2022)	body-scene	laser scans
Synthetic Dataset		
Pose2Room Nie *et al.* (2022)	body-scene (skeleton)	3D scenes (scans)
GTA-IM Cao *et al.* (2020)	body-scene (skeleton)	3D scenes (CAD)
3D-FRONT HUMAN (Ours)	body-scene (static)	3D scenes (CAD)
Loco-3D-FRONT (Ours)	body-scene (motion)	3D scenes (CAD)

Table 2.1: **Comparison of human-scene interactions datasets**.

interactions (mostly walking, sitting, lying) and affects the quality of body fits.

2.1.2 New Synthetic Datasets for Enhanced Training

Composite or synthetic datasets such as Patel *et al.* (2021); Bazavan *et al.* (2021); Cai *et al.* (2021) are also widely used for human mesh recovery, but their meaningful human-scene interaction is fairly limited. To our knowledge, Pose2Room Nie *et al.* (2022) and GTA-IM Cao *et al.* (2020) are the closest to our needs. However, they represent humans with 3D skeletons, which cannot represent realistic contact between the body surface and the scene. Also, the scene arrangement is still not rich enough to train a generative model.

To overcome these deficiencies, we introduce two synthetic datasets: *3D FRONT Human* and **Loco-3D-FRONT** tailored for advanced machine learning models like MIME and TeSMo. The 3D FRONT Human dataset utilizes the 3D FRONT project's scenes Fu *et al.* (2021a), populating them with human models that move and interact with the environment in a dynamic fashion. Complementing this, the Loco-3D-FRONT dataset integrates locomotion sequences from HumanML3D into a variety of 3D environments, generating about 9,500 walking motions, each accompanied by textual descriptions and ten plausible 3D scenes, thus creating approximately 95,000 training pairs Yi *et al.* (2023b); Guo *et al.* (2022). This dataset not only captures realistic motions but also allows for randomized initial translations and orientations of the motions within scenes, augmented

further with left-right mirroring to enhance variety.

These synthetic datasets are created to offer a richer and more dynamic resource for training models that need a detailed understanding of human interactions in complex 3D environments. By overcoming the limitations of existing datasets, these new resources aim to improve machine learning's ability to understand and predict human behavior in real-world scenarios.

2.2 Reconstructing Human-scene Interactions

2.2.1 Reconstruct Single-view 3D Human Pose in "Isolation"

Estimating human pose from an image is a long-standing problem Moeslund *et al.* (2006); Sarafianos *et al.* (2016). Typically, this is cast as estimating 2D or 3D joints of the body Andriluka *et al.* (2018); Martinez *et al.* (2017); Rogez and Schmid (2016); Tekin *et al.* (2016); Tome *et al.* (2017) or a whole-body Cao *et al.* (2021); Jin *et al.* (2020); Weinzaepfel *et al.* (2020) skeletons. Recently, there has been a significant shift in research interest towards reconstructing the 3D human body surface which, in contrast to the joints, interacts directly with objects and can be observed by commodity cameras. To this end, many non-parametric methods Gabeur *et al.* (2019); Kolotouros *et al.* (2019a); Saito *et al.* (2019, 2020); Smith *et al.* (2019); Varol *et al.* (2018); Zheng *et al.* (2019) have been developed, estimating either depth maps Gabeur *et al.* (2019); Smith *et al.* (2019), 3D voxels Varol *et al.* (2018); Zheng *et al.* (2019), 3D distance fields Saito *et al.* (2019, 2020), or free-form 3D meshes Kolotouros *et al.* (2019a). While these methods can reconstruct bodies with details like hair and clothing, they miss semantic information and correspondence information. In contrast, parametric statistical 3D shape models for the body Anguelov *et al.* (2005); Hasler *et al.* (2009); Loper *et al.* (2015) and a whole body Joo *et al.* (2018); Pavlakos *et al.* (2019a); Xu *et al.* (2020) provide this information and allow re-posing. Since parametric models represent the shape and pose in a low-dimensional space, they are a powerful tool to estimate the surface from incomplete data (e.g., 2D images with occlusions) through optimization Bogo *et al.* (2016); Joo *et al.* (2018); Pavlakos *et al.* (2019a); Xiang *et al.* (2019), regression Choutas *et al.* (2020); Kanazawa *et al.* (2018); Kocabas *et al.* (2020); Kolotouros *et al.* (2019b), or hybrid approaches Joo *et al.* (2020).

However, all the above methods reason about the human in "isolation", i.e. without taking the surrounding objects and scenes into account. Thus, they struggle to reconstruct details like contact with objects, and often fail due to occlusions (e.g., bodies standing behind furniture). PARE Kocabas *et al.* (2021) addresses this by leveraging localized features and attention, gaining occlusion robustness. We initialize our approach with Kocabas *et al.* (2021) to refine the 3D scene layout.

Method	GDI	Cam.	C-HOI	N-HOI	FGC
PHOSA Zhang *et al.* (2020a)	✓	✗	✓	✗	✗
Holistic++ Chen *et al.* (2019)	✗	✗	✗	✗	✓
HolisticMesh Weng and Yeung (2020)	✓	✓	✓	✗	✓
Ours	✓	✓	✓	✓	✓

Table 2.2: Comparison of the most relevant methods. GDI: Geometric Detailed Interaction. C-HOI: Contact-Human-Object Interaction. N-HOI: Exploiting free space constraints with no object contact. FGC: Feet-Ground Contact. Cam.: Camera orientation and ground-plane are refined with humans or not.

2.2.2 Reconstructing Single-view 3D Scene in Isolation

3D reconstruction from single views has been addressed in several recent works that leverage learned geometrical priors for specific object classes or entire scenes. Shapes from single views are reconstructed using generative models for specific object classes Choy *et al.* (2016); Groueix *et al.* (2018); Wang *et al.* (2018b); Mescheder *et al.* (2019); Sitzmann *et al.* (2019). The methods differ in the underlying representation, which ranges from volumetric representations like occupancy fields Mescheder *et al.* (2019) and implicit surface functions Park *et al.* (2019); Liu *et al.* (2020), to explicit surface representations like triangular meshes Wang *et al.* (2018b); Gkioxari *et al.* (2019). To reconstruct scenes, single objects can be detected He *et al.* (2017) and reconstructed in isolation. Mesh-RCNN Gkioxari *et al.* (2019) detects the objects in an RGB image and predicts the geometry for each object individually. Instead of a generative mesh model, Izadinia *et al.* (2017) and Kuo *et al.* (2020) propose to retrieve individual CAD models for the detected objects in the scene. Bansal *et al.* (2016) predict a normal map from the input image that is used to align a retrieved CAD model. Instead of predicting normal maps from the input image, there is a series of methods that estimate depth maps Laina *et al.* (2016); Godard *et al.* (2017); Fu *et al.* (2018); Shin *et al.* (2019), or pixel-aligned implicit functions for objects Saito *et al.* (2019, 2020); Xiu *et al.* (2022) and scenes Denninger and Triebel (2020); Dahnert *et al.* (2021). Joint estimation of the room layout and objects with scene context information has been proposed by Choi *et al.* (2013); Huang *et al.* (2018a,b); Zhang *et al.* (2017); Zhao and Zhu (2013); Zhang *et al.* (2021a); Nie *et al.* (2020). However, these methods only consider an isolated 3D scene without a human in it.

Note that there are also methods that predict room layouts with 3D bounding boxes Dasgupta *et al.* (2016); Hedau *et al.* (2009); Lee *et al.* (2009); Mallya and Lazebnik (2015). In contrast, we reconstruct the detailed object geometry to leverage explicit contact point constraints based on the human scene interactions, while optimizing for the scene layout.

2.2.3 Reconstructing 3D Human-Scene Interaction

Humans inhabit 3D scenes. Several methods model this and learn to populate a 3D scene Hassan *et al.* (2021a); Li *et al.* (2019b); Zhang *et al.* (2020b,d). In contrast, our work reasons about the human and its interaction with the 3D scene from RGB observations. There are several methods that explore different kinds of human scene interaction; these can be divided into three categories based on the granularity of the interaction between human and scenes: (1) Hand-Object Cao *et al.* (2021); Yang *et al.* (2021); Chao *et al.* (2021a); Liu *et al.* (2021); Jiang *et al.* (2021); Kwon *et al.* (2021). (2) Body-Object Zhang *et al.* (2020a); Dabral *et al.* (2021); Xie *et al.* (2025). (3) Body-Scene Monszpart *et al.* (2019); Chen *et al.* (2019); Weng and Yeung (2020); Huang *et al.* (2022).

Our proposed method focuses on reconstructing 3D scenes composed of objects and structural elements like the floor plane, using accumulated human scene interactions (body-objects and body-scene). Table 2.2 summarizes the most related work that operates on single-view RGB images/videos. PHOSA Zhang *et al.* (2020a) infers humans and objects together when they are in contact. They do not consider the fact that humans do not need to contact an object to constrain its location; their movement through free space constrains object placement. Zanfir *et al.* (2018) only consider feet-ground contact. iMapper Monszpart *et al.* (2019) maps RGB videos to dynamic "interaction snapshots", by learning "scenelets" from PiGraphs data and fitting them to videos. However, the estimated scene is not aligned with the 2D image, and consists of pre-defined CAD templates with fixed shape and size. Holistic++ Chen *et al.* (2019) takes learned 3D Human Object Interaction (HOI) into account, to reason about the arrangement of bodies and objects jointly. Both Monszpart *et al.* (2019) and Chen *et al.* (2019) do not model geometrically detailed human-scene interactions, due to their simplified representation of the scene and bodies. Weng and Yeung (2020) jointly optimize the reconstructed mesh-based 3D scene and bodies, which are initialized from Nie *et al.* (2020) and Pavlakos *et al.* (2019a). The approach only considers interpenetration between objects and the human, and does not model the explicit human-scene contact. Additionally, both Weng and Yeung (2020); Chen *et al.* (2019) do not model the coherence of human-scene interactions across frames from monocular video. In contrast to the prior work, our contribution lies in incorporating multiple human-scene interactions collectively, such that we can reconstruct a more accurate and consistent scene, with physically plausible human-scene interactions.

2.3 Generate Scenes from Humans

2.3.1 Generative Scene Synthesis (No People)

Most prior work on indoor scene synthesis focuses on generating scenes without considering human interactions. These methods are typically based on: (1) procedural model-

ing with grammars Müller *et al.* (2006); Parish and Müller (2001); Prakash *et al.* (2019); Talton *et al.* (2011); Kar *et al.* (2019); Devaranjan *et al.* (2020); Purkait *et al.* (2020), which employ recursive functions and formal grammar rules to model 3D structures like plants, buildings, cities, and indoor or outdoor scenes. For example, Talton *et al.* (2011) integrates reversible-jump MCMC to control the output of the stochastic context-free grammars, while Meta-sim Kar *et al.* (2019) learns models that modify scene graph attributes sampled from known probabilistic grammars to match real-world data. (2) graph neural networks Li *et al.* (2019a); Wang *et al.* (2019a); Zhou *et al.* (2019,?); Luo *et al.* (2020); Purkait *et al.* (2020); Zhang *et al.* (2020e,c); Keshavarzi *et al.* (2020); Di *et al.* (2020), which represent scenes as graphs. These models typically involve neural message passing or other graph-based techniques to predict relationships between objects. For instance, methods like Zhou *et al.* (2019) synthesize 3D scenes by generating parse trees, adjacency matrices, or scene hierarchies, requiring supervision in the form of relation graphs or scene structures.

(3) autoregressive neural networks Ritchie *et al.* (2019); Wang *et al.* (2018a), that sequentially generate scene elements. Ritchie *et al.* (2019) employs a CNN-based architecture to predict object attributes like category, location, orientation, and size in a specific order.

(4) transformers-based approaches Wang *et al.* (2021c); Paschalidou *et al.* (2021); Para *et al.* (2023), which utilize the powerful modeling capabilities of transformers for scene synthesis. SceneFormer Wang *et al.* (2021c) autoregressively adds objects to a scene, while ATISS Paschalidou *et al.* (2021) generates a 3D scene represented as an unordered set of objects using a permutation-invariant autoregressive transformer. Some works leverage lexical text Chang *et al.* (2015) or a sentence Chang *et al.* (2017b) as input to guide the 3D scene synthesis. Fisher *et al.* (2012) takes 3D scans as input and synthesizes the corresponding 3D object arrangements.

All methods mentioned above do not take human motion into consideration to guide the 3D scene synthesis. In contrast, we generate 3D scenes that are compatible with the humans defined by a given input motion. Specifically, the objects in the generated scene should support human motion (e.g., a chair or couch for sitting) and should not collide with the path of a walking human, To this end, we build upon the autoregressive scene synthesis architecture of ATISS Paschalidou *et al.* (2021) and incorporate contact and free-space information into the pipeline.

2.3.2 Human-aware Scene Generation

Qi *et al.* (2018) propose a method that synthesizes a 3D scene based on a human's affordance map together with a spatial And-Or graph. PiGraphs Savva *et al.* (2016) learns a probability distribution over human pose and object geometry from interactions. It does not model the lack of interaction, i.e. the free space carved out by movement. Similarly, recent methods Mura *et al.* (2021); Nie *et al.* (2022) explore how to estimate a 3D scene from human behaviors and interactions. Mura *et al.* (2021) predict the "3D floor plan"

from a 2D human walking trajectory in a deterministic way. The approach only indicates the room layout and furniture footprints and does not model objects or contact. Nie *et al.* (2022) propose Pose2Room, which predicts 3D objects inside a room from 3D human pose trajectories in a probabilistic way, by learning 3D object arrangement distribution. It only predicts contacted objects and can not generate objects in free space. In addition, it cannot take floor plans as input. We find these crucial in our experiments since object arrangements are highly related to the floor plan; some furniture is designed to go against a wall.

2.4 Generating Humans Motions from Scenes

2.4.1 Scene-aware Motion Generation

Motion synthesis in computer graphics has a rich history, encompassing areas such as locomotion Zhang and Tang (2022); Agrawal and van de Panne (2016); Lee *et al.* (2006); Kovar *et al.* (2023); Guzov *et al.* (2024), human-scene/object interaction Lee *et al.* (2002); Taheri *et al.* (2022); Zhang *et al.* (2024b), and dynamic object interaction Corona *et al.* (2020); Li *et al.* (2024b, 2023). We refer readers to an extensive survey Zhu *et al.* (2023) for an overview and focus on scene-aware motion generation in this section.

A particular challenge in modeling scene-aware motion is the lack of paired, high-quality human-scene datasets. One line of work Wang *et al.* (2021a,b) employs a two-stage method that first predicts the root path, followed by the full-body motion based on the scene and predicted path. However, these methods suffer from low-quality motion generation, attributed to the noise in the training datasets captured from monocular RGB-D videos Hassan *et al.* (2019). Neural State Machine (NSM) Starke *et al.* (2019) proposes the use of phase labeling Holden *et al.* (2017) and local expert networks Eigen *et al.* (2013); Jacobs *et al.* (1991); Yuksel *et al.* (2012) to generate high-quality object interactions, such as sitting and carrying, after training on a small human-object mocap dataset. Nonetheless, it struggles with recognizing walkable regions in 3D scenes, often failing to avoid obstacles. Therefore, later work in this vein requires using the A* algorithm for collision-free path planning Hassan *et al.* (2021b). These and related approaches Zhang *et al.* (2022b, 2024a) are moreover limited by the diversity of the small human-scene interaction datasets with no text annotations.

Various approaches ameliorate the data issue by creating synthetic data with captured Yi *et al.* (2023b); Ye *et al.* (2022) or generated Kulkarni *et al.* (2024) motions placed in scenes heuristically. HUMANISE Wang *et al.* (2022) improves this for text-conditioned scene interactions but relies entirely on short synthetic sequences for training, where the data generation heuristics used limit the realism. The reinforcement learning (RL) approach DIMOS Zhao *et al.* (2023) learns autoregressive policies to reach goal poses in a scene without requiring paired human-scene data for training, but still relies on A* and is constrained by the accuracy of goal pose generation Zhao *et al.* (2022). RL

with physical simulators Chao *et al.* (2021b); Peng *et al.* (2022); Hassan *et al.* (2023); Xiao *et al.* (2024) has been used to produce physically plausible movements but faces challenges in generalizing across varied scenes and objects.

Unlike most prior work, our approach is text-conditioned and leverages a mix of both scene-agnostic and paired human-scene data. Pre-training is done with a diverse scene-agnostic dataset, while scene-aware fine-tuning uses motion data with scene context. For training, we adopt both synthetic data creation with real motions and data augmentation of real-world human-object interactions Hassan *et al.* (2021b).

2.4.2 Diffusion-Based Motion Generation

Recently, diffusion models have demonstrated the ability to generate high-quality human motions, especially when conditioned on a text prompt Tevet *et al.* (2023); Zhang *et al.* (2022a); Petrovich *et al.* (2024); Li *et al.* (2024a). In addition to text, several diffusion models add spatial controllability. Some work Tevet *et al.* (2023); Shafir *et al.* (2023) adopt image inpainting techniques to incorporate dense trajectories of spatial joint constraints into generated motions. OmniControl Xie *et al.* (2024) and GMD Karunratanakul *et al.* (2023) allow control with sparse signals and a pre-defined root path, respectively.

A few diffusion works handle interactions with objects or scenes. TRACE Rempe *et al.* (2023) generates 2D trajectories for pedestrians based on a rasterized street map. SceneDiffuser Huang *et al.* (2023a) conditions generation on a full scanned scene point cloud, but motion quality is limited due to noisy training data Hassan *et al.* (2019). Another approach Pi *et al.* (2023) tackles single-object interactions through the hierarchical generation of milestone poses followed by dense motion, but it lacks text control. A concurrent line of work enables text conditioning for single-object interactions Diller and Dai (2024); Peng *et al.* (2023), but they focus on humans manipulating dynamic objects rather than interactions in full scenes.

We leverage a pre-trained text-to-motion diffusion model Tevet *et al.* (2023) and a fine-tuned scene-aware branch to enable both text controllability and scene-awareness with diffusion. We break motion generation into navigation and interaction with static objects by conditioning on 2D-floor maps and 3D geometry, respectively, and create specialized human-scene data to enable diversity and quality.

Chapter 3

Human-Aware Object Placement for Visual Environment Reconstruction

3.1 Introduction

Human behavior, and the interaction of humans with their environment, are fundamentally about the 3D world. Hence, 3D reconstruction of both the human and scene can facilitate human behavior analysis. Where and how humans interact with a scene can be used to predict future motions and interactions for human-centered AI and robots, or to synthesize these for AR/VR and other computer-graphics applications.

Tremendous progress has been made in reconstructing 3D human bodies Kocabas *et al.* (2021); Guler and Kokkinos (2019); Joo *et al.* (2020); Kanazawa *et al.* (2018); Kocabas *et al.* (2020); Kolotouros *et al.* (2019b); Pavlakos *et al.* (2019a,c); Yuan *et al.* (2022, 2021); Luo *et al.* (2021) and 3D scenes Huang *et al.* (2018a); Nie *et al.* (2020); Zhang *et al.* (2021a); Dahnert *et al.* (2021); Božič *et al.* (2021) from monocular images or videos, typically in isolation from each other. In real life, though, humans always interact with scenes. Consequently, humans (partially) occlude the scene, and the scene (partially) occludes humans. Strong human-scene occlusion can cause challenges for both scene and human reconstruction.

In contrast, recent work on human-scene interaction (HSI), estimates humans and scenes together Hassan *et al.* (2019); Chen *et al.* (2019); Weng and Yeung (2020). PROX Hassan *et al.* (2019) demonstrates how HSI can constrain 3D human pose estimation, but it requires a priori knowledge of a 3D scan of the full scene. This is often impractical and cumbersome, as it requires one to conduct offline 3D reconstruction by walking around the scene with a depth sensor Zollhöfer *et al.* (2018) to observe it from many viewpoints.

What we need, instead, is a method that estimates the scene and humans from images of a single color camera. This is challenging because the lack of depth information leads to inconsistencies in the scale and placement of objects relative to the interacting humans. Methods that reconstruct 3D humans from single views utilize statistical body models Joo *et al.* (2018); Loper *et al.* (2015); Pavlakos *et al.* (2019a); Xu *et al.* (2020) as priors for body shape and pose. However, such tools do not exist for the collective space

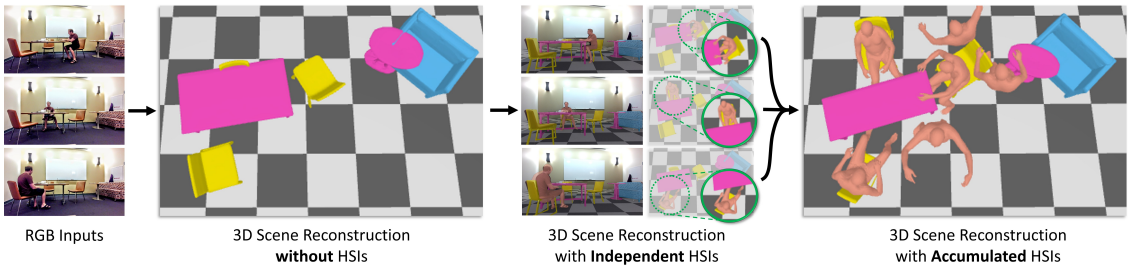


Figure 3.1: From a monocular video sequence, MOVER reconstructs a 3D scene that best affords humans interacting with it. Existing methods for monocular 3D scene reconstruction ignore people and produce non-functional scenes. MOVER takes as input: (1) several images of human-scene interaction (HSI) from a static camera, (2) a rough estimate of 3D object shape and placement in 3D space Nie *et al.* (2020), and (3) estimated 3D human bodies interacting with the scene Pavlakos *et al.* (2019a); Kocabas *et al.* (2021). Each frame contains valuable information about humans, objects, and the proximal relationship between them. MOVER accumulates this information across frames, to optimize for a physically plausible and functional 3D scene. The final 3D scene is more accurate than the input and enables reasoning about human-scene contact.

of 3D scene layouts. This is due to the enormous space of possible object arrangements in indoor 3D scenes, the large number of different object classes, and the huge inter-class (e.g., chairs and desks) and intra-class (e.g., desk chair and club chair) shape variability.

To address the above issues, we present MOVER, which stands for "human Motion driven Object placement for Visual Environment Reconstruction". MOVER leverages information across several HSI frames to estimate both a plausible 3D scene and a moving human that interacts with the scene. Figure 3.1 provides a high-level overview of the approach. MOVER takes as input: (1) a set of color frames from a static monocular camera, (2) a 3D human mesh inferred for each frame Pavlakos *et al.* (2019a); Kocabas *et al.* (2021), and (3) a 3D shape inferred for each object detected in the scene Nie *et al.* (2020); Kirillov *et al.* (2020). As output, MOVER produces a refined 3D scene, consisting of repositioned input objects, ensuring consistency with the estimated 3D human; that is, it satisfies the expected contacts on the body Hassan *et al.* (2021a) while preventing inter-penetration. MOVER employs a novel optimization scheme that jointly optimizes over camera pose, ground-plane pose, and the size and position of 3D objects, subject to various HSI constraints.

MOVER considers three types of HSI constraints into account: (1) humans that move in a scene are occluded or occlude objects, thus, defining the depth ordering of the objects, (2) humans move in free space that is not occupied by objects and do not interpenetrate objects, (3) contact between humans and objects means that the contacting parts of their surfaces occupy the same place in space. Thus, we leverage both explicit (i.e., contact) and implicit (i.e., free space, no penetrations) HSI cues. MOVER is able to use

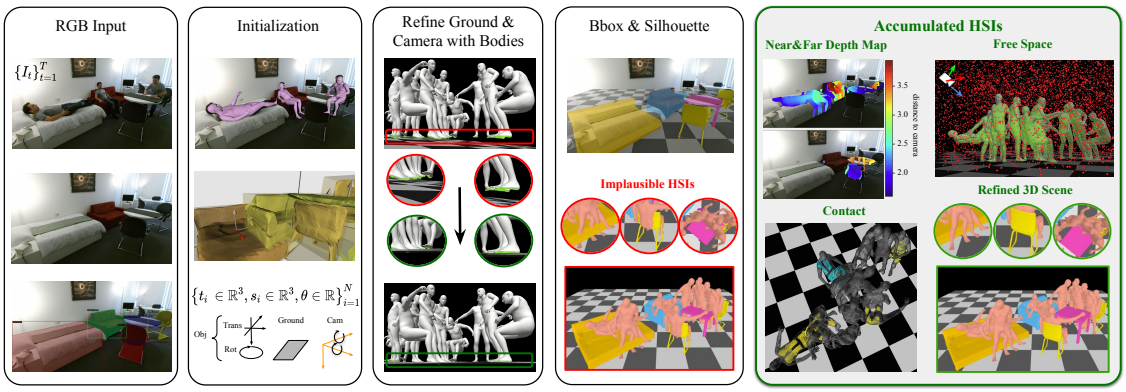


Figure 3.2: Overview of MOVER. Given a video or multiple images, the initialization involves using Nie *et al.* (2020) to reconstruct a 3D scene from labeled or detected 2D instance segmentation masks Kirillov *et al.* (2020), estimating the 3D human poses and shape Pavlakos *et al.* (2019a); Kocabas *et al.* (2021), and extracting the expected contact vertices on the estimated bodies using POSA Hassan *et al.* (2021a). The first step then refines the camera orientation and ground plane using the human bodies and their foot contact. Then we optimize the object layout based on 2D bounding boxes and silhouettes to remove interpenetration between people and objects, e.g., the human sits through the chair, stands into a table, and the legs are in a bed. Finally, incorporating multiple HSIs collectively from the whole video, we can improve the 3D scene further such that the bodies perform more realistic scene interaction.

these because it employs detailed meshes for both the scene and the moving human. In contrast, previous attempts toward similar goals either use oversimplified shapes Chen *et al.* (2019), such as 3D bounding boxes for objects and skeletons for humans, work only for static humans that contact a single object Zhang *et al.* (2020a), or do not integrate information across several interaction frames Weng and Yeung (2020); Zhang *et al.* (2020a); Chen *et al.* (2019).

Comparisons with the state of the art on the PROX Hassan *et al.* (2019) and PiGraphs Savva *et al.* (2016) datasets demonstrate that MOVER estimates more accurate and realistic 3D scene layouts, satisfying expected contacts while minimizing penetrations relative to the moving humans. Interestingly, we find that MOVER's estimated 3D scene can refine human poses using a PROX-like method Hassan *et al.* (2019). Although estimating 3D scenes and 3D humans from a monocular camera is challenging, our results indicate that these tasks are synergistic and benefit from each other.

3.2 Method

MOVER is an optimization-based approach that reconstructs a physically plausible 3D scene that is consistent with predicted human-scene interactions over time (see Fig-

ure 3.2). Specifically, our method takes a single RGB video or multiple images $\{I_t\}_{t=1}^{T}$ as input and reconstructs the human bodies at each time step t along with the numerous static scene objects, all residing in a common 3D space supported by a ground plane. In our experiments, we focus on large objects in indoor scenes that humans frequently interact with, such as, chairs, beds, sofas, and tables.

We initialize our approach using separate estimates for the 3D human poses Kocabas *et al.* (2021); Pavlakos *et al.* (2019a), the 3D scene Nie *et al.* (2020), and the ground plane. Using the estimated body poses, we predict contact vertices $\mathcal{C}$ for all bodies using POSA Hassan *et al.* (2021a), which predicts likely contact vertices on the body conditioned on pose. We further categorize these vertices into foot contacts $\mathcal{C}^{\text{feet}}$ and other body part contacts $\mathcal{C}^{\text{body}}$. The explicit foot contact points $\mathcal{C}^{\text{feet}}$ serve as constraints to refine the camera orientation and ground plane prediction. Based on this initialization, we optimize object alignment by minimizing an objective function incorporating multiple human-scene interactions (HSIs) across the entire input data.

3.2.1 3D Scene Layout Optimization

Our method leverages multiple HSIs to refine the 3D scene. Recall that these HSIs provide the following constraints: (1) humans that move in a scene are occluded or occlude objects, thus, defining the relative depth ordering of the objects (depth order constraint), (2) humans move through free space and do not interpenetrate objects (collision constraint), (3) when humans and objects are in contact, the contact surfaces occupy the same place in space (contact constraint). Using these constraints, our objective $L_{\text{scene-human}}$ is:

$$\mathcal{L}_{\text{scene-human}} = \lambda_1 \mathcal{L}_{\text{bbox}} + \lambda_2 \mathcal{L}_{\text{occ-sil}} + \lambda_3 L_{\text{scale}} + \lambda_4 \mathcal{L}_{\text{depth}} + \lambda_5 \mathcal{L}_{\text{collision}} + \lambda_6 \mathcal{L}_{\text{contact}}. \quad (3.1)$$

We apply an occlusion-aware silhouette term $\mathcal{L}_{\text{occ-sil}}$ from Zhang *et al.* (2020a), a 2D bounding box projection term $\mathcal{L}_{\text{bbox}}$ that constrains the top-left corner and the width of the bounding boxes of the objects, and L_{scale}, an ℓ_2-based regularizer to constraint the variation of the object scales.

The *2D bounding box term* $\mathcal{L}_{\text{bbox}}$ is an ℓ_1 norm between the object's projected 3D bounding box $Proj_i$ and its corresponding detected 2D bounding box Det_i, expressed with the top-left corner coordinate x_{min}, y_{min} and $width$ value.

$$\mathcal{L}_{\text{bbox}} = \sum_i \| Proj_i^{\alpha} - Det_i^{\alpha} \|, \quad \alpha \in \{x_{min}, y_{min}, width\}. \quad (3.2)$$

The *scale term* prevents object scales s deviating far from the initial estimates s^{init}

from Total3D Nie *et al.* (2020):

$$\mathcal{L}_{\text{scale}} = \sum_i \| \frac{s_i}{s_i^{init}} - 1.0 \|_2. \tag{3.3}$$

Depth Order Constraint $\mathcal{L}_{\text{depth}}$. The occlusion between humans and objects can provide clues about the object's depth. We assume that the human's depth is accurate. If a human occludes an object, the far side of the person sets a limit on how close the object can be. Alternatively, if the object occludes the person, the visible side of the person sets a maximum distance for the object. This is summarized in Figure 3.3. In this way, human-object occlusion provides constraints on scene layout even when there is no human-object contact. Directly applying the ordinal depth loss proposed by Jiang et al. Jiang *et al.* (2020) for each image is inefficient, as the required memory increases with the number of images. In contrast, we accumulate all single depth ordering maps into one far depth range map $\hat{D}_{\text{far}}$ and one near depth range map $\hat{D}_{\text{near}}$ as follows:

$$\begin{aligned} \hat{D}_{\text{far}}(p) &= \min\left(D^1_{\text{far}}(p), ..., D^T_{\text{far}}(p)\right), \\ \hat{D}_{\text{near}}(p) &= \max\left(D^1_{\text{near}}(p), ..., D^T_{\text{near}}(p)\right), \end{aligned} \tag{3.4}$$

where the pixel p is in the overlapping region between the human bodies and the objects. Using these accumulated depth range maps, we constrain the depth $D_i(q)$ of a projected pixel q from object i to lie within the corresponding range:

$$\begin{aligned} \mathcal{L}_{\text{depth}} = \sum_i \sum_{q \in Sil_i \cap M_i} [&\text{ReLU}(D_i(q) - \hat{D}_{\text{far}}(q)) \\ &+ \text{ReLU}(\hat{D}_{\text{near}}(q) - D_i(q))], \end{aligned} \tag{3.5}$$

where Sil_i is the rendered silhouette of the object i, M_i is the 2D segmentation mask of it, and $D_i(q)$ is the depth of the object i at the pixel q.

Collision Constraint $\mathcal{L}_{\text{collision}}$. To penalize interpenetrating vertices of objects and bodies in the scene, we use the signed distance field (SDF) of all reconstructed bodies. Specifically, we calculate a signed distance field volume V_j for each body j in a shared 3D world space, and accumulate them into a global SDF volume as $\hat{V} = \min(V_1, ..., V_j, ...)$, This ensures that, at any point in space, the global SDF reflects the smallest signed distance value among the individual volumes. The SDF $\hat{V}$ is stored in a volumetric grid of size 256^3, which spans a padded bounding box of all bodies. For a vertex u_i of an object O_i, we compute the voxel coordinate $f(u_i) = (p(u_i), q(u_i), k(u_i))$ in the global SDF volume, where p, q, k denote the grid indices, and retrieve the corresponding SDF value $\hat{V}_{f(u_i)}$. Based on the SDF values of all vertices of all N objects, we resolve scene-body

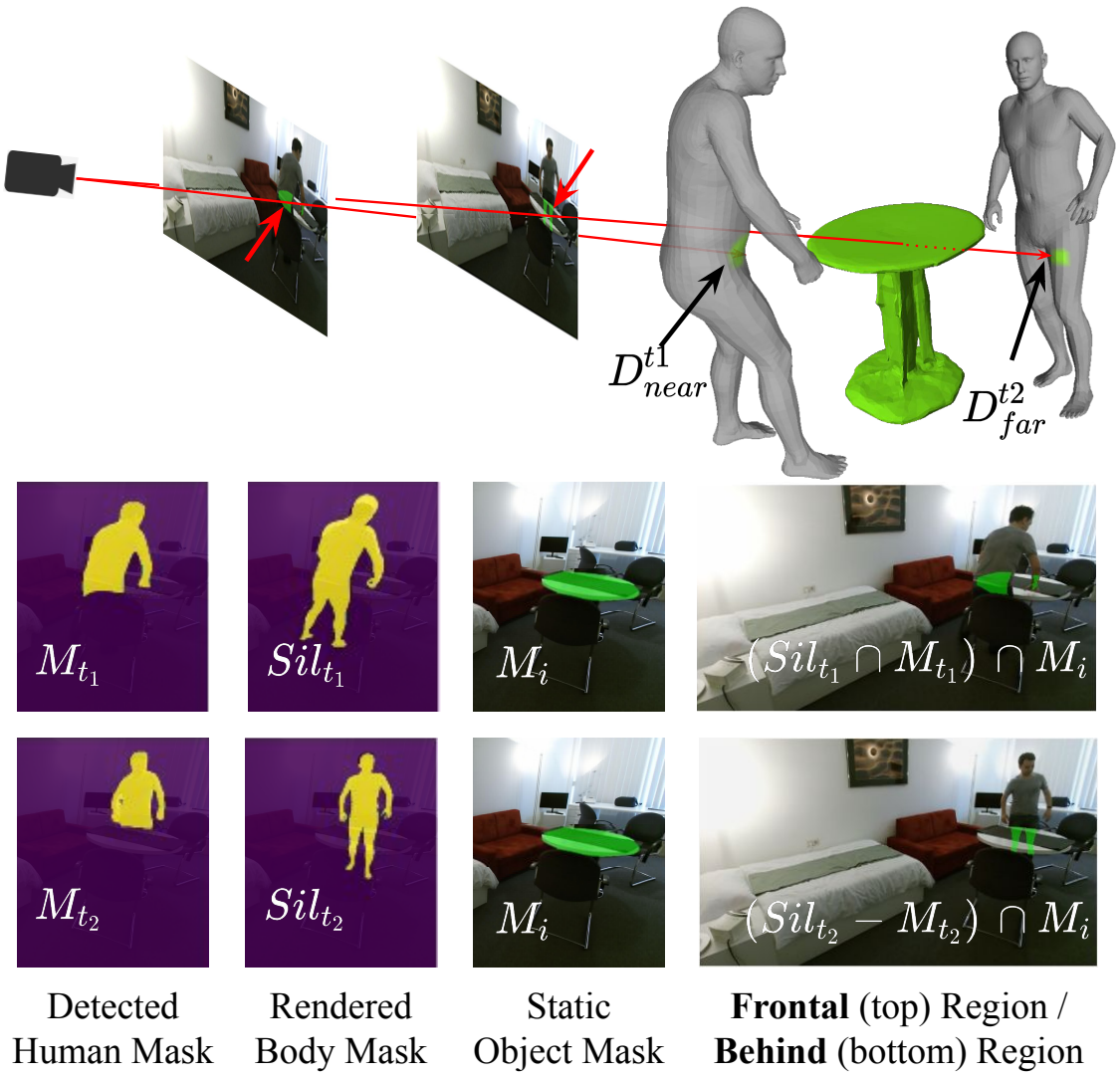


Figure 3.3: Computing depth range maps for the depth order constraint $\mathcal{L}_{\text{depth}}$. Given a detected human mask M_t and a rendered body mask Sil_t, for each object i, we compute the overlap region between M_i and $Sil_t \cap M_t$ as the frontal region and extract the depth of the backside surface of the body as the near depth range D^t_{near} of the object i. Similarly, we compute $(Sil_t - M_t) \cap M_i$, which defines the far depth range D^t_{far} of the object.

inter-penetration by penalizing vertices with a negative SDF value:

$$\mathcal{L}_{\text{collision}} = \sum_i \sum_u \|\hat{V}_{f(u_i)}\|_2^2, \quad \hat{V}_{f(u_i)} < 0. \tag{3.6}$$

Contact Constraint $\mathcal{L}_{\text{contact}}$. When humans and objects are in contact, their contact surfaces occupy the same place. We propose a contact constraint to minimize the distance between the contacted body parts and their corresponding assigned contacted object. PHOSA Zhang *et al.* (2020a) proposes a loss in which they assign a whole body to only one object, whereas humans sometimes interact with multiple objects; for example, a person may sit on a chair while placing their hand on a table. In contrast, we directly assign the contacted body vertices $\mathcal{C}_i^{body}$ of each body to different objects, based on the overlap between the 2D projection of the vertices and the detected object masks, and the 3D distances between them. We consider the vertices of sofa and chair backs and seat bottoms as contactable regions. We minimize the distance between the contacted bodies and the contacted object parts as follows:

$$\mathcal{L}_{\text{contact}} = \sum_i \sum_{v \in \mathcal{C}^{\text{body}}} \mathbb{I}(v, O_i)[\text{CD}(v^y, \mathcal{C}(O_i)^y) + \text{CD}(v^{\perp y}, \mathcal{C}(O_i)^{\perp y})], \tag{3.7}$$

where $\mathcal{C}(O_i)^{\perp y}$ and $\mathcal{C}(O_i)^y$ denote the back and the bottom seat contact part of an object i, respectively. $\mathbb{I}(v, O_i)$ is an indicator function (1 only if the contact vertex v is assigned to the contacted object O_i, 0 otherwise). CD denotes the one-directional ChamferDistance (CD), i.e., from bodies to objects, because for large furniture like a bed or a sofa, a human typically contacts only a small region of the object. In contrast, PHOSA Zhang *et al.* (2020a) uses a bi-directional CD, which tends to shrink the object to match the contacted body parts.

3.2.2 Optimization

We optimize Equation (3.1) for a specific scene with respect to the parameters $\mathbf{s}_i$ (scale), θ_i (rotation), $\mathbf{t}_i$ (translation) of the objects $\{i = 1...N\}$ using the Adam optimizer Kingma and Ba (2014). In the following, we detail the initialization of the 3D scene and the human poses.

Initial 3D Scene. We extract a representative 2D image $\mathbf{I}$ from the input data that contains no human-object occlusion. For this image, we label or compute 2D bounding boxes B_i and an instance masks M_i for all N objects in the scene using PointRend Kirillov *et al.* (2020). We use Nie *et al.* (2020) to get an initial 3D scene $\mathbf{S}_0$, consisting of a ground plane $y = y_p$ and multiple object meshes $\{O_i\}_{i=1}^N$, and a perspective camera with *yaw* and *pitch* orientation. Each object i has a translation $\mathbf{t}_i \in \mathbb{R}^3$, scale $\mathbf{s}_i \in \mathbb{R}^3$, and a rotation along the y-axis $\theta_i^y \in [0, 2\pi)$. Since the predicted meshes of Nie *et al.* (2020) are

incomplete and have holes, we use Occupancy Networks Mescheder *et al.* (2019) and Marching Cubes Lorensen and Cline (1987) to transform each object mesh into a watertight mesh. Based on this preparation, we first optimize the objective function $\mathcal{L}_{\text{scene}}$ without considering the HSIs:

$$\mathcal{L}_{\text{scene}} = \mathcal{L}_{\text{occ-sil}} + \lambda_1 \mathcal{L}_{\text{bbox}} + \lambda_2 \mathcal{L}_{\text{scale}}. \tag{3.8}$$

Initialization of the Ground and the Camera. As shown in the third column of Figure 3.2, the estimated ground plane and camera orientation from Nie *et al.* (2020) violates the reconstructed bodies (e.g., people float in the air). Previous methods either fix the camera orientation and only optimize the ground plane and humans Chen *et al.* (2019), or estimate them independently per image Weng and Yeung (2020), which generates inconsistent camera orientation and ground planes throughout a video. However, the camera orientation and ground plane are essential for producing plausible HSIs. Thus, we jointly estimate the ground, camera and multiple humans together, by applying $\mathcal{L}_{\text{feet}}$:

$$\mathcal{L}_{\text{feet}}(R,p) = \rho_1(R^\top \sum_t \mathcal{C}_t^{\text{feet}} - [0, y_p, 0]^\top), \tag{3.9}$$

where R is the camera rotation matrix calculated from *pitch, yaw*, and ρ denotes a robust Geman-McClure error function Comer *et al.* (2010) for down-weighting outliers.

Initial Estimate of 3D Bodies. As an initial shape and body pose estimate for the input images $\{I_t\}_{t=1}^T$, where a human interacts with a 3D scene, we use OpenPose Cao *et al.* (2021) and SMPLify-X Pavlakos *et al.* (2019a). Specifically, we use a perspective camera projection and estimate the pose parameters θ_t of SMPL-X for each frame with shared body shape parameters β. SMPLify-X requires a good initialization and, for this, we use PARE Kocabas *et al.* (2021) because it is robust to occlusion and our scenes involve significant occlusion. PARE outputs SMPL, which we convert to SMPL-X, and use the resulting 3D joints to initialize SMPLify-X.

We then optimize all SMPL-X parameters to minimize an objective function E_{Body} of multiple terms, as described in SMPLify-X Pavlakos *et al.* (2019a) (see $E_{\text{SMPLify-X}}$) :

$$E_{\text{Body}} = \sum_{t=1}^{T} \left(E_{\text{SMPLify-X}}(t) \right) + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}}. \tag{3.10}$$

To reduce the jitter, we add a constant-velocity motion smoothing term on 3D joints J and their 2D projections J^{Proj}:

$$\mathcal{L}_{\text{smooth}} = \sum_{t=1}^{T-1} \rho_2 \left(\| J_{t-1} + J_{t+1} - 2 \times J_t \| \right) + \rho_3 \left(\| J_{t-1}^{\text{Proj}} + J_{t+1}^{\text{Proj}} - 2 \times J_t^{\text{Proj}} \| \right). \tag{3.11}$$

We also apply the constant-velocity assumption to avoid noisy and unreliable body poses

and, therefore, wrong human-scene interactions during optimization. We calculate the pelvis acceleration ν_t and local joints' acceleration α_t of a person in frame t to describe the global body translation and local pose articulations of the body. We filter out those bodies with either large pelvis translation or incorrect human pose with a large ν or a large α respectively: $\{j : \nu_j < \tau_{\text{pelvis}} \cap \alpha_j < \tau_{\text{local}}, j \in \{1...T\}\}$, where τ_{pel}, τ_{local} are the thresholds for the pelvis acceleration and the local pose acceleration, respectively.

3.3 Datasets

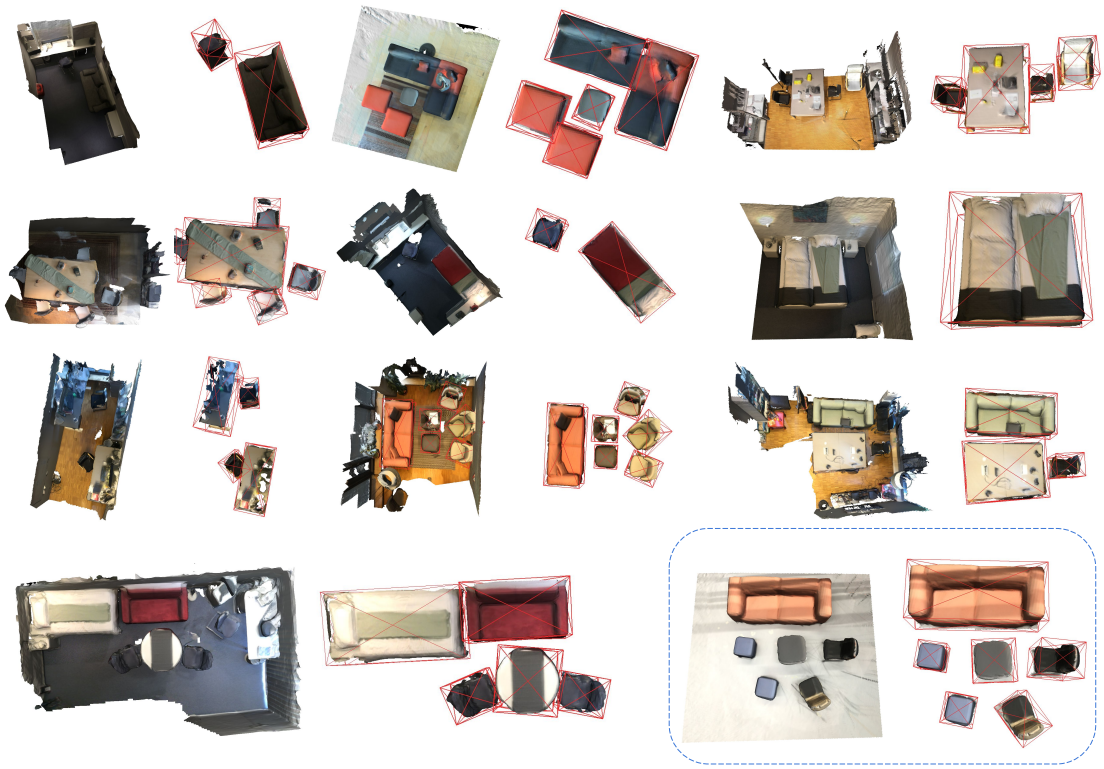
PiGraphs. PiGraphs Savva *et al.* (2016) consists of 60 RGB-D videos of 30 scenes. The dataset is recorded with a *Microsoft Kinect One*, and is designed to capture human and object arrangements in different kinds of interactions. Each video recording is approximately 2 minutes long with 5 fps. It contains labeled 3D bounding boxes of objects in the scene and human poses represented as 3D skeletons. We use this dataset to evaluate the scene reconstruction and compare with Nie *et al.* (2020); Weng and Yeung (2020). Note that the provided human poses are noisy and not suitable for an evaluation of 3D human shape and pose estimation.

PROX Qualitative. PROX *qualitative* contains 61 RGB-D videos at 30 fps of human motion/interaction in 12 scanned static 3D scenes. The data has been recorded using the *Microsoft Kinect One* and *StructureIO* sensor. To enable 3D scene reconstruction evaluation on this dataset, we segment and label each object with its 3D bounding box. Since there are two scenes (i.e., "BasementSittingBooth" and "N0SittingBooth") containing an inseparable object (see Figure 3.4) which is challenging to segment out from the entire 3D scan, we evaluate all methods on the remaining 10 scenes using the corresponding 51 videos as input.

PROX Quantitative. PROX *quantitative* captures a sequence of human-scene interaction RGB-D frames within a synchronized *Vicon* marker-based motion capturing system. In total, the dataset contains 178 frames and provides ground truth body meshes, which are useful for human pose and shape (HPS) evaluation. For fair evaluation on HPS, we input all images into HolisticMesh Weng and Yeung (2020) and ours to get a refined scene and use a refined scene to get refined bodies. In addition, we also label this scene for 3D scene reconstruction evaluation, see Figure 3.4.

3.4 Implementation Details

Contact Regions of Objects. We automatically calculate the contact regions of objects based on vertices normal. Specifically, vertices with normals along the y-axis represent



(A) PROX *qualitative* dataset (B) PROX *quantitative* dataset

Figure 3.4: We crop out each object separately and label the corresponding 3D bounding box for 10 scenes in PROX *qualitative* dataset and one scene in PROX *quantitative* dataset.

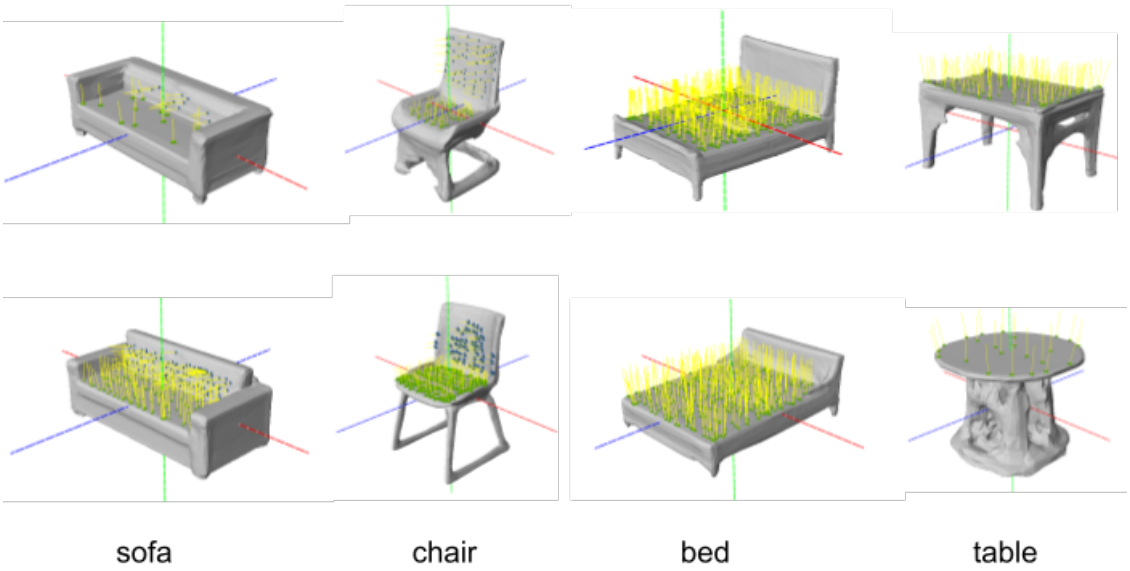


Figure 3.5: Contact regions of different objects. The red, green, and blue axes represent the x, y, and z coordinates, respectively, while the yellow lines indicate the normal direction at each point.

the bottom or top parts of the objects, while vertices with normals along the z-axis represent the back parts of the objects. We define sofas and chairs as having two contact regions (bottom and back parts), while beds and tables have only a top part as the contact region, shown in Figure 3.5.

Since PARE Kocabas *et al.* (2021) uses SMPL Loper *et al.* (2015) model, we use Pavlakos *et al.* (2019b) to transfer it to the SMPL-X Pavlakos *et al.* (2019a) model.

Optimization. We use the Adam optimizer Joo *et al.* (2018) to optimize the final energy term with a step size of 0.002 and 3000 iterations. We set $\lambda_1, \lambda_2, \lambda_3$ to $1000, 0.3, 1000$, respectively, for 2D bounding box term, occlusion-aware term and scale term. The weights of our proposed depth order constraint, collision constraint, and contact constraint are set to $\lambda_4 = 8, \lambda_5 = 1000$, and $\lambda_6 = 1e5$, respectively. We use two robust Geman-McClure error functions, ρ_1, ρ_2 with parameter 0.1 on 3D joints, and one ρ_3 with parameter 100 on 2D projection of 3D joints.

Our method takes around 30 minutes for 3000 iterations to optimize a 3D scene with accumulated HSI constraints. In comparison, HolisticMesh Weng and Yeung (2020), which jointly optimizes human and a 3D scene for one single image, directly trains the parameters of the network in Total3DUnderstanding Nie *et al.* (2020) to regress the 3D scene, which is time-consuming and takes around 40 minutes. For the human optimization, it runs twice (5 minutes), i.e., one is a HPS initialization used to refine the scenes, and the second pass is done using the refined scenes. In total, HolisticMesh takes 45

Methods	Setting					Scene Recon.			HSI	
	BBOX&Mask	Cam.	Contact	Depth	Colli.	IoU_{3D} ↑	P2S↓	IoU_{2D} ↑	Non-Col ↑	Cont. ↑
HolisticMesh Weng and Yeung (2020)	PointRend					0.211	0.410	0.648	0.990	0.369
Total3D Nie *et al.* (2020)	PointRend					0.246	0.319	0.522	0.974	0.510
Ours	PointRend	✓	✓	✓	✓	0.309	0.221	0.777	0.977	0.612
HolisticMesh Weng and Yeung (2020)	2D GT					0.267	0.237	0.745	0.988	0.491
Total3D Nie *et al.* (2020)	2D GT					0.196	0.369	0.227	0.963	0.440
Ours	2D GT	✓	✓	✓	✓	0.383	0.199	0.898	0.986	0.673
Ablation Study	2D GT	✓				0.374	0.206	0.859	0.979	0.738
		✓	✓	✓		0.389	0.199	0.904	0.983	0.697
		✓	✓			0.381	0.205	0.904	0.980	0.773
		✓		✓		0.393	0.194	0.907	0.983	0.638
		✓			✓	0.383	0.199	0.903	0.984	0.674

Table 3.1: Quantitative results for 3D scene understanding (3D object detection) and human-scene interaction on the PROX *qualitative* dataset. P2S, Non-Col and Cont denote *point2surface distance*, Non-Collision and Contactness respectively. In each column, red is the best result among methods that take 2D labeled masks as input; blue is the second best.

minutes for one single image. Our method takes approximately the same time for a scene (around 10 objects) regardless of the number of frames in the input video. The number of frames in a video only affects the time required to calculate the depth map, the SDF volume, and the contact information of each body. However, these calculations can be performed once and are easily parallelized before optimization. In contrast, HolisticMesh Weng and Yeung (2020) processes a video sequentially, i.e., one frame after another. Therefore, the optimization time increases with the number of frames in a video.

3.5 Experiments

To evaluate the influence of accumulated HSIs on the optimized 3D scene layout, we use two datasets, PiGraphs Savva *et al.* (2016) and PROX Hassan *et al.* (2019). In comparison to Nie *et al.* (2020) and Weng and Yeung (2020), we achieve state-of-the-art 3D scene layout reconstruction, both quantitatively (see Section 3.5.1) and qualitatively (see Section 3.5.3). On the PROX *quantitative* dataset, we find that our 3D scene reconstructions result in more accurate human shape and pose estimations than our baselines. In Section 3.5.2, we analyze the contributions of different energy terms to our final results. Qualitative results are shown in Figure 3.6.

3.5.1 Quantitative Analysis

We conduct several experiments to investigate the effectiveness of our proposed method in three areas: 3D scene reconstruction, human-scene interaction (HSI) reconstruction, and human pose and shape (HPS) estimation. The results are listed in Table 3.1.

Methods	IoU_{2D} ↑	IoU_{3D} ↑
Cooperative Huang *et al.* (2018a)	68.6	21.4
Holistic++ Chen *et al.* (2019)	75.1	24.9
HolisticMesh Weng and Yeung (2020)	75.6	26.3
Ours	**79.2**	**27.8**

Table 3.2: Quantitative results for 3D scene understanding (3D object detection) on *PiGraphs* dataset Savva *et al.* (2016).

	Cam. Orien.			Ground Pen	
Methods	pitch ↓	roll ↓	mean ↓	Freq. ↓	Dist. ↓
Total3D Nie *et al.* (2020)	0.059	**0.031**	0.045	0.316	0.167
Ours	**0.042**	0.034	**0.038**	**0.100**	**0.112**

Table 3.3: Errors in the camera orientation and the ground penetration using foot contact on the PROX *qualitative* dataset.

3D Scene Reconstruction. Following Nie *et al.* (2020); Huang *et al.* (2018b); Chen *et al.* (2019); Weng and Yeung (2020), we compute the 3D Intersection over Union (IoU) and 2D IoU of object bounding boxes to evaluate the 3D scene reconstruction and the consistency between the 3D world and 2D image on PROX and PiGraphs. However, the 3D IoU is coarse and does not capture the error in an object's orientation, which is crucial for physically plausible HSI, e.g., a human can not sit on an armed chair with the wrong orientation. Therefore, we introduce the *point2surface distance* (P2S) to measure the distance from a cropped object mesh to the estimated 3D object mesh. It enables 3D scene reconstruction evaluation with more geometric details including orientation and shape. Given 2D labeled or detected Kirillov *et al.* (2020) bounding boxes and masks, our method improves upon the input from Nie *et al.* (2020) and outperforms Weng and Yeung (2020) on all scene-reconstruction metrics across different datasets, as shown in Table 3.1 and Table 3.2.

Additionally, we also evaluate the error in camera orientation and ground plane penetration Rempe *et al.* (2021) using the estimated foot contact vertices (see Table 3.3). We find that jointly optimizing the camera orientation and the ground plane using foot contact significantly improves accuracy compared to the initial estimate from Nie *et al.* (2020).

Human-scene Interaction Reconstruction. To evaluate the estimated HSI or functionality of the scene (i.e., how well the estimated scene supports human motion), we compute the metrics as Zhang *et al.* (2020d,b); Hassan *et al.* (2021a). Specifically, for each reconstructed body and 3D scene, we calculate (1) the *non-collision score* which

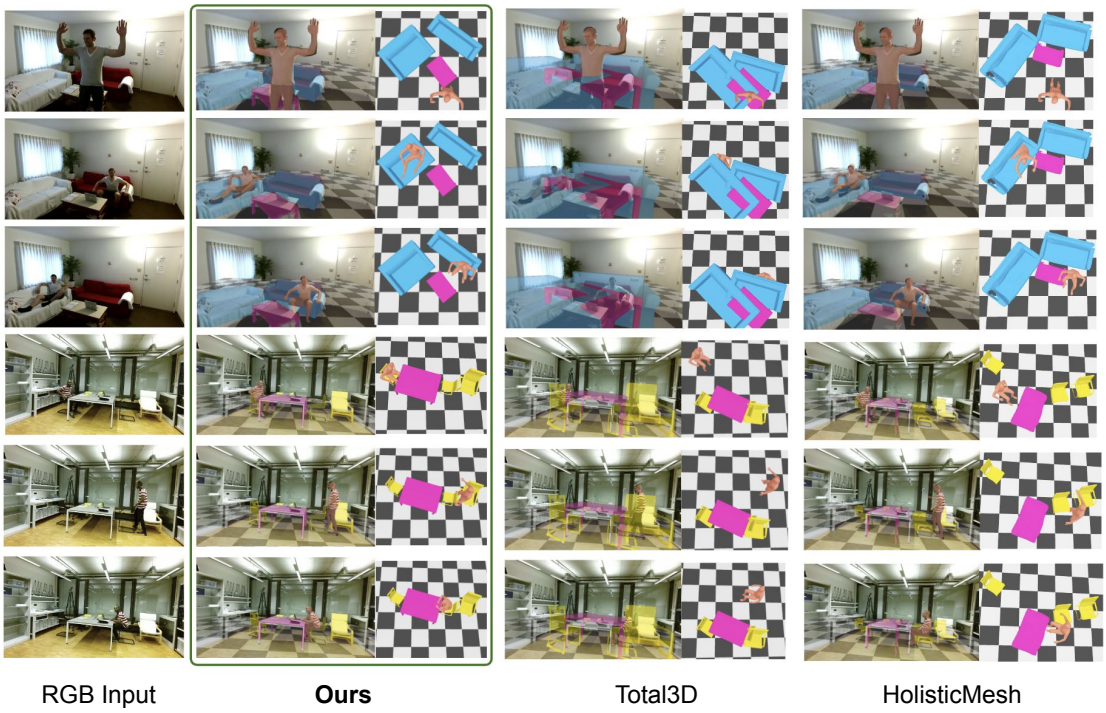


Figure 3.6: Qualitative results on PiGraphs (top) and PROX. Our method recovers better 3D scenes and HPS, which supports more plausible HSIs, compared with our baseline Nie *et al.* (2020) (Separated Composition) and another single-image baseline (Sequentially Joint Optimize) Weng and Yeung (2020).

measures the ratio of body mesh vertices without collision with the estimated 3D scene to the total number of body mesh vertices, and (2) the *contact score*, which indicates whether the body is in contact with the 3D scene. The *contact score* is 1 if at least one vertex of a body interpenetrates with the 3D scene. We report the mean *non-collision score* and mean *contact scores* among all videos and all bodies. In Table 3.1, MOVER (ours) achieves the best balance between non-collision and contact scores.

The estimated scenes with detected 2D boxes and masks Kirillov *et al.* (2020) yield lower HSI scores than with 2D GT. It is primarily due to the mis-detected objects from Kirillov *et al.* (2020). Since the reconstructed scenes of Weng and Yeung (2020) do not support human-scene contact well, e.g., a sitting body often floats, due to the lack of explicit human-scene contact modeling, it has a better non-collision score but a lower contact score.

More Evaluation Results on PROX Quantitative Dataset. We also evaluate 3D scene reconstruction and human-scene interaction on the PROX *quantitative* dataset, as shown in Table 3.4. Our method improves our input baseline Nie *et al.* (2020) significantly and outperforms the previous method Weng and Yeung (2020) by a large margin in both 3D

Methods	Scene Recon.			HSI	
	IoU_{3D} ↑	P2S↓	IoU_{2D} ↑	Non-Col ↑	Cont. ↑
HolisticMesh Weng and Yeung (2020)	0.239	0.133	0.533	0.948	0.951
Total3D Nie *et al.* (2020)	0.063	0.409	0.342	0.940	0.436
Ours	**0.390**	**0.095**	**0.862**	**0.972**	**0.934**

Table 3.4: Quantitative results for 3D scene understanding (3D object detection) and human-scene interaction on the PROX *quantitative* dataset. P2S, Non-Col and Cont denote *point2surface distance*, Non-Collision and Contactness respectively.

scene reconstruction metrics and human-scene interaction metrics.

Human Pose and Shape (HPS) Estimation. Can we use the estimated 3D scene to, in turn, improve 3D HPS? Here, we follow PROX but replace the scanned 3D scene of PROX with our estimated 3D scene. In Table 3.5, we evaluate the HPS estimation on PROX *quantitative* using the same metrics as Hassan *et al.* (2019). Specifically, we report (1) the mean per-joint error (PJE) and (2) the mean vertex-to-vertex distance (V2V). For completeness, we also compute these metrics on the Procrustes-aligned predictions (denoted as p.PJE and p.V2V, respectively). But note that the metrics w./o. Procrustes alignment (PJE and V2V) are more meaningful, since we want to evaluate the translation, rotation and scaling of the human body. As shown in Table 3.5, with estimated camera orientation and ground plane constraints (+CamGP), the PJE and V2V are both improved by a significant margin $+43.21$ and $+42.41$ respectively, with respect to our baseline. We also see that our refined scene can further refine our estimated bodies by applying the SDF loss (+SDF) and the contact loss (+Contact) from Hassan *et al.* (2019). Our final body estimation outperforms HolisticMesh Weng and Yeung (2020) and is similar to PROX, without having access to a scanned 3D scanned scene.

3.5.2 Ablation Study

To analyze the contribution of the accumulated HSIs and the influences of the different constraints, we conducted multiple ablation studies; see Table 3.1. All three proposed HSI constraints (depth, contact, and collision) help improve 3D scene reconstruction in different ways. The *contact* constraint produces the human-scene contact scores, but decreases the non-collision score. The *collision* and *depth* both contribute to the non-collision score. However, using only the *depth* achieves a slightly better 3D scene evaluation than our full model, but leads to worse human-scene contact scores. By applying all constraints, our method can generate a more accurate 3D scene, which supports more physically plausible HSI.

With **G.T Captured 3D Scene Scans**				
Methods	PJE↓	V2V ↓	p.PJE ↓	p.V2V↓
RGB Hassan *et al.* (2019)	220.27	218.06	73.24	60.80
PROX Hassan *et al.* (2019)	167.08	166.51	71.97	61.14
With **Image2Mesh Models**				
HolisticMesh Weng and Yeung (2020)	190.78	192.21	**72.72**	**61.01**
baseline*	219.62	222.50	75.92	68.34
+CamGP	176.41	180.09	73.41	67.33
+CamGP+SDF	175.98	179.98	73.96	68.29
Ours	**174.37**	**178.31**	73.60	67.89

Table 3.5: Quantitative results for human pose estimation on PROX *quantitative* dataset (baseline*: batch-wise SMPLify-X, **Ours**: +CamGP+SDF+Contact.)

3.5.3 Qualitative Analysis

In Figure 3.6, we show reconstructed 3D scenes and humans along with RGB videos, to demonstrate the effectiveness and generality of our approach on different datasets (PROX and PiGraphs). MOVER recovers better 3D scenes and HPS compared to our baseline Nie *et al.* (2020) (Separated Composition) and another single-image baseline Weng and Yeung (2020) (Sequentially Joint Optimize).

In Figure 3.7 and Figure 3.8, we present additional qualitative results on the PROX Hassan *et al.* (2019) qualitative dataset and the PiGraphs Savva *et al.* (2016) dataset, respectively. As can be seen, our method performs well across a variety of scenes and predicts physically plausible and functional scene layouts.

3.5.4 Sensitivity Analysis.

Our approach utilizes HSIs observed in a video. Longer videos typically contain more HSIs, providing additional constraints for our objective function. In Table 3.6, we analyze the impact of video length on scene reconstruction by reporting the 3D intersection-over-union (IoU) metric. We use 10 sequences from the PROX qualitative dataset, one sequence per scene, and randomly sample segments of 10s, 20s, and 30s length from each sequence. We find that longer sequences correlate with improved performance, as indicated by higher IoU values and lower standard deviation. The analysis suggests that performance depends on the number of HSIs rather than video length itself; a shorter video with numerous HSIs can yield better reconstruction than a longer video with fewer unique HSIs.

We also conducted a sensitivity study with respect to noise in the initialization. In Table 3.7, we add uniform noise on the initial scale, translation and orientation of objects predicted by Total3D Nie *et al.* (2020), and report the 3D IoU. MOVER is robust to noisy

	10s	20s	30s	entire videos (51s)
3D IoU mean ↑	0.389	0.395	0.407	0.424
3D IoU std. ↓	0.018	0.015	0.010	-

Table 3.6: Ablation study on different length of videos as input. The average length of entire videos is 51s.

Scale Noise	± 25%	± 15%	± 0.05%
3D IoU ↑	0.345	0.3805	0.4105
Translation	± 30 cm	± 20 cm	± 10 m
3D IoU ↑	0.4175	0.416	0.415
Orientation	± 45°	± 30°	±15°
3D IoU ↑	0.4205	0.418	0.4205

Table 3.7: Sensitivity analysis on scene reconstruction with uniform noise on input scale, translation, and orientation from Total3D Nie *et al.* (2020) (*Werkraum_03301_01* video). Scene without noise has a 3D IoU of 0.417.

orientation and translation estimates from Total3D, but sensitive to scale variations. This is because we currently regularize the optimization to the initial scale relatively strongly; meaning we cannot deviate much from a noisy estimate to "correct" it.

3.5.5 Failure Cases

In this section, we discuss and show the failure cases of our method. Besides optimizing the 3D scene layout, we do not modify the initial shape estimate of an object. As a result, an incorrectly estimated shape can still disrupt human interaction, as shown in (A) in Figure 3.9. A more flexible and adjustable geometry representation, e.g., an implicit representation, would be necessary. Human motion reconstruction struggles with severe occlusions in the input, leading to incorrect body poses and poor estimations of HSIs, which affect our 3D scene layout prediction, see (B) in Figure 3.9. While not the scope of our work, the robustness and accuracy of human motion estimation can be improved by incorporating human motion priors or learning-based probabilistic human pose and estimation network. Severe occlusion can also cause missing objects in the scene like the chair in Figure 3.9(C).

In our pipeline, we currently consider the contact between *detected* objects and bodies. As a potential future extension of our method, one can also leverage the information from a 2D learning-based human-object interaction (HOI) detection network Zou *et al.* (2021), by using contacted bodies to discover missing objects; or learn a model that jointly regresses human-object interaction and their shape.

3.6 Discussion

Based on single-view inputs, our proposed method optimizes the 3D alignment of objects in a *static scene*. However, humans also move objects, resulting in a dynamic scene layout. While our approach uses individual mesh models for each object, we assume a static scene. Nevertheless, we believe that our proposed constraints based on HSIs will be beneficial for future work on the reconstruction of dynamic scenes. Besides optimizing the 3D scene layout, we do not change the initial shape estimate of an object. A more flexible and adjustable geometry representation, e.g., an implicit representation, would be needed since the initial mesh could have a wrong topology.

Human motion reconstruction and 2D instance segmentation struggle with severe occlusions in the input, which leads to poor estimation of HSIs, influencing our 3D scene layout prediction. While not the scope of our work, the robustness and accuracy of human motion estimation can be improved by incorporating human motion priors Zhang *et al.* (2021b); Rempe *et al.* (2021). Also, jointly predicting human motion and the 3D scene with HSIs in a probabilistic framework can be another interesting direction for future work.

3.6.1 Discussion of Potential Misuse

Our approach is not intended for any surveillance application. Our goal is to understand how humans interact and move in scenes from videos (e.g., from TV sitcoms), to this end both the scene geometry and the human pose need to be reconstructed. Our method could be misused in potential surveillance applications that curtail human rights and civil liberties, but we restrict the usage of our method for surveillance.

3.6.2 Conclusion

We have introduced MOVER, which reconstructs a 3D scene by exploiting 3D humans interacting with it. We have demonstrated that accumulated HSIs, computed from a monocular video, can be leveraged to improve the 3D reconstruction of a scene. The reconstructed scene, in turn, can be used to improve 3D human pose estimation. In contrast to the state of the art, MOVER can reconstruct a consistent, physically plausible 3D scene layout.

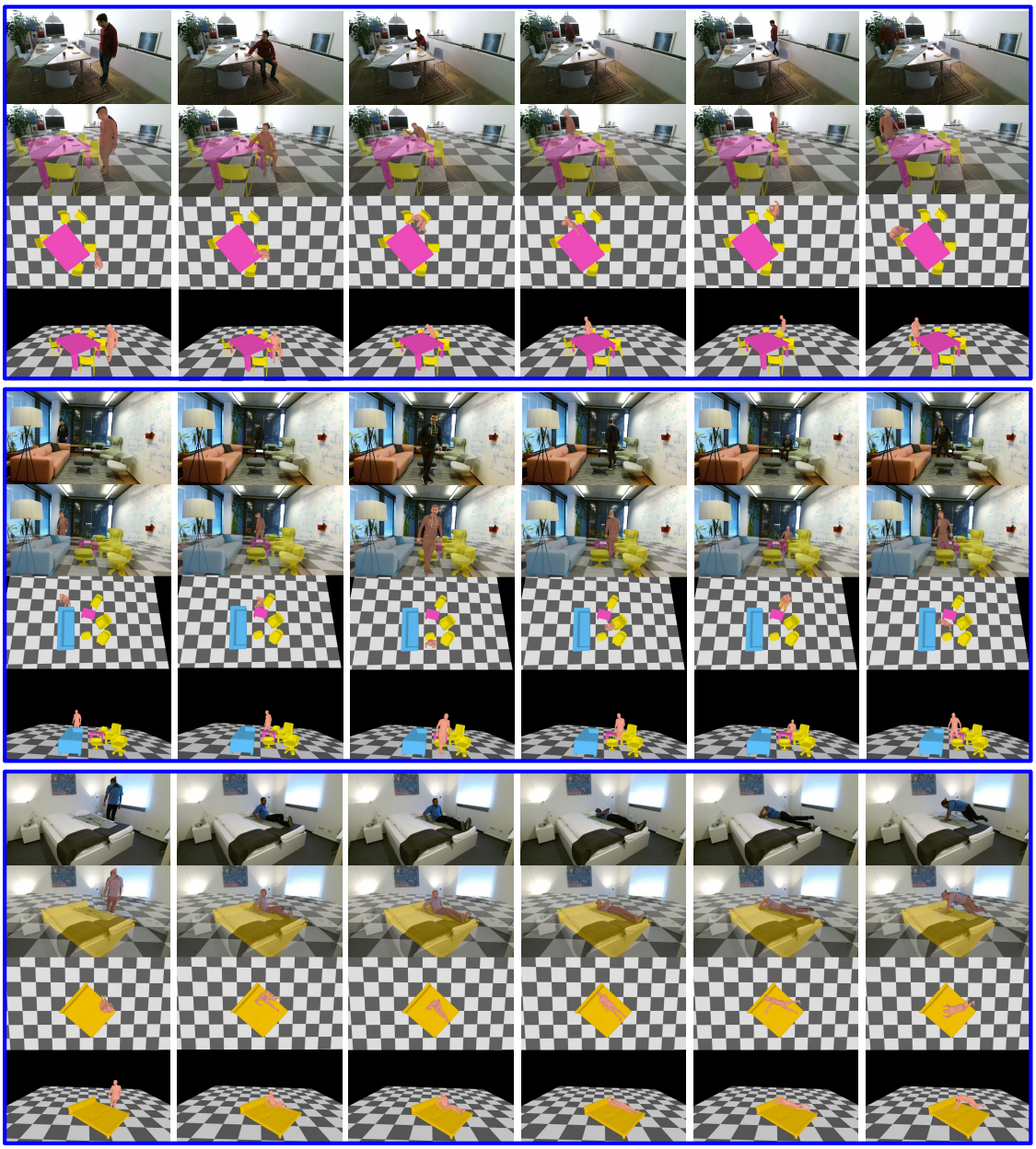


Figure 3.7: More qualitative results on the PROX qualitative dataset.

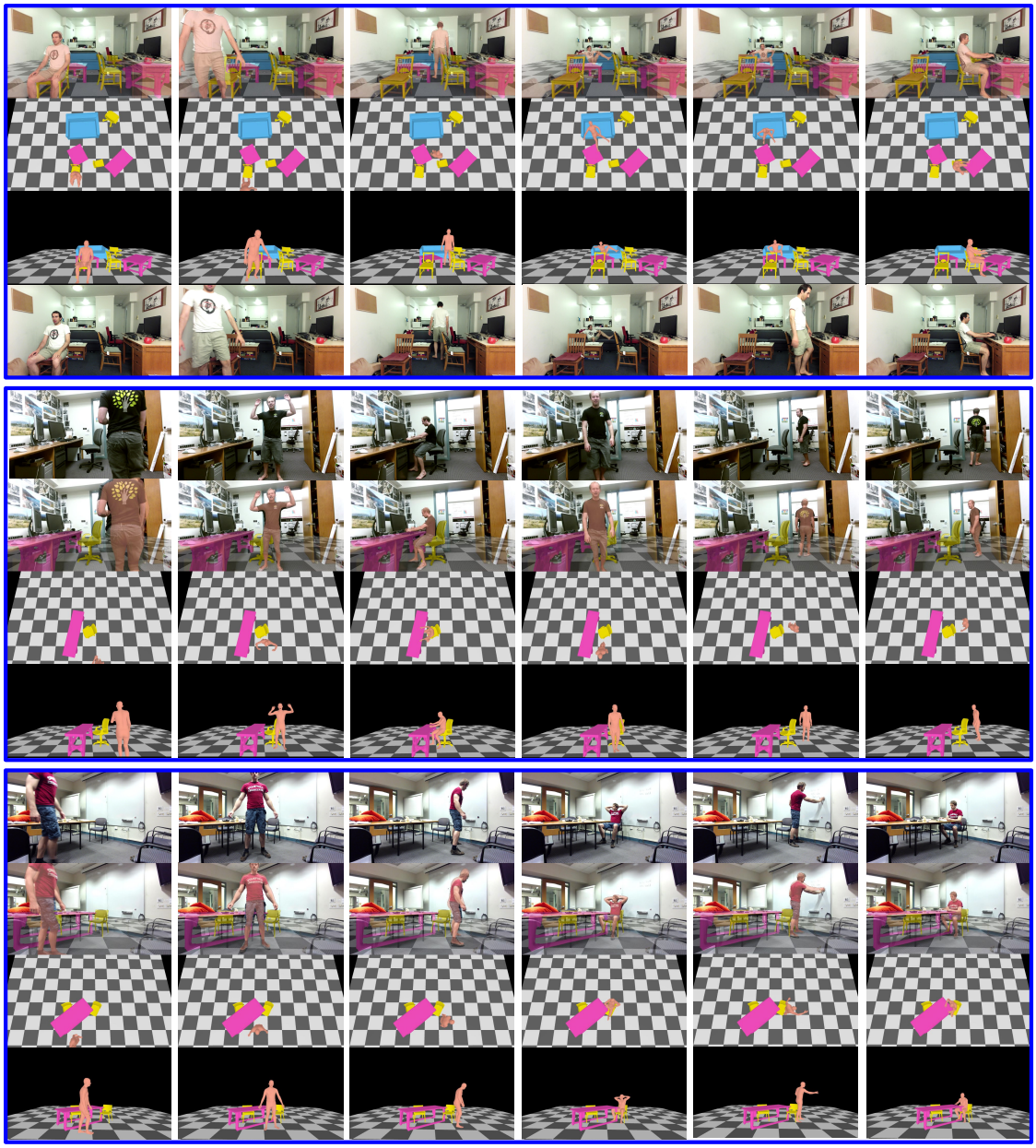
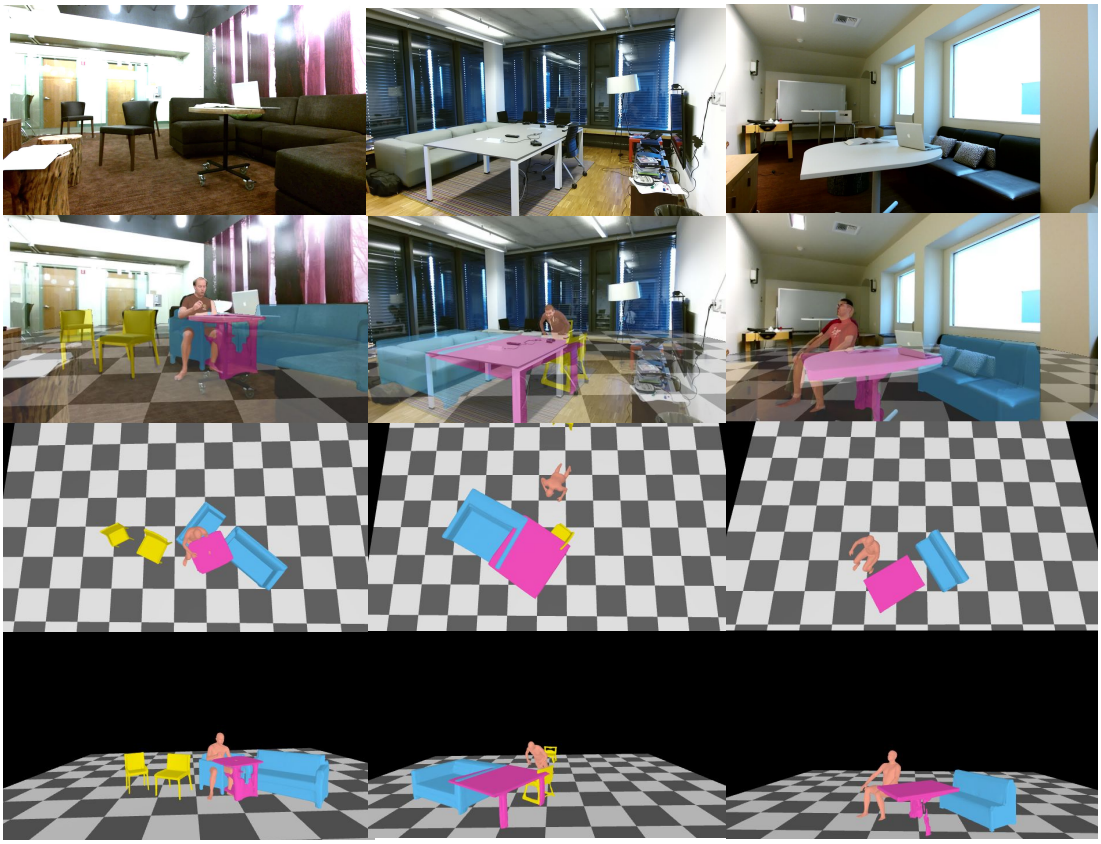


Figure 3.8: More qualitative results on the PiGraphs dataset.



(A) (B) (C)

Figure 3.9: Failure cases. (A) The estimated sofa has arms, which does not match the armless sofa in the input image. (B) The lower half of the body is occluded, leading to incorrect pose estimation and HSI observation. Additionally, the body appears to be 'sitting in the air' because the chair is missing.

Chapter 4

Human-Aware 3D Scene Generation

4.1 Introduction

Humans constantly interact with their environment. They walk through a room, touch objects, rest on a chair, or sleep in a bed. All these interactions contain information about the scene layout and object placement. In fact, a mime is a performer who uses our understanding of such interactions to convey a rich, imaginary, 3D world using only their body motion. Can we train a computer to take human motion and, similarly, conjure the 3D scene to which it belongs? Such a method would have many applications in synthetic data generation, architecture, games, and virtual reality. For example, there exist large datasets of 3D human motion like AMASS Mahmood *et al.* (2019) and such data rarely contains information about the 3D scene in which it was captured. Could we take AMASS and generate plausible 3D scenes for all the motions? If so, we could use AMASS to generate training data containing realistic human-scene interaction.

To answer such questions, we train a new method called MIME (Mining Interaction and Movement to infer 3D Environments) that generates plausible indoor 3D scenes based on 3D human motion. Why is this possible? The key intuitions are that (1) a human's motion through free space indicates the lack of objects, effectively *carving out* regions of the scene that are free of furniture. And (2) when they are in contact with the scene, this constrains both the type and placement of 3D objects; e.g., a sitting human must be sitting on something, such as a chair, a sofa, a bed, etc.

To make these intuitions concrete, we develop MIME, which is a transformer-based auto-regressive 3D scene generation method that, given an empty floor plan and a human motion sequence, predicts the furniture that is in contact with the human, as shown in Figure 4.1. It also predicts plausible objects that have no contact with the human but fit with the other objects and respect the free-space constraints induced by the human motion. To condition the 3D scene generation with human motion, we estimate possible contact poses using POSA Hassan *et al.* (2021a) and divide the motion into contact and non-contact snippets. The non-contact poses define free space in the room, which we encode as 2D floor maps, by projecting the foot vertices onto the ground plane. The contact poses and corresponding 3D human body models are represented by 3D bounding boxes of the contact vertices predicted by POSA. We use this information as input

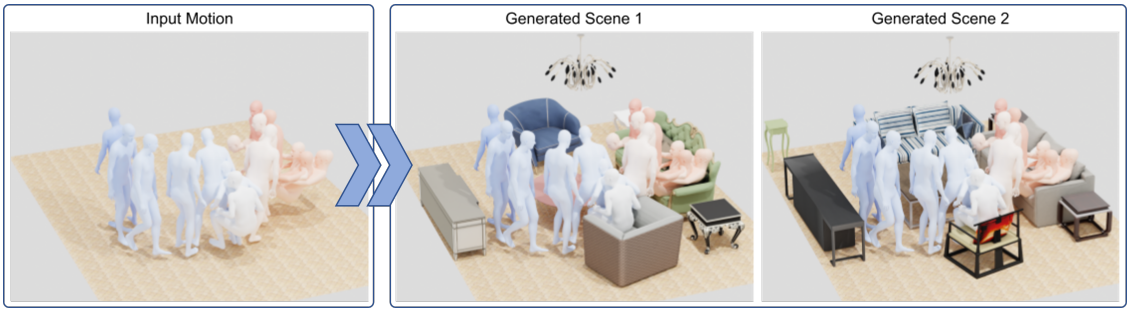


Figure 4.1: **Estimating 3D scenes from human movement.** Given 3D human motion, e.g., from motion capture or body-worn sensors, we reconstruct plausible 3D scenes in which the motion could have taken place. Our generative model is able to produce multiple realistic scenes that take into account the locations and poses of the person, with appropriate human-scene contact.

to the transformer and auto-regressively predict the objects that fulfill the contact and free-space constraints.

To train MIME, we built a new dataset called *3D-FRONT Human*, which extends the large-scale synthetic scene dataset 3D-FRONT Fu *et al.* (2021a). Specifically, we automatically populate the 3D scenes with humans, i.e., non-contact humans (a sequence of walking motions and standing humans) as well as contact humans (sitting, touching, and lying humans). To this end, we leverage motion sequences from AMASS Mahmood *et al.* (2019), as well as static contact poses from AGORA Patel *et al.* (2021) scans.

At inference time, MIME generates a plausible 3D scene layout for the input motion, represented as 3D bounding boxes. Based on this layout, we select 3D models from the 3D-FUTURE dataset Fu *et al.* (2021c) and refine their 3D placement based on geometric constraints between the human poses and the scene.

In comparison to pure 3D scene generation baselines like ATISS Paschalidou *et al.* (2021), our method generates a 3D scene that supports human contact and motion while placing plausible objects in free space. In contrast to Pose2Room Nie *et al.* (2022), which is a recent pose-conditioned generative model, our method enables the generation of objects that are not in contact with the human, thus predicting the entire scene instead of isolated objects. We demonstrate that our method can be directly applied to real captured motion sequences such as PROX-D Hassan *et al.* (2019) *without fine-tuning*.

In summary, we make the following contributions:

- a novel motion-conditioned generative model for 3D room scenes that auto-regressively generates objects in contact with the human or avoids free-space defined by the motion.
- a new 3D scene dataset with interacting humans and free space humans which is constructed by populating 3D FRONT with static contact/standing poses from

AGORA and motion data from AMASS.

4.2 Method

Given input motion of a human and an empty or partially occupied room of a specific kind (e.g., bedroom, living room, etc.) with its floor plan, we learn a generative model that can populate the room with objects that do not collide with the input humans and also support them. To this end, we propose a human-aware autoregressive model that represents scenes as *one* unordered set of objects. We divide the objects into two kinds, i.e., contact objects and non-contact objects, based on the human-object interaction. Contact objects are ones that humans interact with. Non-contact objects can be placed anywhere in the free space of a room that makes semantic sense. These objects enrich the content and potential functionality of a room.

In the following, we describe our human-aware scene synthesis model, MIME, which consists of two components: (1) a generative scene synthesis method based on 3D bounding boxes with object labels, and (2) a 3D refinement method that takes 3D human-scene interactions into account to optimize the rotation and placement of the generated objects. In Section 4.3, we detail the dataset generation process used to train our model.

4.2.1 Generative Human-aware Scene Synthesis

Given humans $\mathcal{H}$ and a floor plan $\mathcal{F}$, our goal is to generate a "habitat" $\mathcal{X} = \{\mathcal{H}, \mathcal{F}, \mathcal{S}\}$ where the 3D scene $\mathcal{S}$ can support all human interactions and motions. In contrast to the pure 3D scene generation methods Paschalidou *et al.* (2021); Para *et al.* (2023), we focus on leveraging information from human motion to guide the 3D scene generation. To this end, we extract two types of information from the input motion and the corresponding human bodies: (i) contact humans $\mathcal{C}$ and (ii) free-space humans. We use POSA Hassan *et al.* (2021a), to take posed human meshes and automatically label which of their vertices are potentially in contact with an object. Free-space humans are those that are only in contact with the floor plane $\mathcal{F}$. These define a binary mask that we call free-space mask $\mathcal{FS}$, which is constructed by the union of all projected foot contact points on $\mathcal{F}$. This free-space mask $\mathcal{FS}$ defines the region of a room that is free from objects, as a human can stand and walk there. Given all contact humans, we compute the bounding boxes of their contact vertices and keep only the non-overlapping boxes using non-maximum suppression; we denote these as c_i. The collection of contact boxes is referred to as $\mathcal{C} = \{c_i\}_{i=1}^{N}$. Instead of storing all contact vertices of all bodies, our features are compact and encode complementary information. The contact humans, represented by $\mathcal{C}$, indicate where to locate an object. See Figure 4.3 top and middle rows for an illustration.

We represent a 3D scene $\mathcal{S}$ as an unordered set of objects, consisting of two kinds of objects based on human-object interaction. Objects in contact with the input human are referred to as contact objects $\mathcal{O} = \{o\}_{i=1}^{N}$, while non-contact objects $\mathcal{Q} = \{q\}_{i=1}^{M}$

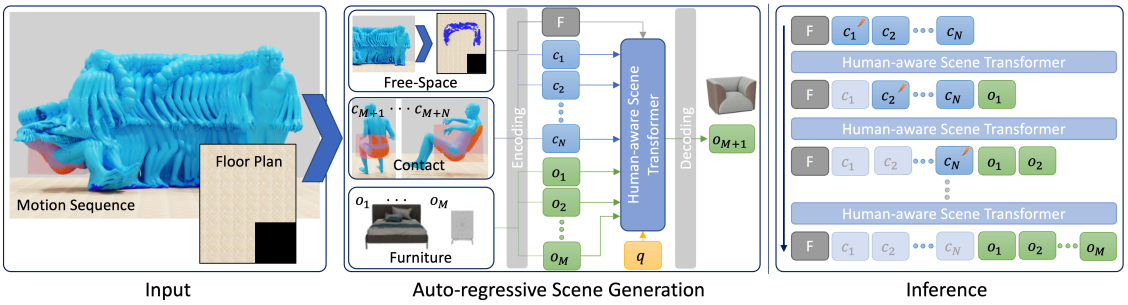


Figure 4.2: Method overview. In training, our method generates object $M+1$ through a transformer encoder and a decoding module, conditioned on the free space concatenated with the floor plan, contact humans $c_{j=1}^{N}$, other existing objects $o_{j=1}^{M}$ and a learnable query q. We minimize the negative log-likelihood between the distribution of the generated object $M+1$ and the ground truth. In inference, we start from the floor plane, the free space, and input contact humans $c_{i=1}^{N}$ and assign the contact label of the first human as 1 by default, to autoregressively generate objects. At each step, we remove the contact humans that overlap the previously generated object and generate the next objects until the *end symbol* is generated.

are without any human interaction. Formally, a 3D scene is the union of contact and non-contact objects: $\mathcal{S} = \mathcal{O} \cup \mathcal{Q}$.

The free-space mask $\mathcal{FS}$, the floor plan $\mathcal{F}$, the contact humans $\mathcal{C}$ as well as the already existing objects $\mathcal{S}$ are input to an auto-regressive transformer model. Each input is encoded with a respective encoder, detailed below.

The log-likelihood of the generation of scene $\mathcal{S}$ including contact objects and non-contact objects is:

$$\log p(\mathcal{S}) = \log p(\mathcal{O}|\mathcal{F},\mathcal{FS},\mathcal{C}) + \log p(\mathcal{Q}|\mathcal{F},\mathcal{FS},\mathcal{C}). \tag{4.1}$$

To calculate the likelihood of all generated contact objects $\mathcal{Q}$, we accumulate the likelihood of every contact object:

$$p(\mathcal{O}|\mathcal{F},\mathcal{FS},\mathcal{C}) = \sum_{\hat{\mathcal{O}} \in \pi(\mathcal{O})} \prod_{j \in \hat{\mathcal{O}}} p\left(o_j \mid o_{<j}, \mathcal{F}, \mathcal{FS}, c_{\geq j}\right), \tag{4.2}$$

where $p\left(o_j \mid o_{<j}, \mathcal{F}, \mathcal{FS}, c_{\geq j}\right)$ is the probability of generating the j_{th} object conditioned on the input floor plan, free-space humans, the rest of contact humans and the previously generated objects, and π is the random permutation function for those generated contact objects in the scene. The likelihood of all non-contact objects $\mathcal{Q}$ is computed by replacing the input contact humans with the corresponding generated contact objects. During the training, we remove all contact humans inside the room; thus, all contact objects $\mathcal{O}$ can

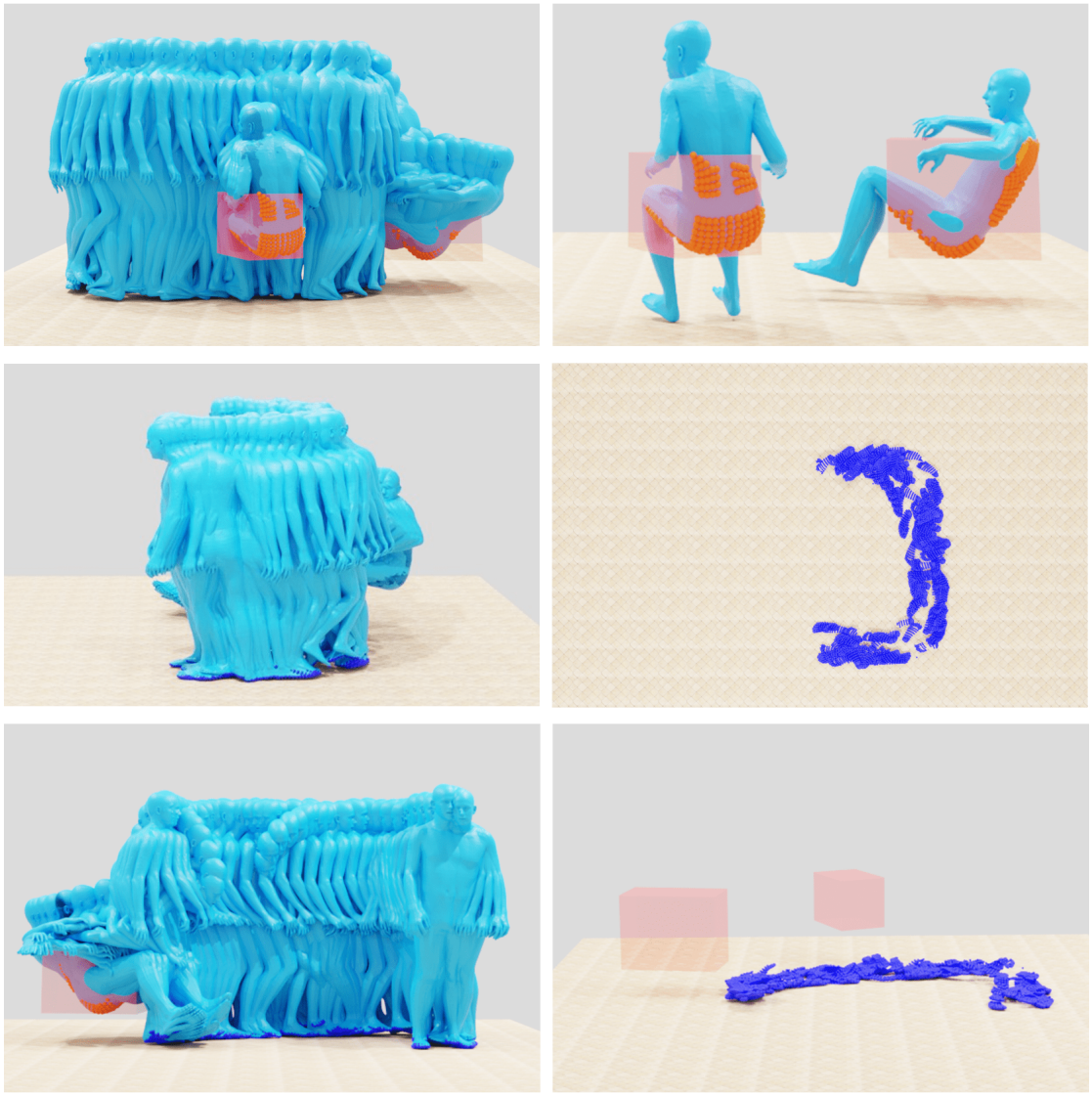


Figure 4.3: We divide input humans into two parts: contact humans and free-space humans. We extract the 3D bounding boxes for each contact human and use non-maximum suppression on the 3D joint union to aggregate multiple humans in the same 3D space into a single contact 3D bounding box (orange boxes). We project the foot vertices of free-space humans on the floor plane, to obtain the 2D free-space mask (dark blue).

be treated as non-contact objects $\mathcal{Q}'$:

$$\begin{aligned} p(\mathcal{Q}|\mathcal{F},\mathcal{FS},\mathcal{C}) &= p(\mathcal{Q}|\mathcal{F},\mathcal{FS},\mathcal{O}) \\ &= p(\mathcal{Q}|\mathcal{F},\mathcal{FS},\mathcal{Q}') \\ &= \sum_{\hat{\mathcal{Q}}\in\pi(\mathcal{Q}+\mathcal{Q}')} \prod_{j\in\hat{\mathcal{Q}}} p\left(q_j \mid q_{<j},\mathcal{F},\mathcal{FS}\right). \end{aligned} \tag{4.3}$$

We follow Paschalidou *et al.* (2021) to use Monte Carlo sampling to approximate all different object permutations during training, to make our model invariant to the order of generated objects.

Free-Space Encoder. The 2D free-space mask $\mathcal{FS}$ is encoded together with the 2D floor plan $\mathcal{F}$ using a ResNet-18 He *et al.* (2016). The encoded feature provides the information to the transformer encoder about where an object can be placed.

Contact Encoder. We represent the contact humans as 3D bounding boxes, which consist of the contact label I, the contact class category k (sitting, touching, lying), the translation t, the rotation r, and the size s. During generation of a scene, we set the contact label I of one contact human to 1 while the others are labeled 0. This label highlights the contribution of the specific contact human to the next generated contacted object. Note that we remove contact humans from the input set if they are already in contact with an existing object in the scene. Otherwise, we encode the j_{th} input contact human by applying:

$$E_\theta : (I_j,k_j,t_j,r_j,s_j) \rightarrow (I_j,\lambda(k_j),p(t_j),p(r_j),p(s_j)), \tag{4.4}$$

where $\lambda(\cdot)$ is a learnable embedding for the contact class category k, and $p(\cdot)$ Vaswani *et al.* (2017) is the positional encoding for the translation t, rotation r and size s.

Furniture Encoder. The furniture encoder computes the embedding of existing objects in the room:

$$E_\theta : (I_j=0,k_j,t_j,r_j,s_j) \rightarrow (0,\lambda(k_j),p(t_j),p(r_j),p(s_j)). \tag{4.5}$$

Note that the furniture encoder shares the same weight as the contact encoder. The contact labels of the objects are all zero, where $j \in [1,M]$.

Scene Synthesis Transformer. We pass the free-space feature F, context embedding $T_{i=1}^{M+N}$, and a learnable query vector $q \in \mathbb{R}^{64}$ into a transformer encoder τ_θ Vaswani *et al.* (2017); Devlin *et al.* (2018) without any positional encoding Vaswani *et al.* (2017), to

predict the feature $\hat{q}$ that is used to generate the next object:

$$\tau_\theta(F, T_{i=1}^{M+N}, q) \rightarrow \hat{q}. \tag{4.6}$$

To decode the attribute distribution $(\hat{k}, \hat{t}, \hat{r}, \hat{s})$ of the generated object o_{M+1} from $\hat{q}$, we follow the same design from ATISS Paschalidou *et al.* (2021). Specifically, we employ an MLP for each attribute in a consecutive fashion. Given $\hat{q}$, we first predict the class category label $\hat{k}$, then we predict the $\hat{t}$, $\hat{r}$ and $\hat{s}$ in this specific order, where the previous attribute will be concatenated with the input $\hat{q}$ for the next attribution prediction.

4.2.2 Training and Inference.

We train our model on the training set of *3D FRONT HUMAN*, by maximizing the log-likelihood of each generated scene $\mathcal{S}$ in Equation (4.1). During training, we select a human-populated scene in *3D FRONT HUMAN* and add a random permutation $\pi(\cdot)$ on all N contact and M non-contact objects. We randomly select the $m_{th}+1$ as the generated object, where $m \in [0, N+M]$. Note that, $m=0$ represents an empty scene, while $m=N+M$ indicates the generated scene is already full, and the class label of the predicted object is an extra *end symbol*. Our model predicts the attribute distribution of the generated object, conditioned on the floor plane $\mathcal{F}$, free space $\mathcal{FS}$, previous m objects and contact humans $\mathcal{C}$; see Figure 4.2. To enable our model to generate both contact objects and non-contact objects, we augment the data by adding input contact humans or dropping them out with equal frequency.

During inference, we start with an empty floor plane F and input humans, including free-space humans $\mathcal{FS}$ and contact humans $\mathcal{C}$. We autoregressively sample the attributes of the next generated object to place one object into a scene. By default, we set the contact label of the first contact human to 1 and the rest are 0. After each generation step, we remove contact humans that are already in contact, by computing the 2D IoU of the human bounding box and the generated object by projecting them onto the ground plane. Specifically, if the IoU is larger than 0.5, we remove the contact human from the input. Once the *end symbol* is generated, the scene generation is complete.

4.2.3 3D Scene Refinement

The generated scene from our model is represented with 3D bounding boxes. Based on the bounding box size and class category label, we retrieve the closest mesh model from 3D FUTURE Fu *et al.* (2021c). To improve the human-scene interaction between the generated scenes and input humans, we optimize the collision loss and the contact loss from MOVER Yi *et al.* (2022) to refine the object position, as can be seen in Figure 4.4. We calculate a unified SDF volume and accumulate all contact vertices for all humans in the 3D space, then jointly optimize the object alignment to improve human-object

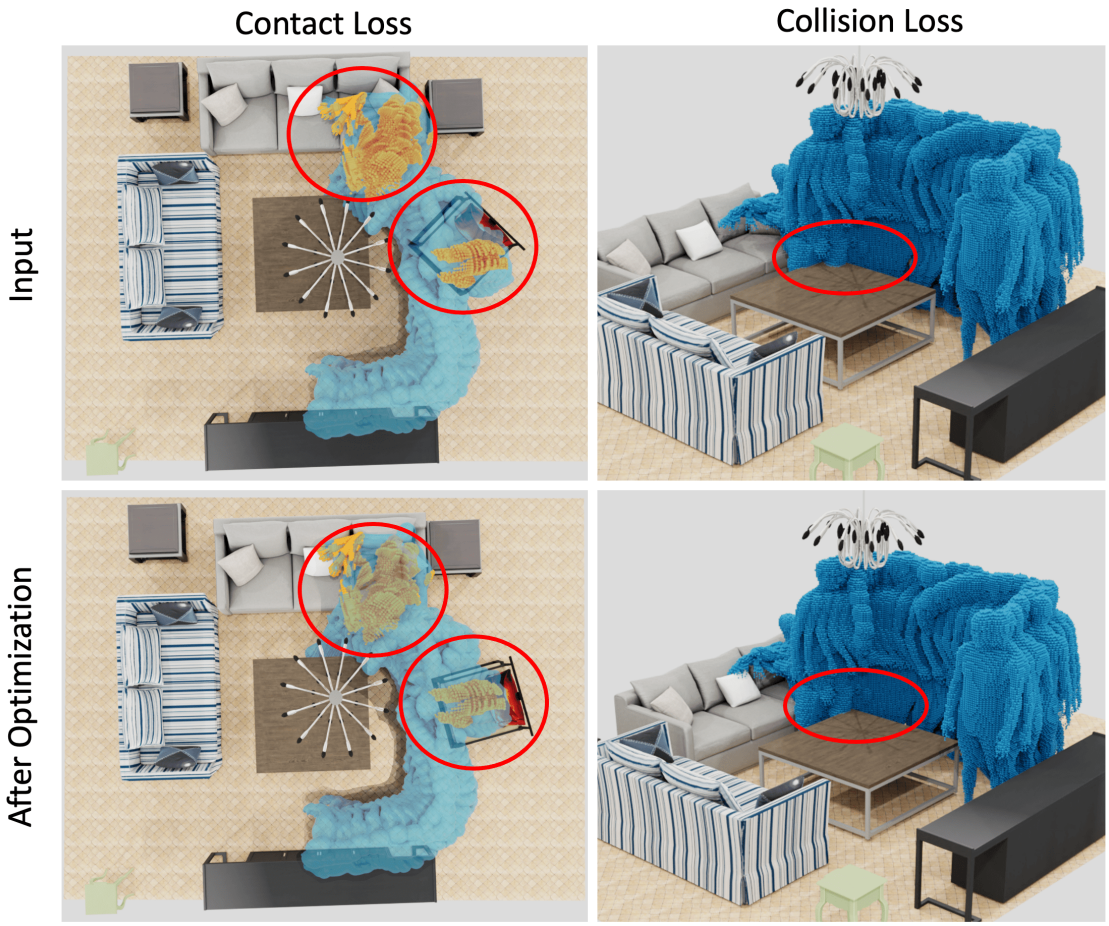


Figure 4.4: Scene refinement with the collision and contact loss from MOVER Yi *et al.* (2022). In the contact loss, all contact vertices (orange color) are accumulated from all bodies into 3D space and the sofa and chair are refined by minimizing the one-directional Chamfer Distance with the contact vertices. In the collision loss, we compute one uniform SDF volume for all bodies, where the inside of bodies is denoted as blue voxels. The table gets refined with the collision loss.

contact and resolve 3D interpenetrations between humans and the scene. The MOVER contact loss weight and the collision loss weight are $1e5$ and $1e3$ respectively.

4.2.4 Training Details

During training, we apply the Adam optimizer Kingma and Ba (2014) with a learning rate $1e^{-4}$ and no weight decay. In the Adam optimizer, we use the default PyTorch implemented parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 1e-8$. We train MIME with a batch size of 128 for $100k$ iterations. We perform random global rotation augmentation

	Interpenetration(↓)		2D IoU(↑)		3D IoU(↑)	
	ATISS Paschalidou *et al.* (2021)	Ours	ATISS Paschalidou *et al.* (2021)	Ours	ATISS Paschalidou *et al.* (2021)	Ours
Bedroom	0.348	**0.129**	0.472	**0.939**	0.376	**0.756**
Living	0.129	**0.050**	0.480	**0.971**	0.360	**0.920**
Dining	0.121	**0.047**	0.163	**0.959**	0.122	**0.769**
Library	0.139	**0.106**	0.351	**0.725**	0.390	**0.570**

Table 4.1: Quantitative comparison on the *test* split of the *3D FRONT HUMAN* dataset for human-scene interaction. Interpenetration loss, 2D IoU, and 3D IoU are used to evaluate the interaction quality in generated scenes.

	FID Score (↓)		Category KL Div. (↓)	
	ATISS Paschalidou *et al.* (2021)	Ours	ATISS Paschalidou *et al.* (2021)	Ours
Bedroom	**70.21**±1.80	74.18±2.19	**0.028**	0.044
Living	**130.61**±1.27	150.03± 1.00	**0.004**	0.053
Dining	**45.99** ± 0.90	76.75 ± 1.45	**0.004**	0.037
Library	**93.16** ± 2.59	118.34±2.94	**0.066**	0.093

Table 4.2: Evaluation of generative model performance on the *test* split of the *3D FRONT HUMAN* dataset. FID score and category KL divergence are used to assess the realism and diversity of generated scenes compared with ATISS.

between $[0, 360]$ degrees on the holistic populated scene, including the floor plane, all objects, the free space, and all contact humans.

4.3 Dataset Generation of 3D FRONT HUMAN

To enable 3D scene generation from humans, we need a dataset that consists of a large number of rooms with a wide variety of human interactions. Since no such dataset exists, we generate a new synthetic dataset by populating the 3D rooms in the 3D FRONT Fu *et al.* (2021a) with interactive humans. We name the resulting dataset *3D FRONT HUMAN*. To populate the rooms of 3D FRONT with people, we insert humans with contact and humans that stand or walk in free space, as shown in Figure 4.5. We represent people using the SMPL-X model Pavlakos *et al.* (2019a) and add contact humans from AGORA Patel *et al.* (2021) by randomly assigning plausible interactions to different contactable objects in the room. Specifically, we allow three types of contact interactions: touching, sitting, and lying. In Figure 4.5 (bottom), we place a lying down person on a bed, and multiple humans interact with a nightstand or wardrobe. In the free space, we put a random number of static standing people and add multiple walking motion clips from AMASS Mahmood *et al.* (2019) with random start positions and directions to the scene, and remove humans that intersect with objects in the scene.

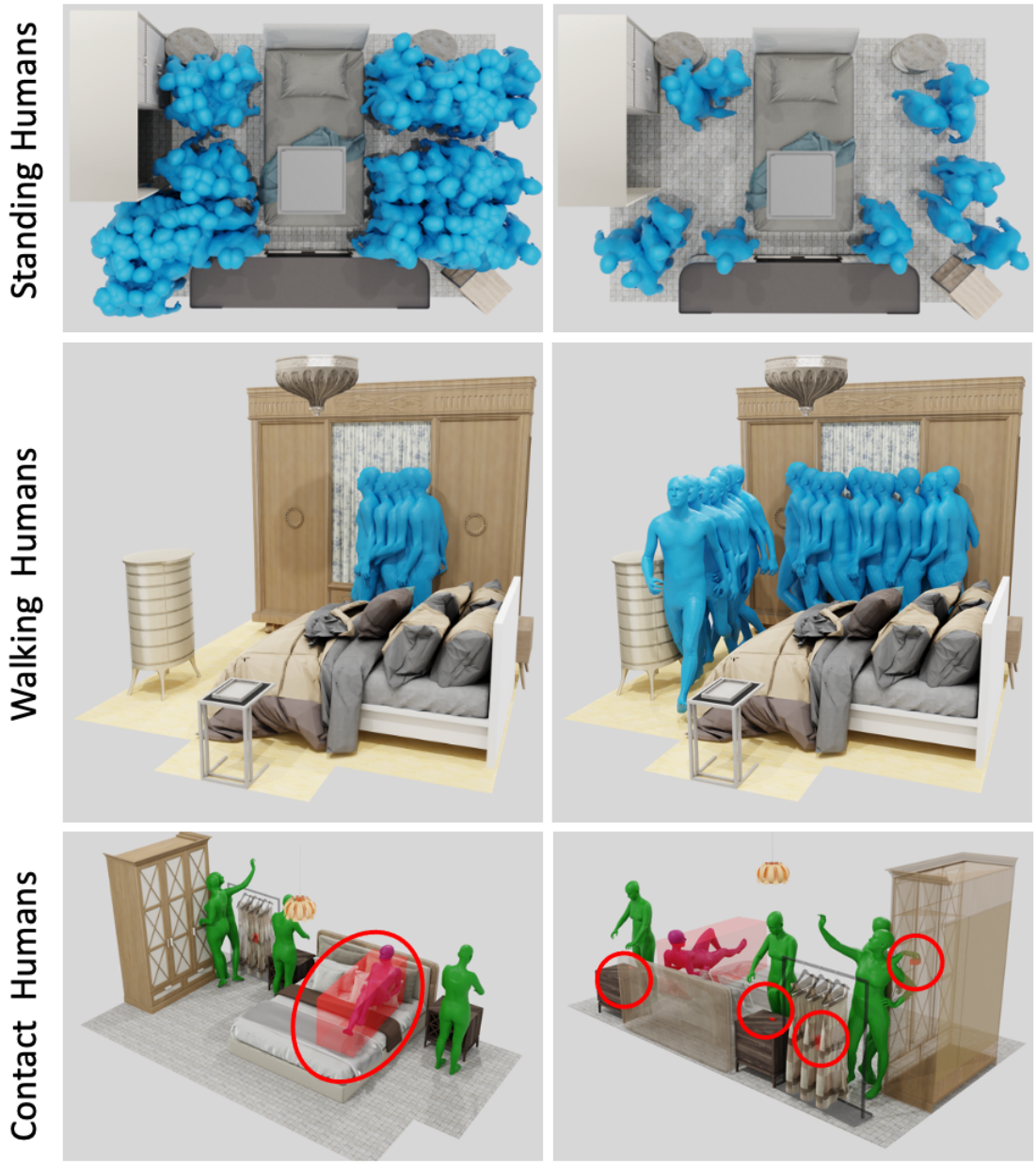


Figure 4.5: The illustration of populated 3D scenes in *3D FRONT HUMAN*. Given a room, we place random numbers of static "standing" people and add multiple "walking" motion sequences with varying start positions and directions in the free space. We also place various "contact humans" into the scene so that their interaction with the objects makes sense, e.g., "touching" and "lying". The red boxes represent the bounding boxes of the contact vertices of each interactive body.

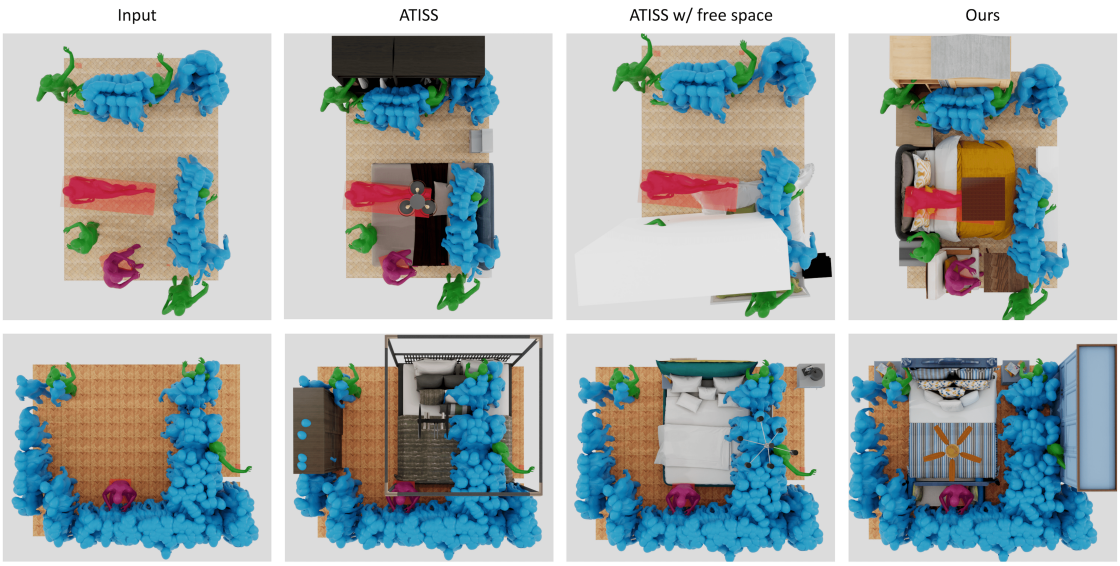


Figure 4.6: Qualitative comparison on the test split in *3D FRONT HUMAN*. Given free space and contact humans as input, MOVER generates more plausible scenes in which the contact humans interact with the contact objects and the free space humans have fewer collisions with all the generated objects. We also show the original ATISS with or without the free space mask as input. All results are without refinement. Top and bottom rows represent two different example inputs.

4.4 Experiments

We qualitatively and quantitatively evaluate our method and compare with two baselines. Specifically, we compare to the 3D scene generation method ATISS Paschalidou *et al.* (2021) and the human-aware scene reconstruction method Pose2Room Nie *et al.* (2022).

Evaluation Datasets. Our human-populated dataset *3D FRONT HUMAN* contains four room types: 1) 5689 bedrooms, 2) 2987 living rooms, 3) 2549 dining rooms and 4) 679 libraries. We use 21 object categories for the bedrooms, 24 for the living and dining rooms, and 25 for the libraries. We independently train our model four times for the four kinds of rooms. Following our baseline, ATISS Paschalidou *et al.* (2021), for each kind of room, we split the data 80%, 10%, 10% into training, validation and test sets respectively. We train and validate MIME on the training and validation sets respectively, and evaluate it on the test set. Since ATISS Paschalidou *et al.* (2021) does not provide a pre-trained model, we retrain it with the official code[1], following the same training strategy on the original 3D FRONT dataset as one of our baseline.

To evaluate the effectiveness and generalization of our method, we test MIME on

[1]https://github.com/nv-tlabs/ATISS/commit/6b46c11.

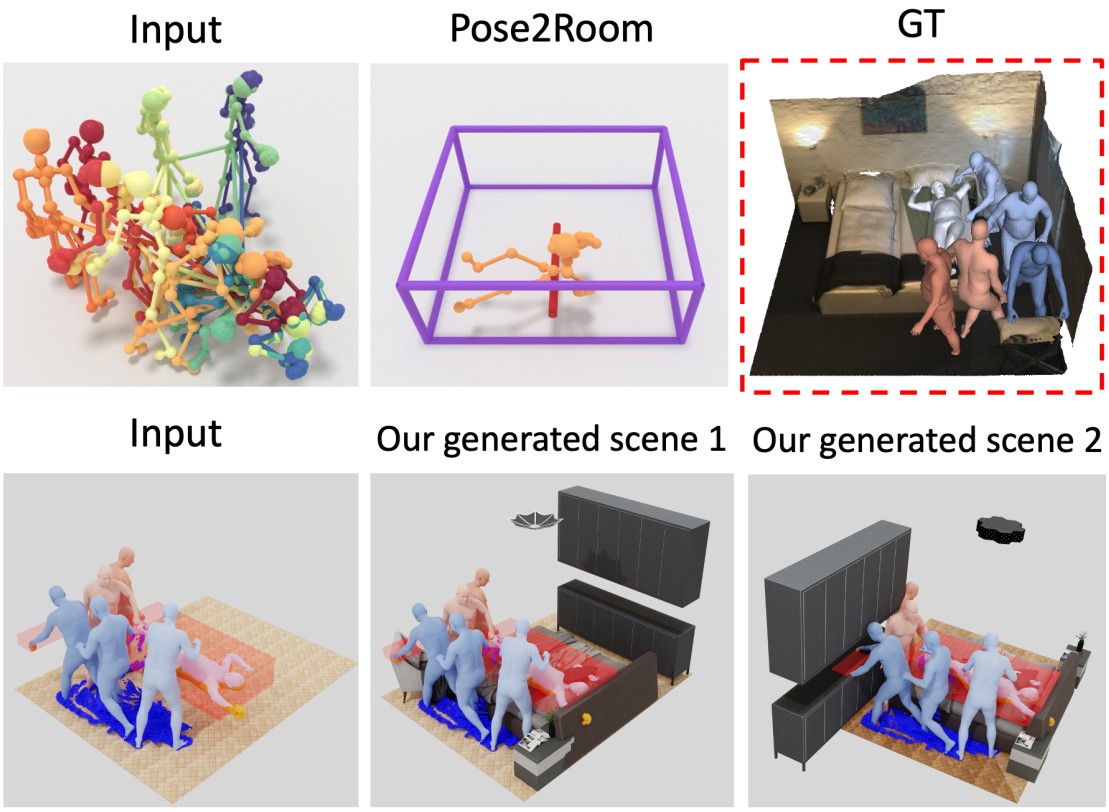


Figure 4.7: Evaluation on PROX Hassan *et al.* (2019); Yi *et al.* (2022). Compared with Pose2Room Nie *et al.* (2022), which uses the 3D skeletons of the same input motion as MOVER, MOVER (w/o finetuning and w/o refinement) can not only generate more accurate contact objects, but it also generates objects appropriately in free space. GT = ground truth.

a RGB-D based motion capture 3D motion dataset PROX-D Hassan *et al.* (2019) and compare it with Pose2Room Nie *et al.* (2022). Pose2Room requires a sequence of human motions that are in contact with objects. Our *3D FRONT HUMAN* does not provide these interactive human-object motions, so we cannot finetune and evaluate Pose2Room on our *3D FRONT HUMAN*.

Evaluation Metrics. We compare MIME with the baselines in two different ways: (i) the plausibility between human-scene interaction and (ii) the realism of the generated scenes only. We propose an *interpenetration loss* (↓) to evaluate the collisions between the generated objects and the free space, by computing the ratio of the violated free space

versus the 2D projection of the generated objects:

$$L_{\text{inter}} = \left(\sum_{j=1}^{M} \sum_{p \in O_j} \mathcal{FS}(p) \right) / \sum_{p \in \mathcal{FS}} \mathcal{FS}(p), \tag{4.7}$$

where p denotes each pixel on the floor plane image. We calculate the *2D IoU* and *3D IoU* between generated objects and input contact bounding boxes to measure human-object interaction. To evaluate the realism and diversity of generated scenes, we follow common practice Paschalidou *et al.* (2021); Zhang *et al.* (2020c) and calculate the FID Heusel *et al.* (2017) (at 256^2 resolution) score between birds-eye view orthographic projections of generated scenes and real scenes from the *test* set, as well as the category KL divergence. We compute the FID score 10 times and report its mean and variance. All these evaluation experiments are conducted on the *test* split of the *3D FRONT HUMAN* dataset.

4.4.1 Human-aware Scene Synthesis.

In Figure 4.6, we visualize the ability of our method to generate plausible 3D scenes from input motion and floor plans for different kinds of rooms; we also show our baseline methods for comparison.

We present more qualitative examples for different kinds of rooms, in Figure 4.8, Figure 4.9, and Figure 4.10. Compared with our baseline methods Paschalidou *et al.* (2021), our method can generate more plausible 3D scenes that input motions can interact with.

Note that the original ATISS Paschalidou *et al.* (2021) model generates a 3D scene only based on the floor plan, without taking the humans into account. Thus, scenes generated by ATISS violate the free space constraints and are not consistent with human contact. For a more fair comparison, we extend ATISS to take information about the human motion as input. Specifically, we adapt the 2D input floor plan to also contain the free space information of the walking and standing humans. However, ATISS with input free space still generates objects in free space, while also generating implausible object configurations such as the white closet inside the bed (Figure 4.6, top). In contrast, MIME generates plausible 3D scenes that have less interpenetration with the free space and support interacting humans; e.g. a bed beneath a lying person and a chair under a sitting person.

The observations in the qualitative comparison are also confirmed by a quantitative evaluation in Table 4.1 and Table 4.2. MIME achieves significant improvements on human-scene interaction evaluation metrics compared with ATISS. Note, since our scene generation is constrained by the input motion, the diversity scores (FID, KL divergence) are lower than of ATISS, which is not human-aware. This is not a failure/limitation of MIME.

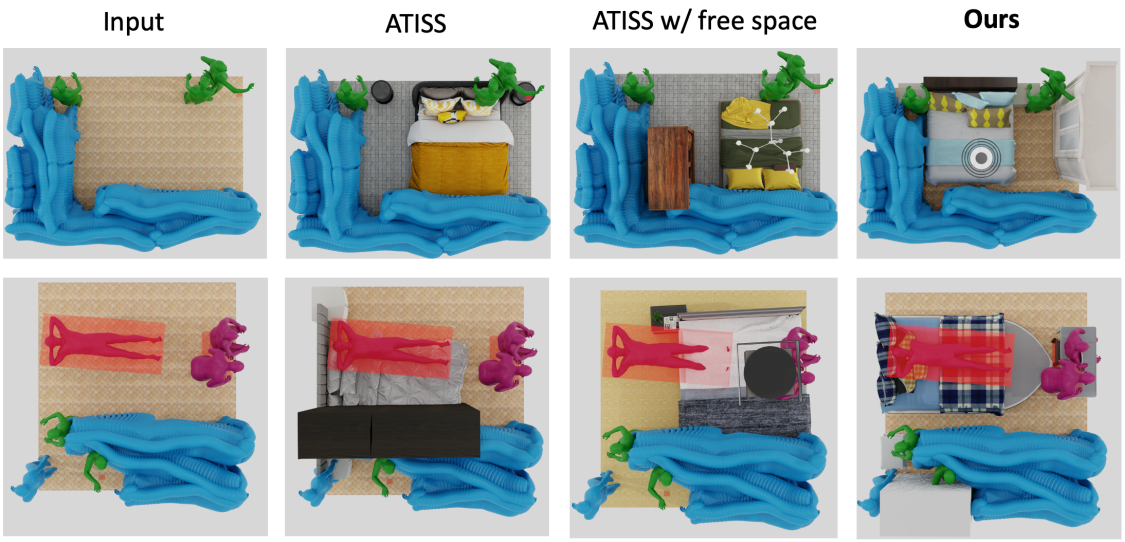


Figure 4.8: Qualitative comparison on bedrooms in the test split of *3D FRONT HUMAN*. Given free space and contact humans as input, MIME generates more plausible scenes in which the contact humans interact with the contact objects and the free space humans have fewer collisions with all the generated objects. We also show the original ATISS w/ or w/o the free space mask as input. All results are w/o refinement. Each row represents an example input.

To evaluate the generalization of our method, we test it on a motion capture (mocap) dataset of human motion. We consider the PROX-D Hassan *et al.* (2019) dataset and the 3D bounding box annotation from Yi *et al.* (2022). We use it *without scene refinement*, and use the motions to generate scenes. We compare our method with Pose2Room Nie *et al.* (2022), which predicts 3D objects from a motion sequence of 3D skeletons. Note that Pose2Room can only predict contact objects, it does not predict an entire scene which is the goal of our method. Figure 4.7 presents a qualitative comparison of the methods and we report the quantitative metrics in Table 4.3.

Specifically, we compute the mean average precision with 3D IoU 0.5 (mAP@0.5) to evaluate the 3D object detection accuracy for those contact objects only. Both methods are probabilistic generative models that predict the distribution of object attributes.

Method	3D IoU
P2R-Net Nie *et al.* (2022) w/o pretrain	5.36
Ours (MIME) w/o pretrain	**8.47**

Table 4.3: Comparisons on 3D object detection accuracy (mAP@0.5) using the PROX-D *qualitative* dataset Hassan *et al.* (2019).

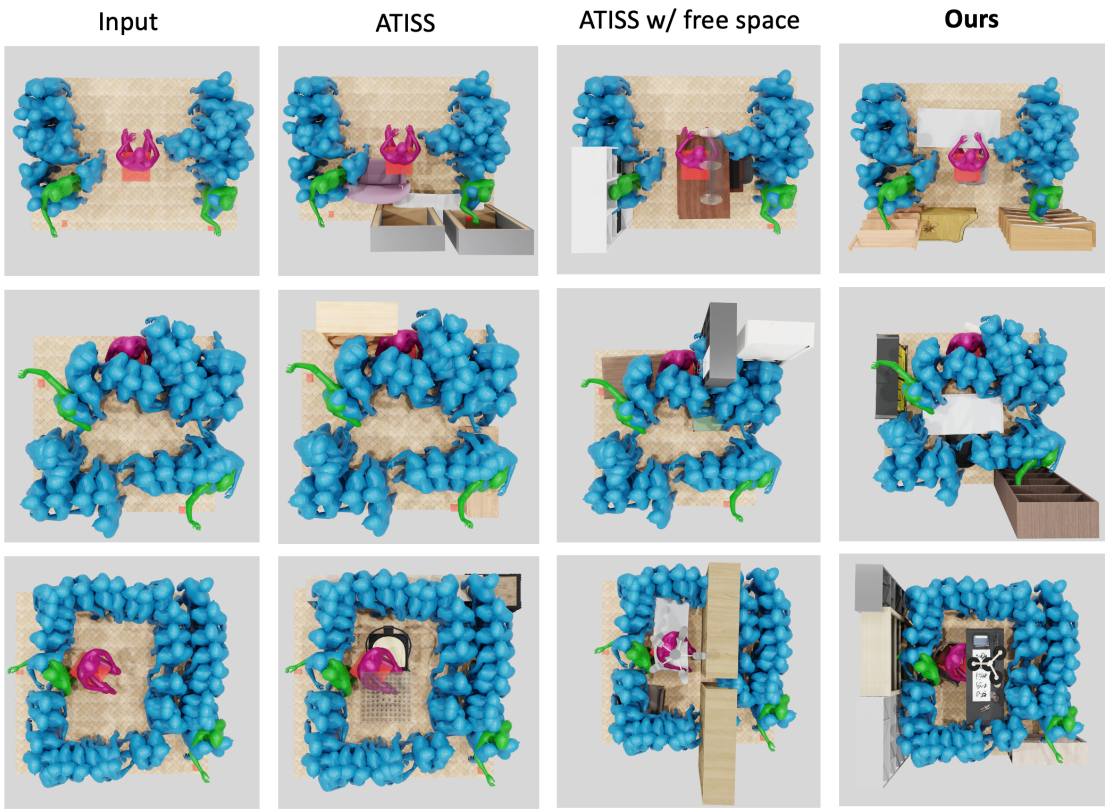


Figure 4.9: Qualitative comparison on the class "library" in the test split of *3D FRONT HUMAN*. Given free space and contact humans as input, MIME generates more plausible scenes in which the contact humans interact with the contact objects, and free space humans experience fewer collisions with all generated objects. We also show the original ATISS with or without the free space mask as input. All results are without refinement. Each row represents an example input case.

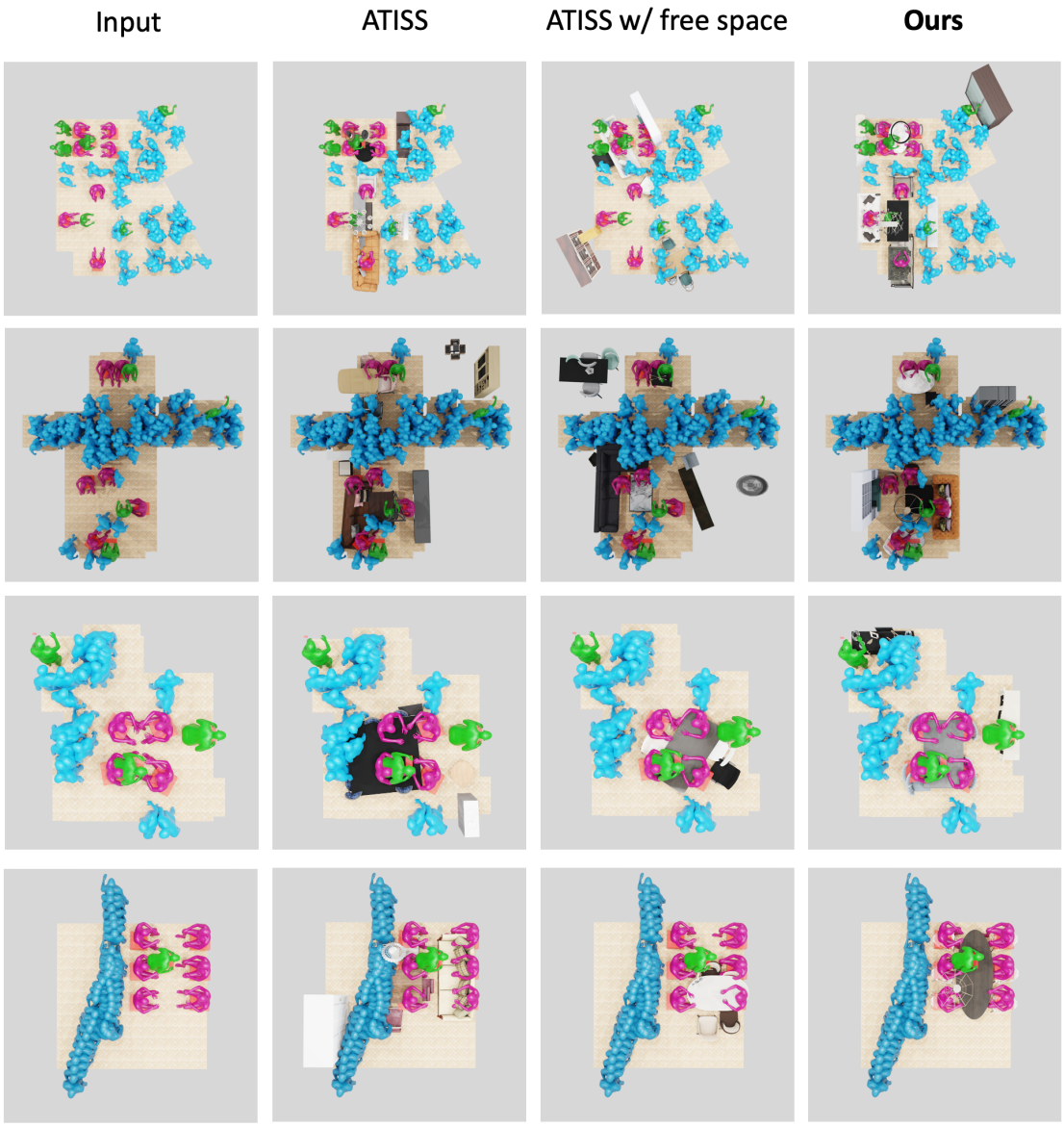


Figure 4.10: Qualitative comparison on living rooms (the first two rows) and dining rooms (the last two rows) in the test split of *3D FRONT HUMAN*. Given free space and contact humans as input, MIME generates more plausible scenes in which the contact humans interact with the contact objects and the free space humans have fewer collisions with all the generated objects. We also show the original ATISS w/ or w/o the free space mask as input. All results are w/o refinement. Each row represents an example input.

Following Pose2Room, we use the same 5 input motions and sample 10 scenes for each motion sequence, and report the mean value of the 3D IoU. Our method achieves better 3D object detection accuracy compared to Pose2Room *without pretraining Pose2Room on our dataset.*

4.4.2 Ablation Study on Input Humans

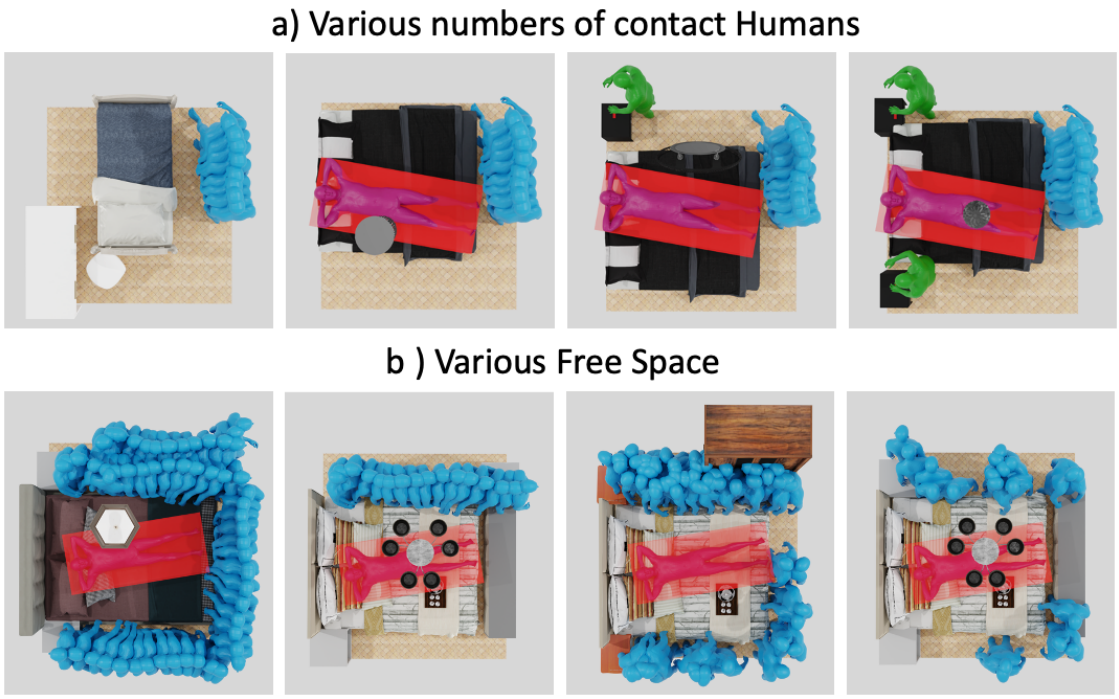


Figure 4.11: Ablation study on different numbers of contact humans and different density of free space humans. In (a), with more contact humans as input, the generated scenes contain more occupied objects. In (b), the more free space humans have in a room, the fewer objects are generated in a scene.

In Figure 4.11, we evaluate the influence of the density of free-space humans, and the number of contact humans, that we provide as input to MIME. We observe that MIME generates contact objects according to the number of contact humans and, as the density of free-space humans increases, MIME generates fewer objects in scenes. This is as expected.

4.5 Discussion

Given a sequence of human motions, MIME generates diverse and plausible scenes with which the humans interact. We assume that the generated scenes are static, and future

work should explore generating moving objects by investigating the interaction between humans and moving objects, such as moving a chair, grasping a cup, opening a door, and etc.

MIME, like ATISS, needs a pre-defined floor plan room layout as input. The resolution of the 2D floor plan is coarse; for instance, 1 pixel represents approximately 10 centimeters, which is extracted as a 512-dimension feature by ResNet-18. Introducing a finer floor plan representation, such as dividing a floor plan into multiple patches (cf. ViTRanftl *et al.* (2021)) or simply increasing the size of the feature dimension, could improve the generated object placement, resulting in less collision between the humans and the free space. Another interesting direction is to jointly estimate a floor plan, room category, and 3D object layout from input humans alone.

During inference, MIME uses a hand-crafted 2D IoU metric between the generated objects and the input contact humans to factor out which human is in contacted with which object. A simple extension would be to use the network to learn this information. Our model directly estimates 3D bounding boxes as a 3D scene representation, followed by a scene refinement that places the mesh models into the scene. Learning to directly estimate mesh models from interacting humans is another promising direction.

4.6 Conclusion

We introduced MIME, a method that generates diverse furniture layouts consistent with input human movements and contacts. To train MIME, we built a new dataset called *3D FRONT HUMAN*, by populating humans into the large-scale synthetic scene dataset Fu *et al.* (2021a). We demonstrated that by incorporating input human motion into free space and contact boxes, our method can generate multiple realistic scenes where the input motion can occur. MIME has numerous applications, particularly for generating synthetic training data at scale. MIME provides a means of taking existing human motion capture data and "upgrading" it to include plausible 3D scenes that are consistent with it.

Chapter 5

Generating Human Interaction Motions in Scenes with Text Control

5.1 Introduction

Apart from capturing human-scene interaction (chapter 3), we can generate human-scene interaction by synthesizing 3D scenes from input human motions (chapter 4). However, these input motions still need to be captured, and their variety and numbers are limited. Can we take the opposite approach, i.e., generating 3D humans directly from 3D scenes? Generating realistic human movements that can interact with 3D scenes is crucial for many applications, ranging from gaming to embodied AI. For example, character animators for games and films need to author motions that successfully navigate through cluttered scenes and realistically interact with target objects, while still maintaining artistic control over the style of the movement. One natural way to control style is through text, e.g., "skip happily to the chair and sit down". Recently, diffusion models have shown remarkable capabilities in generating human motion from user inputs. Text prompts Tevet *et al.* (2023); Zhang *et al.* (2022a) let users control style, while methods incorporating spatial constraints enable more fine-grained control, such as specifying desired joint positions and trajectories Xie *et al.* (2024); Shafir *et al.* (2023); Karunratanakul *et al.* (2023). However, these works have predominantly focused on characters in isolation, without considering environmental context or object interactions.

In this work, we aim to incorporate scene-awareness into user-controllable human motion generation models. However, learning to generate motions involving scene interactions is challenging, even without text prompts. Unlike large-scale motion capture datasets that depict humans in isolation Mahmood *et al.* (2019), datasets with paired examples of 3D human motion and scene/object geometry are limited. Prior work uses small paired datasets without text annotations to train VAEs Hassan *et al.* (2021b); Zhang *et al.* (2022b); Starke *et al.* (2019) or diffusion models Huang *et al.* (2023a); Pi *et al.* (2023) that generate human scene interactions with limited scope and diversity. Reinforcement learning methods are able to learn interaction motions from limited supervision Hassan *et al.* (2023); Zhao *et al.* (2023); Lee and Joo (2023), and can generate behaviors that are not present in the training motion dataset. However, designing reward

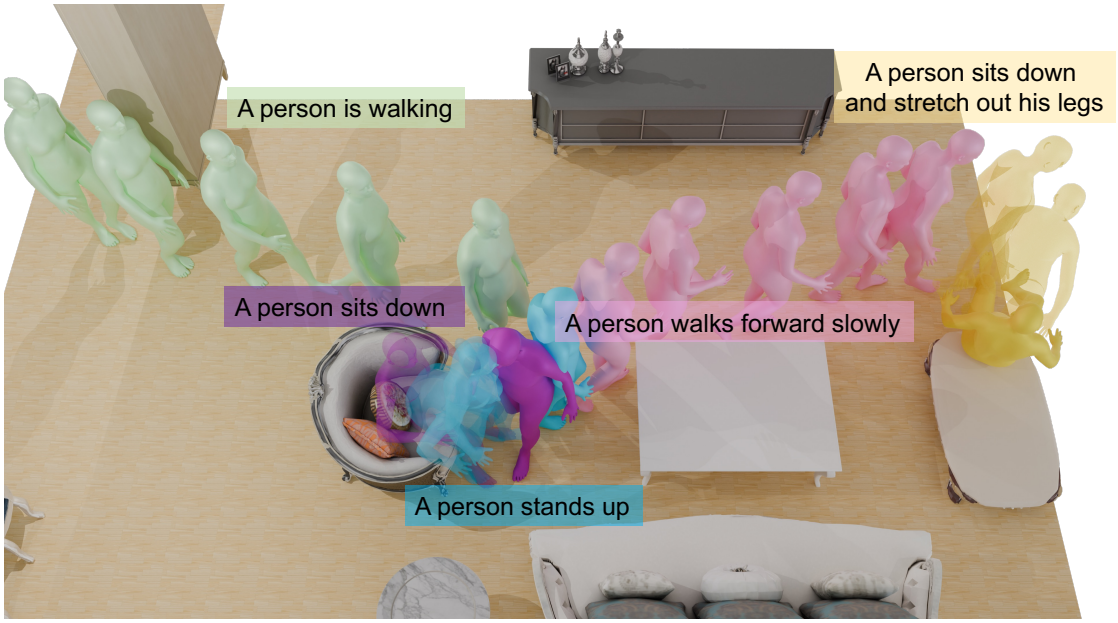


Figure 5.1: We present TeSMo, a method for generating diverse and plausible human-scene interactions from text input. Given a 3D scene, TeSMo generates scene-aware motions, such as walking in free space and sitting on a chair. Our model can be easily controlled using textual descriptions, start positions, and goal positions.

functions that lead to natural movements for a diverse range of interactions is difficult and tedious.

To address these challenges, we introduce a method for Text-conditioned Scene-aware Motion generation, called TeSMo. As shown in Figure 5.1, our method generates realistic motions that navigate around obstacles and interact with objects, while being conditioned on a text prompt to enable stylistic diversity. Our key idea is to combine the power of general, but scene-agnostic, text-to-motion diffusion models with paired human-scene data that captures realistic interactions. First, we pre-train a text-conditioned diffusion model Tevet *et al.* (2023) on a diverse motion dataset with no objects (e.g., HumanML3D Guo *et al.* (2022)), allowing it to learn a realistic motion prior and the correlation with text. We then fine-tune the model with an augmented scene-aware component that takes scene information as input, thereby refining motion outputs to be consistent with the environment.

Given a target object with which to interact and a text prompt describing the desired motion, we decompose the problem of generating a suitable motion in a scene into two components, *navigation* (e.g., approaching a chair while avoiding obstacles) and *interaction* (e.g., sitting on the chair). Both stages leverage diffusion models that are pre-trained on scene-agnostic data, then fine-tuned with an added scene-aware branch. The *navigation* model generates a pelvis trajectory that reaches a goal pose near the inter-

action object. During fine-tuning, the scene-aware branch takes, as input, a top-down 2D floor map of the scene and is trained on our new dataset containing locomotion sequences Mahmood *et al.* (2019) in 3D indoor rooms Fu *et al.* (2021b). The generated pelvis trajectory is then lifted to a full-body motion using motion in-painting Shafir *et al.* (2023). Next, the *interaction* model generates a full-body motion conditioned on a goal pelvis pose and a detailed 3D representation of the target object. To further improve generalization to novel objects, the model is fine-tuned using augmented data that re-targets interactions Hassan *et al.* (2021b) to a variety of object shapes while maintaining realistic human-object contacts.

Experiments demonstrate that our navigation approach outperforms prior work in terms of goal-reaching and obstacle avoidance, while producing full-body motions on par with scene-agnostic diffusion models Xie *et al.* (2024); Karunratanakul *et al.* (2023). Meanwhile, our interaction model generates motions with fewer object penetrations than the state-of-the-art approach Zhao *et al.* (2023), being preferred 71.9% of the time in a perceptual study. The central contributions of this method includes: (**1**) a novel approach to enable scene-aware and text-conditioned motion generation by fine-tuning an augmented model on top of a pre-trained text-to-motion diffusion model, (**2**) a method, TeSMo, that leverages this approach for navigation and interaction components to generate high-quality motions in a scene from text, (**3**) data augmentation strategies for placing navigation and interaction motions with text annotations realistically in scenes to enable scene-aware fine-tuning.

5.2 Text-Conditioned Scene-Aware Motion Generation

5.2.1 Overview

Given a 3D scene and a target interaction object, our goal is to generate a plausible human-scene interaction, where the motion style can be controlled by a user-specified text prompt. Our approach decomposes this task into two components, *navigation* and *interaction*, as illustrated in Figure 5.2. Both components are diffusion models that leverage a fine-tuning routine to enable scene-awareness without losing user controllability, as introduced in Section 5.2.2. To interact with an object, the character must first navigate to a location in the scene near the object, which is easily calculated heuristically or specified by the user, if desired. As described in Section 5.2.3, we design a hierarchical *navigation* model, which generates a root trajectory starting from an initial location that moves to the goal location while navigating around obstacles in the scene. The generated root trajectory is then lifted into a full-body motion using in-painting techniques Shafir *et al.* (2023); Xie *et al.* (2024). Since the navigation model gets close to the object in the first stage, to generate the actual object *interaction*, we can focus on scenarios where the character is already near the object. This allows a one-stage motion generation model that directly predicts the full-body motion from the starting pose (i.e, the last pose of

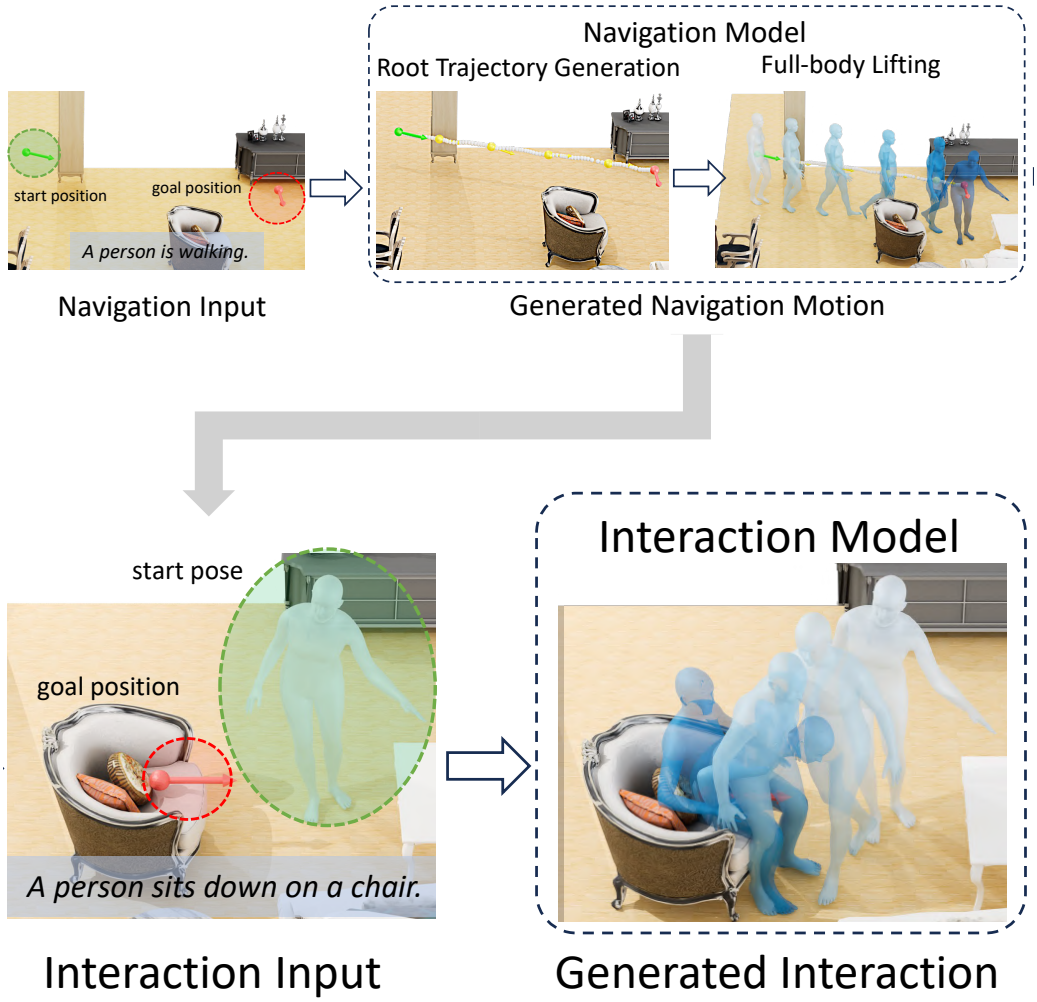
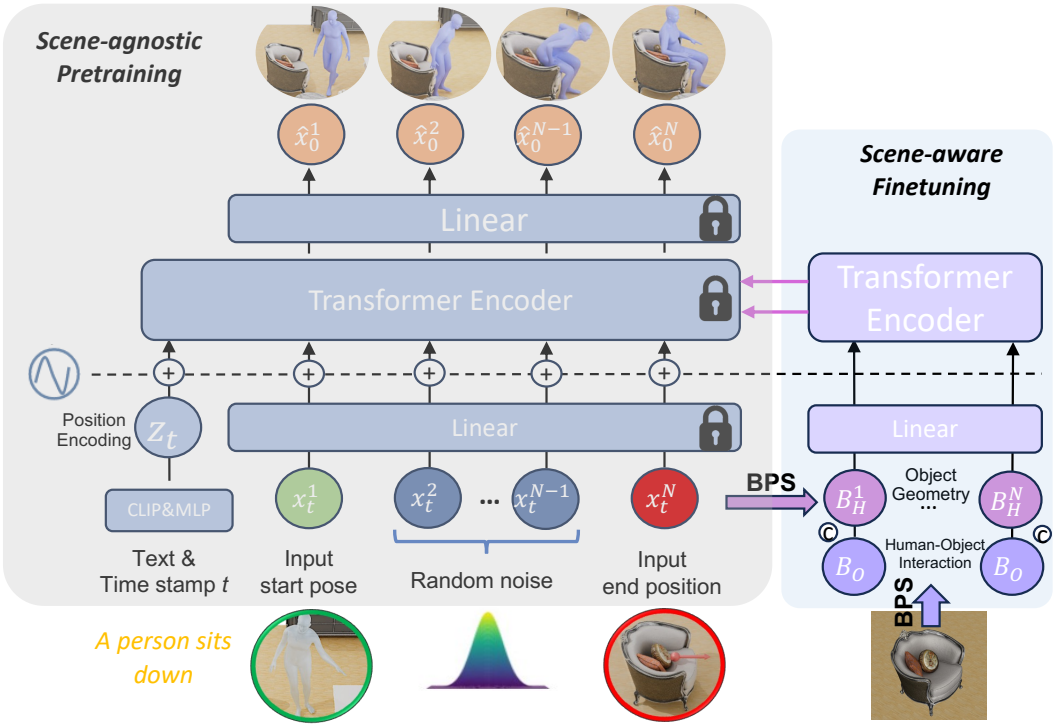


Figure 5.2: Pipeline overview: given the start position (green arrow), goal position (red arrow), 3D scene, and text description, the navigation root trajectory is first generated and then the full-body motion is completed through in-painting. Subsequently, the interaction is generated from a start pose (i.e., the end pose from navigation), the goal position, and the target object, enabling the generation of object-specific motion.

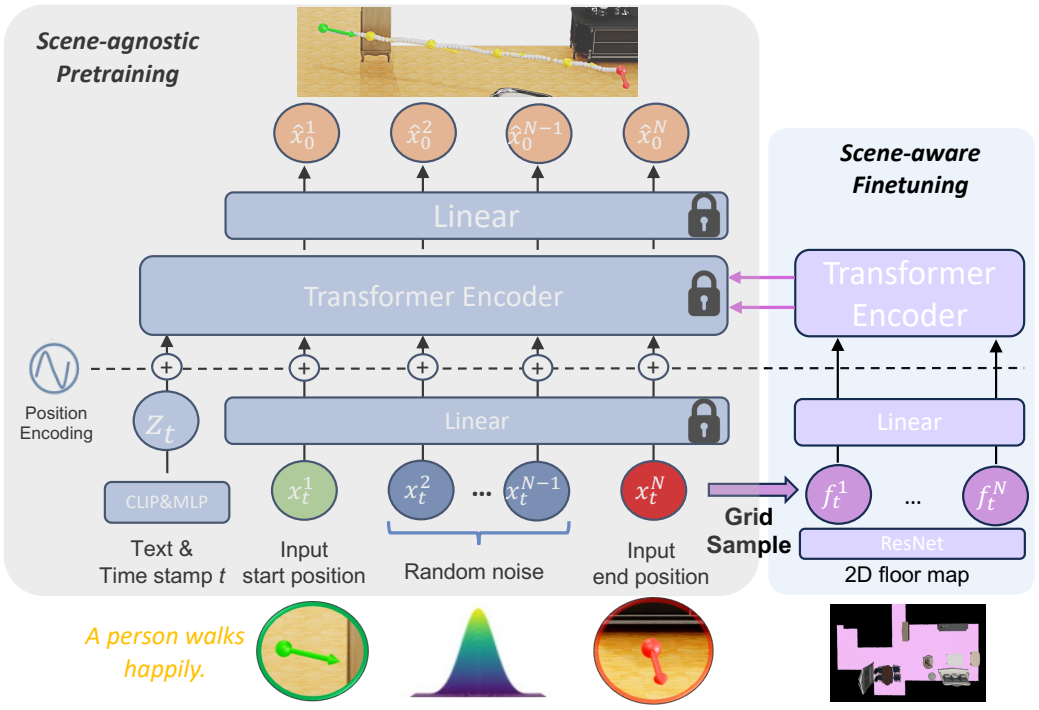
navigation), a goal pelvis pose, and the object (as detailed in Section 5.2.4).

5.2.2 Background: Controllable Human Motion Diffusion Models

Motion Diffusion Models. Diffusion models have been successfully used to generate both top-down trajectories Rempe *et al.* (2023) and full-body motions Tevet *et al.* (2023); Zhang *et al.* (2022a). These models generate motions by iteratively denoising



b) Interaction model



a) Root trajectory model

Figure 5.3: Network architecture of the (**a**) root trajectory model and (**b**) interaction motion model. Initially, the base transformer encoder is trained on scene-agnostic motion data using the start pose, target pose, and text as input. Subsequently, a scene-aware component is fine-tuned, which incorporates the 2D floor map (a) or 3D object (b).

a temporal sequence of N poses (e.g., root positions or full-body joint positions/angles) $\mathbf{x} = \left[\mathbf{x}^1, \ldots, \mathbf{x}^N\right]$. During training, the model learns to reverse a forward diffusion process, which starts from a clean motion $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, sampled from the training data, and after T diffusion steps is approximately Gaussian $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then at each step t of motion denoising, the reverse process is defined as:

$$p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c}) = \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\phi(\mathbf{x}_t, \mathbf{c}, t), \beta_t \mathbf{I}\right) \tag{5.1}$$

where $\mathbf{c}$ is some conditioning signal (e.g., a text prompt), and β_t depends on a pre-defined variance schedule. The denoising model μ_ϕ with parameters ϕ predicts the denoised motion $\hat{\mathbf{x}}_0$ from a noisy input motion $\mathbf{x}_t$ Ho *et al.* (2020). The model is trained by sampling a motion $\mathbf{x}_0$ from the dataset, adding random noise, and supervising the denoiser with a reconstruction loss $\|\mathbf{x}_0 - \hat{\mathbf{x}}_0\|^2$.

Augmented Controllability. In the image domain, general pre-trained diffusion models are specialized for new tasks using an augmented ControlNet Zhang *et al.* (2023) branch, which takes in a new conditioning signal and is fine-tuned on top of the frozen base diffusion model. OmniControl Xie *et al.* (2024) adapts this idea to the human motion domain. For motion diffusion models with a transformer encoder architecture, they propose an augmented transformer branch that takes in kinematic joint constraints (e.g., pelvis or other joint positions) and, at each layer, connects back to the base model through a linear layer that is initialized to all zeros.

As described in Secs. 5.2.3 and 5.2.4, our key insight is to use an augmented control branch to enable scene awareness. We first train a strong scene-agnostic motion diffusion model to generate realistic motion from a text prompt, and then fine-tune an augmented branch that takes scene information as input (e.g., a 2D floor map or 3D geometry). This new branch adapts generated motion to be scene-compliant, while still maintaining realism and text controllability.

Test-time Guidance. At test time, diffusion models can be controlled to meet specific objectives through guidance. We directly apply the guidance to the clean motion prediction from the model $\hat{\mathbf{x}}_0$ Rempe *et al.* (2023); Ho *et al.* (2022). At each denoising step, the predicted $\hat{\mathbf{x}}_0$ is perturbed with the gradient of an analytic objective function $\mathcal{J}$ as $\tilde{\mathbf{x}}_0 = \hat{\mathbf{x}}_0 - \alpha \nabla_{\mathbf{x}_t} \mathcal{J}(\hat{\mathbf{x}}_0)$ where α controls the strength of the guidance and $\mathbf{x}_t$ is the noisy input motion at step t. The predicted mean μ_ϕ is then calculated with the updated motion prediction $\tilde{\mathbf{x}}_0$ as in Rempe *et al.* (2023); Ho *et al.* (2022). As detailed later, we define guidance objectives for avoiding collisions and reaching goals.

5.2.3 Navigation Motion Generation

The goal of the navigation stage is for the character to reach a goal location near the target object using realistic locomotion behaviors that can be controlled by the user via text. We design a hierarchical method that first generates a dense root trajectory with a diffusion model, then leverages a powerful in-painting model Shafir *et al.* (2023) to generate a full-body motion for the predicted trajectory. This approach facilitates accurate goal-reaching with the root-only model while allowing diverse text control through the in-painting model.

Root Trajectory Generation. Our root trajectory diffusion model, shown in Figure 5.3 (a), operates on motions where each pose is specified by $\mathbf{x}^n = [x, y, z, \cos\theta, \sin\theta]_n$, with (x, y, z) being the pelvis position and θ the pelvis rotation, both of which are represented in the coordinate frame of the *first* pose in the sequence. The model is conditioned on a text prompt along with starting and ending goal positions and orientations. In contrast to the representation from prior work Guo *et al.* (2022), which uses relative pelvis velocity and rotation, our representation using absolute coordinates facilitates constraining the outputs of the model with goal poses.

Inspired by motion in-painting models Tevet *et al.* (2023); Shafir *et al.* (2023), given a start pose $\mathbf{s}$ and end goal pose $\mathbf{g}$, at each denoising step, we mask out the input $\mathbf{x}_t$ such that $\mathbf{x}_t^1 = \mathbf{s}$ and $\mathbf{x}_t^N = \mathbf{g}$, thereby providing clean goal poses directly to the model. To achieve this, a binary mask $\mathbf{m} = \left[\mathbf{m}^1, \ldots, \mathbf{m}^N\right]$ with the same dimensionality as $\mathbf{x}_t$ is defined, where $\mathbf{m}^1$ and $\mathbf{m}^N$ are a vector of 1's and all other $\mathbf{m}^n$ are 0's. During training, overwriting occurs with $\tilde{\mathbf{x}}_t = \mathbf{m} * \mathbf{x}_0 + (\mathbf{1} - \mathbf{m}) * \mathbf{x}_t$ where $*$ indicates element-wise multiplication and $\mathbf{x}_0$ is a ground truth root trajectory.

We then concatenate the mask with the overwritten motion $[\tilde{\mathbf{x}}_t; \mathbf{m}]$ and use this as input to the model to indicate which frames have been overwritten.

At test time, goal-reaching is improved using a guidance objective $\mathcal{J}_g = (\hat{\mathbf{x}}_0^N - \mathbf{g})^2$ that measures the error between the end pelvis position and orientation of the predicted clean trajectory $\hat{\mathbf{x}}_0^N$ and the final goal pose.

Incorporating Scene Representation. The model as described so far is trained on a locomotion subset of the HumanML3D dataset Guo *et al.* (2022) to enable generating realistic, text-conditioned root trajectories. However, it will be entirely unaware of the given 3D scene. To take the scene into account and avoid degenerating the text-following and goal-reaching performance, we augment the base diffusion model with a control branch that takes a representation of the scene as input. This scene-aware branch is a separate transformer encoder that is fine-tuned on top of the frozen base model. As input, we extract the walkable regions from the 3D geometry of the scene and project them to a bird's-eye view, yielding a 2D floor map $\mathcal{M}$. Following Rempe *et al.* (2023), a Resnet-18 He *et al.* (2016) encodes the map $\mathcal{M}$ as a feature grid, and at denoising step t, each 2D projected pelvis position $(x, z) \in x_t^n$ is queried in the feature grid $\mathcal{M}$ to get

the corresponding feature $\mathbf{f}_t^n$. The resulting sequence of features $\mathbf{f}_t = [\mathbf{f}_t^1, \ldots, \mathbf{f}_t^N]$, along with the text prompt and noisy motion $\mathbf{x}_t$, become the input to the separated transformer branch.

At test time, a collision guidance objective further encourages scene compliance. This is defined as $\mathcal{J}_c = \text{SDF}(\hat{\mathbf{x}}_0, \mathcal{M})$ where SDF calculates the 2D transform distance map from the 2D floor map, then queries the 2D distance value at each time step of the root trajectory. Positive distances, indicating pelvis positions outside the walkable region, are averaged to get the final collision loss.

Scene-aware training and data. To train the scene-aware branch, it is important to have a dataset featuring realistic motions navigating through scenes with corresponding text prompts. For this purpose, we create the **Loco-3D-FRONT** dataset by integrating locomotion sequences from HumanML3D into diverse 3D environments from 3D-FRONT Fu *et al.* (2021b). Each motion is placed within a different scene with randomized initial translation and orientation, following the methodology outlined in Yi *et al.* (2023b), as depicted in Figure 5.4(a). Additionally, we apply left-right mirroring to both the human motions and the 3D scenes with which they interact to augment the dataset Guo *et al.* (2022). This results in a dataset of approximately 9,500 walking motions, each motion accompanied by textual descriptions and 10 plausible 3D scenes on average, resulting in 95k locomotion-scene training pairs.

Added Control with Trajectory Blending. Our root trajectory diffusion model generates scene-aware motions and, unlike many prior works Hassan *et al.* (2021b); Zhao *et al.* (2023), does not require a navigation mesh to compute A* Hart *et al.* (1968) paths to follow. However, a user may want a character to take the shortest path to an object by following the A* path, or to control the general shape of the path by drawing a 2D route themselves. To enable this, we propose to fuse an input 2D trajectory $\mathbf{p} \in \mathbb{R}^{N\times 2}$ with our model's predicted clean trajectory at every denoising step. At step t, we extract the 2D (x,z) components $\hat{\mathbf{p}}_0$ from the predicted root trajectory $\hat{\mathbf{x}}_0$ and interpolate them with the input trajectory $\tilde{\mathbf{p}}_0 = s * \hat{\mathbf{p}}_0 + (1-s) * \mathbf{p}$ where s is the blending scale that controls how closely the generated trajectory matches the input. We then overwrite the 2D components of $\hat{\mathbf{x}}_0$ with $\tilde{\mathbf{p}}_0$ and continue denoising. This blending procedure ensures outputs roughly follow the desired path but still maintain realism inherent to the trained diffusion model.

Lifting to Full-body Poses. To lift the generated pelvis trajectory to a full-body motion, we leverage the existing text-to-motion in-painting method PriorMDM Shafir *et al.* (2023), which takes a dense 2D root trajectory as input. By using this strong model that is pre-trained for text-to-motion, we can effectively generate natural and scene-aware full-body motion, while offering diverse stylistic control through text.

5.2.4 Object-Driven Interaction Motion Generation

After navigation, the character has reached a location near the target object and next should execute a desired interaction motion. Due to the fine-grained relationship between the body and object geometry during interactions, we propose a single diffusion model to directly generate full-body motion, unlike the two-stage navigation approach from Section 5.2.3.

Interaction Motion Generation. The interaction motion model operates on a sequence of full-body poses and is shown in Figure 5.3(b). Our pose representation extends that of HumanML3D Guo *et al.* (2022) to add the absolute pelvis position and heading $(x, y, z, \cos\theta, \sin\theta)$, similar to our navigation model. Each pose in the motion is $\mathbf{x}^n = \left[x,\ y,\ z,\ \sin\theta,\ \cos\theta,\ \dot{r}^a,\ \dot{r}^x,\ \dot{r}^z,\ r^y,\ \mathbf{j}^p,\ \mathbf{j}^v,\ \mathbf{j}^r,\ \mathbf{c}^f\right]_n \in \mathbb{R}^{268}$ with $\dot{r}^a$ root angular velocity, $(\dot{r}^x, \dot{r}^z)$ root linear velocity, r^y root height, $\mathbf{c}^f$ foot contacts, and $\mathbf{j}^p$, $\mathbf{j}^v$, $\mathbf{j}^r$ the local joint positions, velocities, and rotations, respectively.

The model is conditioned on a text prompt along with a starting full-body pose (i.e, the final pose of the navigation stage) and a final goal pelvis position and orientation. The goal pelvis pose can usually be computed heuristically, but may also be provided by the user or predicted by another network Hassan *et al.* (2021b). The same masking procedure described in Section 5.2.3 is used to pass the start and end goals as input to the model. At test time, we also use the same goal-reaching guidance to improve the accuracy of hitting the final pelvis pose.

Object representation. The base interaction diffusion model is first trained on a dataset of interaction motions from HumanML3D and SAMP Hassan *et al.* (2021b) without any objects, which helps develop a strong prior on interaction movements driven by text prompts. Similar to navigation, we then augment the base model with a new object-aware transformer encoder and fine-tune this encoder separately.

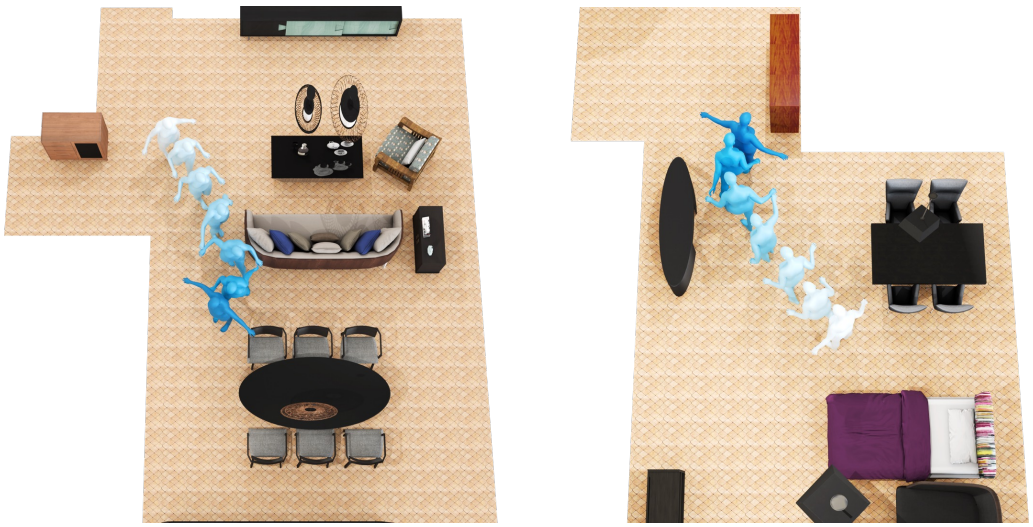
For the input to this branch at each denoising step t, we leverage Basis Point Sets (BPS) Prokudin *et al.* (2019) to calculate two key features: object geometry and the human-object relationship. First, a sphere with a radius of 1.0m is defined around the object's center, and 1024 points are randomly sampled inside this sphere to form the BPS. The distance between each point in the BPS and the object's surface is then calculated, capturing the object's geometric features and stored as $\mathbf{B}_O \in \mathbb{R}^{1024}$. Next, for each body pose $\mathbf{x}_t^n$ at timestep n in the noisy input sequence, we calculate the minimum distance from each BPS point to any body joint, giving $\mathbf{B}^n \in \mathbb{R}^{1024}$. The resulting sequence of features $\mathbf{B}_H = \left[\mathbf{B}^1, \ldots, \mathbf{B}^N\right]$ represents the human-object relationship throughout the entire motion. Finally, the object and human-object interaction features are concatenated with the original pose representation at each timestep $[\mathbf{x}_t^n;\ \mathbf{B}^n;\ \mathbf{B}_O]$ and fed to an MLP to generate a merged representation, which serves as the input to the scene-aware branch.

At test time, a collision objective is used to discourage penetrations between the human and object. This is very similar to the collision loss described in Section 5.2.3, but

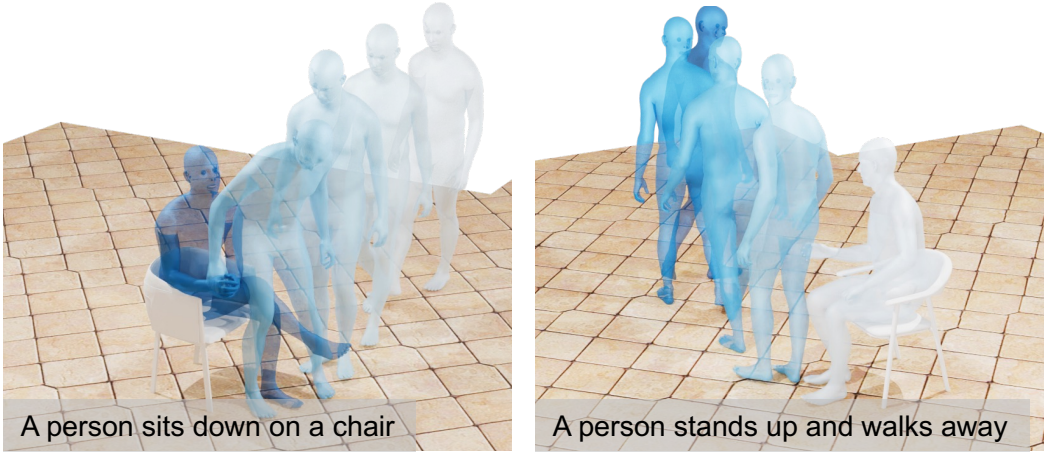
the SDF volume is computed for the 3D object and body vertices that are inside the object are penalized. Noteably, our interaction motion generation model outputs 3D joint positions. We begin by selecting random points (vertices) on the surface of the SMPL mesh, which represents a human body in its default A-pose. These selected surface vertices are then mapped to corresponding positions on an internal skeletal structure (or inside skeleton), which is also in the A-pose. This linking creates a connection between specific points on the mesh surface and the underlying skeletal framework. Once this initial linking is complete, we can apply any new pose to the internal skeleton. As the skeleton changes pose, the connected vertices on the SMPL mesh surface follow the movements of the skeleton, resulting in the mesh surface adopting the new pose as well. This setup allows us to efficiently update the positions of these sampled vertices for any new pose without recalculating the mapping each time, ensuring consistent movement with the inside skeleton. This is defined as $\mathcal{J}_c = \text{SDF}(\hat{\mathbf{x}}_0, \mathcal{S}_{\mathcal{O}})$ where SDF calculates the SDF volume of the object O, then queries the sign distance value at each time step of the body vertices. Positive distances, indicating body vertices inside the interactive object, are averaged to get the final collision loss.

Scene-aware Training and Data. To train the scene-aware branch, we utilize the SAMP dataset Hassan *et al.* (2021b), which captures motions and objects simultaneously. Specifically, we focus on "sitting" and "stand-up" interactions extracted from 80 sitting motion sequences in the SAMP dataset involving chairs of varying heights, as shown in Figure 5.4(b). To diversify the object geometry, we randomly select objects from 3D-FRONT Fu *et al.* (2021b) to match the contact vertices on human poses in the original SAMP motion sequences. This matching is achieved using the contact loss and collision loss techniques outlined in MOVER Yi *et al.* (2022).

The original SAMP motions are often lengthy ($\sim$100 sec) and lack paired textual descriptions. For instance, a "sit" motion sequence involves walking to an object, sitting down, standing up, and moving away. To effectively learn individual skills, we extract sub-sequences containing specific interactions that begin or end with a sitting pose, such as "walk then sit", "stand up then sit", "stand up from sitting", and "walk from sitting." Furthermore, we annotate textual descriptions for each sub-sequence, which often incorporate the style of sitting poses, such as "a person walks and sits down on a chair while crossing their arms." Applying left-right data augmentation to motion and objects results in approximately 200 sub-sequences for each motion sequence, each paired with corresponding text descriptions and featuring various objects.



a) Loco-3D-Front: locomotion in different rooms



b) Interaction with different objects and text description

Figure 5.4: (**a**) Loco-3D-FRONT contains locomotion placed in 3D-FRONT Fu *et al.* (2021b) scenes without collisions. (**b**) We augment SAMP Hassan *et al.* (2021b) by randomly selecting chairs from 3D-FRONT to match the motions and annotating a text description for each sub-sequence.

5.3 Experimental Evaluation

5.3.1 Implementation Details

Training. The scene-agnostic branch of our navigation model is trained on the 3D motions and text descriptions from the Loco-3D-FRONT dataset for 420k optimization steps. Subsequently, we freeze the base model weights and fine-tune the scene-aware branch, with additional 2D-floor map inputs, for a further 20k steps. Similarly, the scene-agnostic base of our interaction model is first trained on a mix of HumanML3D Guo *et al.* (2022) and SAMP Hassan *et al.* (2021b) data without objects for 400k steps. Then, the object-aware branch is fine-tuned on our text-annotated SAMP data with 3D object inputs for an additional 20k steps.

Test-time Guidance. For the navigation model, we set the guidance weight α to 30 for goal-reaching guidance and 1000 for collision guidance. In the interaction model, we utilize weights of 1000 for goal-reaching loss and 10 for the collision SDF loss. To ensure smooth generation results, we exclude the inference guidance at the final time step of denoising. For a fair comparison with baselines, we do *not* use inference guidance unless explicitly stated in the experiment.

5.3.2 Evaluation Data and Metrics

Navigation. Navigation performance is assessed using the test set of Loco-3D-FRONT, comprising roughly 1000 sequences. Our metrics evaluate the generated root trajectory and the full-body motion after in-painting separately. For the root trajectory, we measure goal-reaching accuracy for the 2D (horizontal *xz*) root **position** (m), **orientation** (rad), and root **height** (m). The **collision ratio**, the fraction of frames within generated trajectories where a collision occurs, evaluates the consistency of root motions with the environment. For the full-body motion after in-painting, we use common metrics from prior work Guo *et al.* (2022). **FID** measures the realism of the motion, **R-precision** (top-3) evaluates the consistency between the text and motion, and **diversity** is computed based on the average pairwise distance between sampled motions. Additionally, the **foot skating** ratio Karunratanakul *et al.* (2023) evaluates the physical plausibility of motion-ground interaction by the proportion of frames where either foot slides a distance greater than a specified threshold (2.5 cm) while in contact with the ground (foot height $<$5 cm).

Interactions. To evaluate full-body human-object interactions, we use the established test split of the SAMP dataset Hassan *et al.* (2021b), which contains motions related to sitting. Similar to navigation, we analyze goal-reaching accuracy through position, orientation, and height errors. Furthermore, we assess physical plausibility by computing average **penetration values** and **penetration ratios** between the generated motion and

interaction objects. The penetration value is the mean SDF value across all interpenetrated body vertices of the generated motions, while the ratio is the fraction of generated poses containing penetrations (i.e., SDF values < -3 cm) over all generated motion frames.

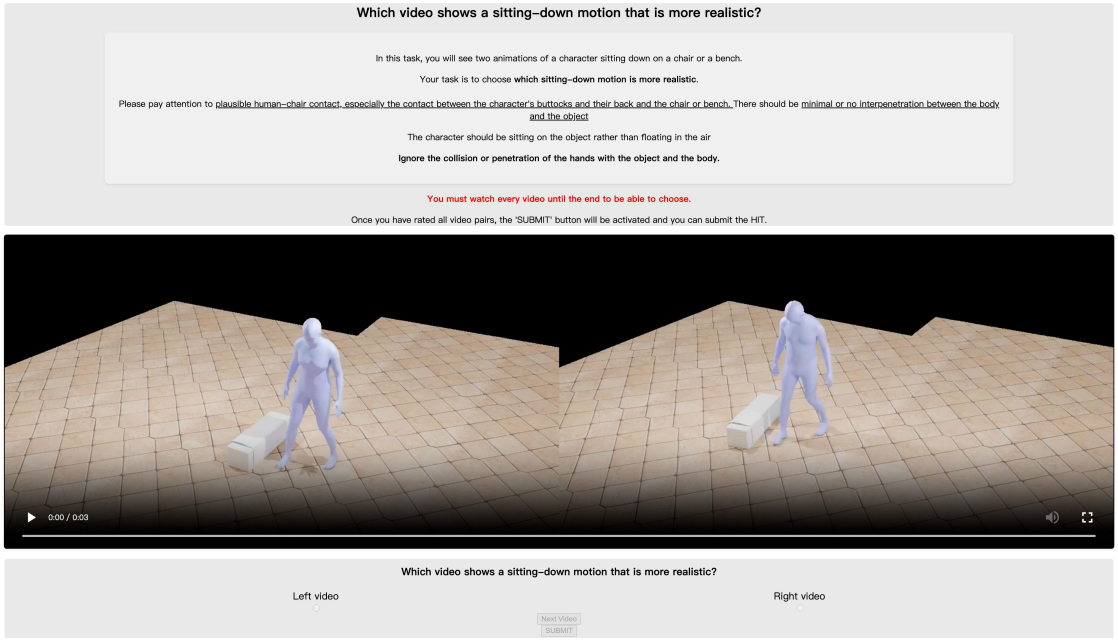


Figure 5.5: The layout of our perceptual study for evaluating the plausibility of human-object interaction.

To evaluate the plausibility of human-object interaction, we perform a **user study** to compare our method and DIMOS Zhao *et al.* (2023). We employ Amazon Mechanical Turk (AMT) Amazon Web Services, Inc. (2024) to solicit assessments from 30 individuals. Raters are presented with two side-by-side videos of generated interactions and asked which is more realistic, particularly focusing on the contact between the character's buttocks and their back with the chair or bench, and the presence of minimal or no interpenetration between the body and the object. We present 70 test videos with the positions of our generated videos and DIMO's results randomly shuffled horizontally. In order to filter out poor responses, we duplicated our 5 test examples where clear preferences between two video results were evident, serving as catch trials. Ultimately, we obtained 65 useful responses out of 70 raters. The full survey page is illustrated in Figure 5.5. The perceptual study reveals a distinct preference for motions generated by our approach (preferred 71.9%) over those produced by DIMOS.

5.3.3 Comparisons

Navigation. We conduct a comparative analysis of our method with previous scene-aware and scene-agnostic motion generation approaches, shown in Table 5.1. Every method is conditioned on a text prompt along with a start and end goal pose, as described in Section 5.2.3. The TRACE baseline and our method MOVER also receive the 2D-floor map as input.

	Root trajectory evaluation				**Full-body motion evaluation**			
	Goal-reaching error ↓							
Method	Pos.	Orient.	Height	Collision ↓	FID ↓	R-precision ↑	Diversity ↑	Foot skating ↓
Ground Truth	-	-	-	-	0.010	0.672	7.553	0.000
GMDKarunratanakul *et al.* (2023)	0.374	1.231	-	-	**13.160**	0.114	4.488	0.181
OmniContolXie *et al.* (2024)	1.226	1.018	1.159	-	22.930	**0.458**	**7.128**	0.094
TRACE Rempe *et al.* (2023)	0.205	0.152	0.010	0.055	22.669	0.144	6.501	0.058
Ours (1-stage train)	0.197	0.132	0.013	**0.028**	22.372	0.152	6.347	0.062
Ours	**0.169**	**0.119**	**0.008**	0.031	20.465	0.376	6.415	**0.056**

Table 5.1: Evaluation of navigation motion generation on the Loco-3D-FRONT test set. (**Left**) For generated pelvis trajectories, our approach achieves the best goal-reaching accuracy with low collision rate. (**Right**) After in-painting the full-body motion, our method maintains diverse and realistic motion that aligns with the given text prompt, and is competitive with diffusion-based scene-agnostic GMD and OmniControl.

	Goal-reaching error ↓			**Object penetration ↓**		**User study**
Method	Pos.	Height	Orient.	Value	Ratio	**preference ↑**
DIMOS Zhao *et al.* (2023)	0.2020	0.1283	0.4731	0.0193	0.1076	29.1%
Ours	**0.1445**	**0.0120**	**0.2410**	**0.0043**	**0.0611**	**71.9%**

Table 5.2: Evaluation of human-object interaction motion generation on the SAMP Hassan *et al.* (2021b) sitting test set. Compared to DIMOS, our approach reaches the goal pose more accurately and exhibits fewer object penetrations, leading to superior performance in the user perceptual study.

We first compare to GMD Karunratanakul *et al.* (2023) and OmniControl Xie *et al.* (2024), previous scene-agnostic text-to-motion diffusion models trained on HumanML3D to follow a diverse range of kinematic motion constraints. GMD utilizes the horizontal pelvis positions (x, z) of both the start and end goals to generate a dense root trajectory and subsequently the full-body motion. OmniControl takes as input the horizontal pelvis positions (x, z) along with the height y to directly generate full-body motion in a single stage. Our navigation model achieves better goal-reaching accuracy, e.g., 16.9 cm for root position, since it is trained specifically for the goal-reaching locomotion task. More importantly, in the right half of Table 5.1 the full-body motion from our method after in-painting is comparable in terms of realism, text-following, and diversity, while achieving

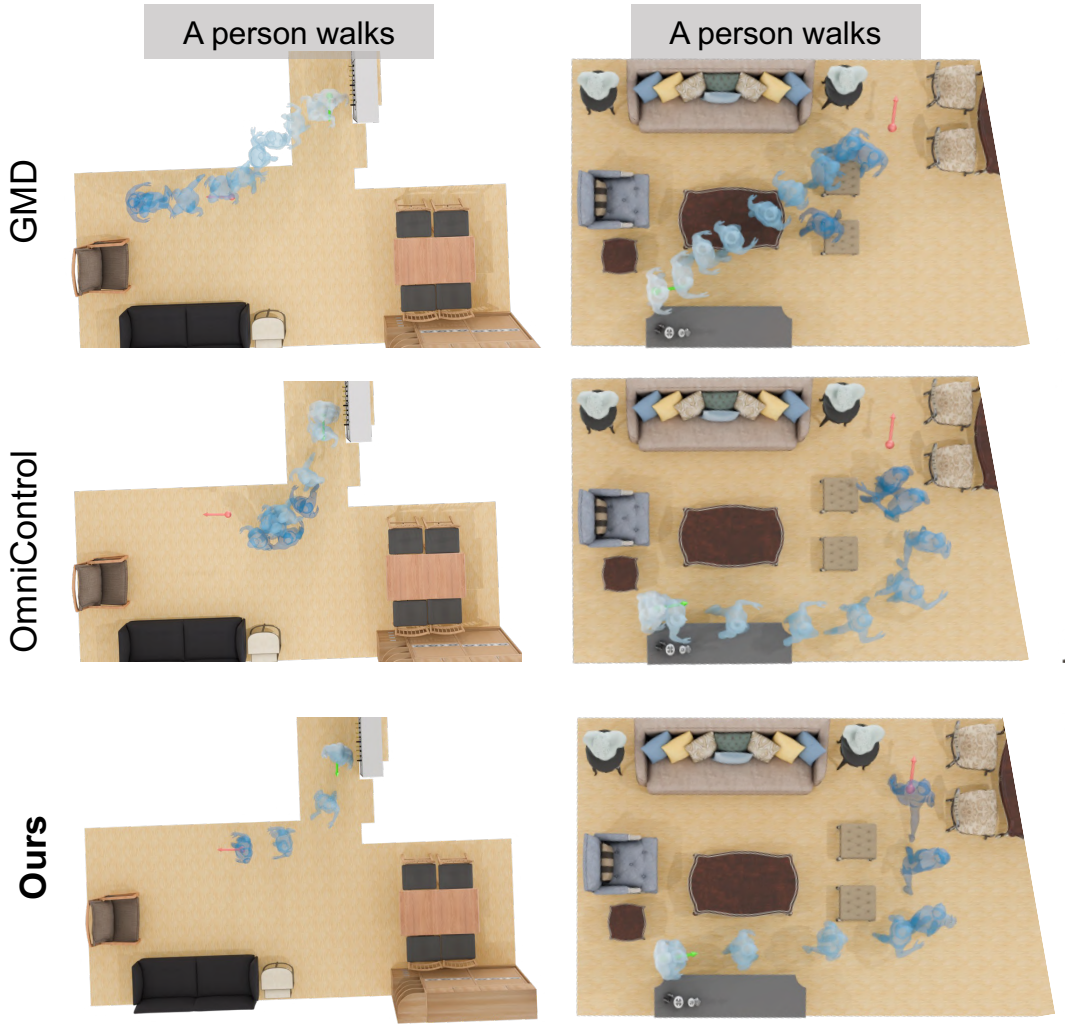


Figure 5.6: Navigation generation performance. The start pose is the green arrow, and the goal pose is the red arrow. Our method more accurately reaches the goal and avoids obstacles while the style is controlled by a text prompt.

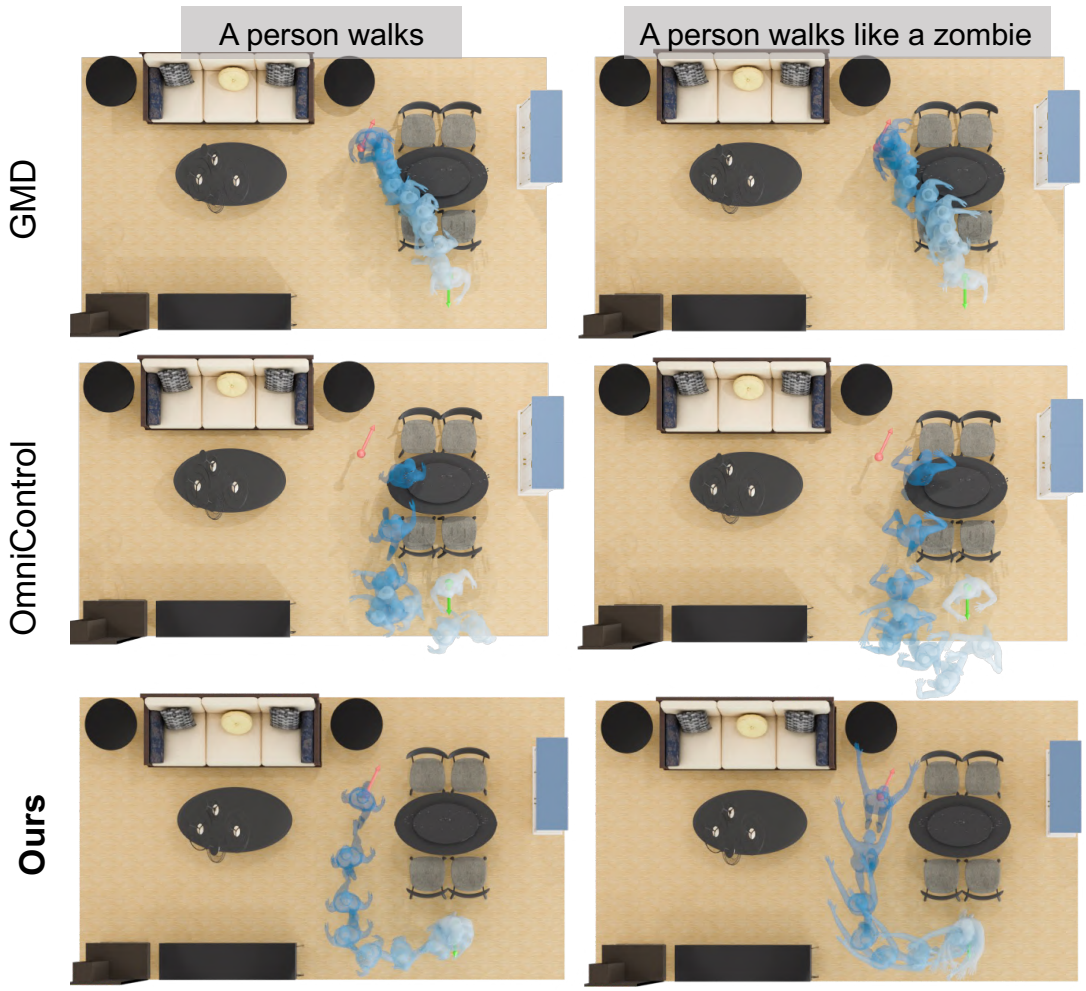


Figure 5.7: More results of navigation generation. The start pose is the green arrow, and the goal pose is the red arrow. Our method more accurately reaches the goal and avoids obstacles while the style is controlled by a text prompt.

the best foot skating results. This demonstrates that our approach adds scene-awareness to locomotion generation, without compromising realism or text control.

To evaluate the importance of our two-branch model architecture, we adopt TRACE Rempe *et al.* (2023), a recent root trajectory generation model designed to take a 2D map of the environment as input. The adapted TRACE architecture is very similar to our model in Figure 5.3(a), but instead of using a separate scene-aware branch, the base transformer directly takes the encoded 2D-floor map features as input. This results in a single-branch architecture that must be trained from scratch, as opposed to our two-branch fine-tuning approach. Table 5.1 reveals that our method generates more plausible root trajectories with fewer collisions and more accurate goal-reaching. We also see that training our full two-branch architecture from scratch (*1-stage train* in Table 5.1),

instead of using pre-training and then fine-tuning, degrades both goal-reaching and final full-body motion after in-painting.

A qualitative comparison of generated motions in different rooms is shown in Figure 5.6 and Figure 5.7. GMD tends to generate simple walking-straight trajectories. OmniControl and GMD do not reach the goal pose accurately and ignore the surroundings, leading to collisions with the environment. Our method TeSMo is able to generate diverse locomotion styles controlled by text in various scenes, achieving superior goal-reaching accuracy compared to other methods.

Interaction. Table 5.2 compares our approach to DIMOS Zhao *et al.* (2023), a state-of-the-art method to generate interactions trained with reinforcement learning. DIMOS requires a full-body final goal pose as input to the policy, unlike our approach which uses just the pelvis pose. Despite this, DIMOS struggles to reach the goal accurately, likely due to error accumulation during autoregressive rollout. Our method exhibits fewer instances of interpenetration with interaction objects and the perceptual study reveals a distinct preference for motions generated by our approach (preferred 71.9%) over those produced by DIMOS. Figure 5.8 compares the approaches qualitatively, where we see that more accurate goal-reaching reduces floating or penetrating the chair during sitting. Moreover, the interactions generated by DIMOS lack diversity, and cannot be conditioned on text.

Guidance		**Navigation**		**Interaction**		
Goal Reach	Collision	Goal Pos.	Collision	Goal Pos.	Pen. Val.	Pen. Ratio
✗	✗	0.1568	0.0294	0.1445	0.0043	0.0611
✓	✗	**0.118**	0.0342	0.1453	0.0050	0.0554
✗	✓	0.1550	0.0013	0.1407	**0.0040**	**0.0414**
✓	✓	0.1241	**0.0012**	**0.1404**	0.0045	0.0494

Table 5.3: Test-time guidance evaluation. Adding guidance to reach goal poses and avoid collisions during inference improves performance. Lower is better for all metrics.

5.3.4 Analysis of Capabilities

In Figure 5.1, our method carries out a sequence of actions, enabling traversal and interaction with multiple objects within a scene. Figure 5.9 demonstrates additional key capabilities. In the top section, our method is controlled through a variety of text prompts. For interactions in particular, diverse text descriptions disambiguate between actions like sitting or standing up, and allow stylizing the sitting motion, e.g., with crossed arms. In the middle section, we enable user control over trajectories by adhering to a predefined A* path. By adjusting the blending scale, users can adjust how closely the generated

trajectory follows A*. At the bottom of Figure 5.9, we harness guidance at test time to encourage motions to reach the goal while avoiding collisions and penetrations. As shown in Table 5.3, combining guidance losses gives improved results both for navigation and interactions.

5.3.5 Ablation Study

Method	**Goal-reaching error ↓**			**Collision ↓**
	Pos.	**Orient.**	**Height**	
Ours (OmniControl in-painting)	0.459	0.999	0.090	0.073
Ours (full-body rep)	0.844	0.016	0.110	0.124
Ours	**0.169**	**0.119**	**0.008**	**0.031**
	FID ↓	**R-precision ↑**	**Diversity ↑**	**Foot skating ↓**
Ours (OmniControl in-painting)	17.927	0.396	6.288	0.0308
Ours (full-body rep)	24.642	0.189	6.967	0.169
Ours	20.465	0.376	6.415	0.056

Table 5.4: Ablation study comparing various full-body infilling methods and different representations of navigation motion generation using the Loco-3D-FRONT test set. (**Left**) For generated pelvis trajectories, our approach achieves the best goal-reaching accuracy with low collision rate. (**Right**) After in-painting the full-body motion, our method preserves diverse and realistic movements that align with the provided text prompt, much like the model employing an alternative OminiControl full-body inpainting technique. However, our approach distinctly outperforms the model utilizing full-body representation.

Alternative Full-Body In-painting Approach. While our root trajectory generation approach can integrate with several motion in-painting techniques, here we use PriorMDM Shafir *et al.* (2023). As an alternative, we evaluate our method using OminiControl Xie *et al.* (2024) for in-painting in Table 5.4. However, OmniControl overrides our generated dense pelvis trajectory and jointly generates full-body locomotion with a new pelvis trajectory. This severely degrades the goal-reaching ability (from 0.169 cm to 0.459 cm) as demonstrated in Table 5.4. Therefore, we choose to utilize PriorMDM as our body motion in-painting method. It aligns well with our generated trajectory, resulting in the generation of plausible locomotion while maintaining adherence to the goal position.

One-stage Navigation Motion Generation. To evaluate the efficacy of our two-stage navigation model design, we compare to a single-stage full-body motion generation abla-

tion of our model. This model operates on the same input data but directly generates full-body locomotion. However, as shown in Table 5.4, this approach limits goal-reaching ability and does not produce motion styles that align with the input text. The local poses are somewhat dissociated from the global pelvis trajectories, allowing trajectory variations while maintaining the same motion style. For instance, individuals can walk along different paths while maintaining consistency in their motion style.

5.4 Discussion

We introduced TeSMo, a novel method for text-controlled scene-aware motion generation. By first pre-training a scene-agnostic text-to-motion diffusion model on large-scale motion capture data and subsequently fine-tuning with a scene-aware component, our text-conditioned method enables generating realistic and diverse human-object interactions within 3D scenes. To support such training, we introduced the new Loco-3D-FRONT dataset containing realistic navigation motions placed in 3D scenes, and extended the SAMP dataset with additional objects and text annotations. Experiments demonstrate that our generated motion is on par with state-of-the-art diffusion models, while improving the plausibility and realism of interactions compared to prior work.

Limitations & Future Work. While our navigation model enables accurate goal-reaching and text-to-motion controllability, the two-stage process can sometimes lead to a disconnect between the generated pelvis trajectory and in-painted full-body poses. Exploring new one-stage models, capable of simultaneously generating pelvis trajectories and poses, would streamline the process. Additionally, our current approach, which operates on 2D-floor maps, restricts the ability to handle intricate interactions, such as a person stepping over a small stool.

Our current approach is aimed at controllability to allow users to specify text prompts or goal objects and locations. However, our method may also fit into recently proposed pipelines Xiao *et al.* (2024) that employ LLM planners to specify a sequence of actions and contact information that could be used to guide our motion generation. Looking ahead, we also aim to broaden the spectrum of actions modeled by the system, to encompass activities such as lying down and touching. Furthermore, enabling interactions with dynamic objects will allow more interactive and realistic scenarios.

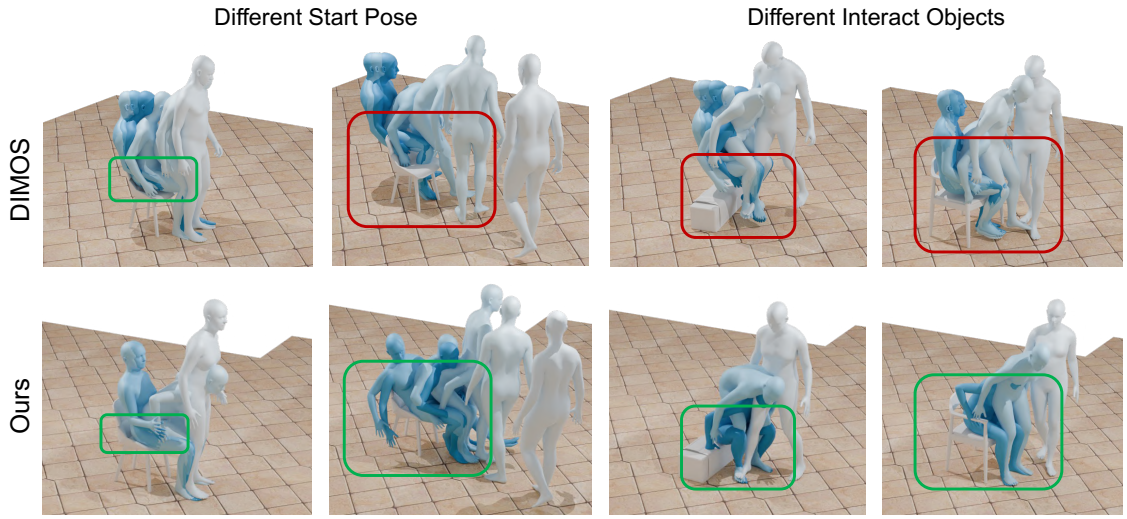


Figure 5.8: Compared with DIMOS Zhao *et al.* (2023), our method generates more realistic human-object interactions with reduced floating and interpenetrations.

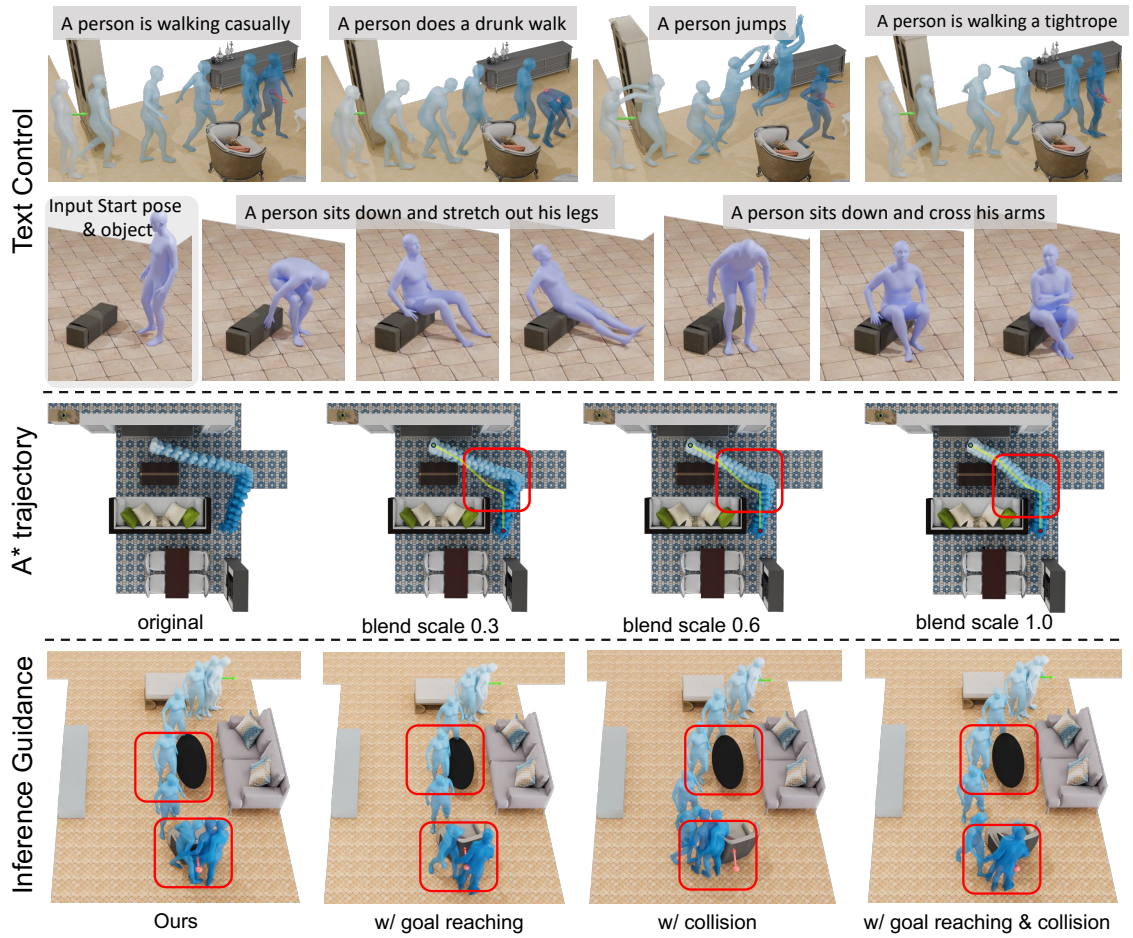


Figure 5.9: TeSMo capabilities. (**Top**) Diverse text control; (**Middle**) Following an A* path, the blending scale controls adherence by blending the A* path with the randomly initial noise in the input. (**Bottom**) Test-time guidance encourages locomotion to reach the goal accurately without colliding with the environment.

Chapter 6

Conclusion and Future Work

Throughout this thesis, we have explored human-scene interaction by focusing on the constraints that influence movement and behavior within 3D environments. Through the study of depth ordering, collision avoidance, and contact consistency, we have developed models that significantly enhance the accuracy and realism of scene reconstructions in computer vision. These constraints inform the spatial arrangement of elements within a scene and guide the generation of plausible human actions and interactions.

Capturing Human-scene Interactions. In Chapter 3, we established a robust framework for understanding and reconstructing 3D environments through the lens of human-scene interactions (HSIs). By harnessing HSIs derived from monocular video footage, we have developed a method (MOVER) that not only reconstructs a scene with enhanced accuracy but also refines 3D human pose estimations within that scene. The primary achievement of this work lies in its ability to produce a consistent, physically plausible scene layout, which significantly surpasses the capabilities of existing methods.

Our approach's core revolves around optimizing three specific HSI constraints: depth ordering, non-interpenetration, and contact consistency. These constraints are vital as they allow for the logical arrangement of objects in relation to human interaction, thereby ensuring a reconstruction that respects both the spatial and physical realities of the scene.

For future work, several promising directions exist. First, extending the model to incorporate dynamic interactions and transient scene elements could extend MOVER to more complex and variable environments. Second, enhancing the computational efficiency of the model could allow for real-time processing applications, which are crucial for areas such as augmented reality and robotic navigation. Lastly, expanding the datasets to include more varied interaction scenarios could improve the generalization of the model, making it adaptable to a wider range of real-world applications. Continuing to build on the foundation laid by this thesis will enable significant advances in both theoretical and practical aspects of computer vision and human-scene interaction studies.

Generating Human-scene Interactions by Scenes from Humans. Capturing human-scene interactions is labor-intensive, we explore to generate human-scene interaction. In Chapter 4, we introduce MIME, a novel method capable of generating 3D indoor

scenes informed by human motion. Utilizing an autoregressive transformer architecture, MIME effectively synthesizes furniture layouts that are coherent with captured human movements and interactions. This approach leverages human motion as a dynamic input to model free space and object interactions, such as sitting or touching, which helps generate more realistic and functionally accurate 3D scenes.

While MIME represents a significant advance in scene generation from human motion, it currently operates under the constraint of static scenes. Future iterations could explore the inclusion of dynamic objects to simulate more complex interactions like moving a chair or opening a door. Additionally, the resolution of the 2D floor plan used as input is relatively coarse, and enhancing this by employing a finer floor plan representation or enlarging the feature dimensions could further improve object placement and reduce collisions.

Another promising direction is to develop methods for joint estimation of floor plans and 3D object layouts directly from human inputs, potentially eliminating the need for predefined layouts. Enhancing the model's capability to directly estimate mesh models from interacting humans could streamline the generation process, providing a more integrated and accurate scene reconstruction.

Our contributions not only pave the way for generating synthetic training data at scale but also have potential applications in gaming, architecture, and virtual reality, where realistic human-scene interactions are crucial. The current model uses a hand-crafted metric for object placement, but a more sophisticated approach would involve learning this placement directly through the network, enhancing the model's ability to adapt to various scene configurations and human interactions.

Overall, MIME offers a transformative approach to 3D scene generation, providing a foundation for further research in creating interactive and dynamically rich virtual environments. Future work will focus on these expansions and refinements to push the boundaries of what is achievable with generative scene modeling technologies.

Generating Human-scene Interactions by Humans from Scenes. Apart from generating scenes based on human-scene interactions, we also explore the opposite approach: generating human motions from 3D scenes. In Chapter 5, we introduced TeSMo, a method for text-controlled scene-aware motion generation leveraging denoising diffusion models. This approach represents an advancement over prior methods by integrating a scene-agnostic text-to-motion model pre-trained on extensive motion capture datasets with a scene-aware component fine-tuned on our new Loco-3D-FRONT dataset. Our experiments demonstrate that TeSMo achieves a high degree of realism and diversity in human-object interactions, matching and, in some aspects, surpassing current state-of-the-art diffusion models.

Despite these achievements, there are notable limitations and areas for improvement. The current two-stage process occasionally results in discrepancies between the generated pelvis trajectory and the full-body poses, suggesting the potential benefits of

developing a more integrated one-stage model. Additionally, the reliance on 2D floor maps limits the system's ability to handle more complex interactions, such as navigating around or over small objects.

Looking forward, we plan to explore several avenues to enhance the capabilities of TeSMo. One primary focus will be the development of models that allow for seamless generation of both pelvis trajectories and full-body poses in a single step. We also aim to extend the variety of actions the system can model, such as incorporating motions like lying down and interacting with dynamic objects, which would increase the realism and interactivity of the generated scenes. Furthermore, integrating our method into frameworks that utilize large language model planners for action sequencing could provide more detailed and contextually appropriate motion generation.

6.1 Long-term Future Work

In the future, the overarching aim of this thesis is to establish a robust feedback loop between reconstruction and generation to enhance both the realism of generated interactions and the accuracy of reconstructed human motions and scenes. Specifically, the goal is to leverage 3D human-scene interaction generation to create realistic videos that feed back into the reconstruction process. This feedback would act as a flywheel: accurate reconstruction provides richer 3D data on human-scene interactions, which, in turn, helps train a more sophisticated 3D human-scene interaction generation model. The missing element in this cycle is the generation of realistic videos that incorporate these interactions. By integrating 3D human-scene interactions into the generated videos, we can improve our understanding of human-scene interaction perception. Furthermore, this approach will enable us to extend the scope from human-scene interactions to human-human interactions, thereby demonstrating nuanced social abilities such as conversational dynamics and assisting behaviors, which will further enrich the system. This iterative process will lead to an ever-improving system where both reconstruction and generation inform and refine each other. This section outlines our long-term research objectives and the methods we propose to explore in order to achieve these ambitious goals.

Synthesizing Realistic Videos in Physics Rendering Engine with HSIs. To synthesize realistic videos, we aim to integrate human-scene interactions into physics-based rendering. This integration will enhance the realism of human motions within various 3D environments, effectively bridging the gap between digital human representations and their physical surroundings. Leveraging the achievements of the BEDLAM dataset, as shown in Figure 6.1, which incorporates highly realistic simulations of diverse body shapes, motions, skin tones, and detailed clothing animated through physics simulation, we aim to further enhance the authenticity of these interactions. By extracting and analyzing human motion data, and rendering these motions within physics-simulated scenes,

Figure 6.1: BEDLAM is a large-scale synthetic video dataset designed to train and test algorithms on the task of 3D human pose and shape estimation (HPS). BEDLAM contains diverse body shapes, skin tones, and motions. Beyond previous datasets, BEDLAM has SMPL-X bodies with hair and realistic clothing animated using physics simulation.

we expect to produce videos that not only display realistic human movements but also show authentic interactions with the environment. Future work will focus on refining these interactions to capture complex dynamics, such as variable lighting, texture interactions, and the physical impact of human activities on the surroundings, thus pushing the boundaries of current video realism in synthetic datasets In our ongoing efforts to synthesize realistic videos, we aim to integrate human-scene interactions into physics-based rendering. This integration will enhance the realism of human motions within various 3D environments, effectively bridging the gap between digital human representation and their physical surroundings. Leveraging the achievements of the BEDLAM dataset, as shown in Figure 6.1, which incorporates highly realistic simulations of diverse body shapes, motions, skin tones, and detailed clothing animated through advanced physics simulations, we aim to enhance the authenticity of these interactions further. By extracting and analyzing human motion data, and rendering these motions within physics-simulated scenes, we expect to produce videos that not only display realistic human movements but also show authentic interactions with the environment. Future work will focus on refining these interactions to capture complex dynamics, such as variable lighting, texture interactions, and the physical impact of human activities on the surroundings, thus pushing the boundaries of current video realism in synthetic datasets

Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.

Prompt: Several giant wooly mammoths approach treading through a snowy meadow, their long wooly fur lightly blows in the wind as they walk, snow covered trees and dramatic snow capped mountains in the distance, mid afternoon light with wispy clouds and a sun high in the distance creates a warm glow, the low camera view is stunning capturing the large furry mammal with beautiful photography, depth of field.

Figure 6.2: Sora OpenAI (2024) is an AI video generation model that generates realistic and imaginative scenes from textual instructions.

Synthesizing Realistic Videos based on Diffusion Models. Recent diffusion-based video generation methods have achieved notable advances, enabling the creation of high-

quality videos. Sora OpenAI (2024), a text-to-video model, excels in generating intricate scenes with detailed, accurate representations of multiple characters and dynamic motions, as shown in Figure 6.2. Sora also demonstrates realistic interaction capabilities within videos, though it faces challenges with simulating complex physical interactions and cause-and-effect scenarios, such as not reflecting physical changes when an object is manipulated. Nonetheless, its strengths in handling diverse and complex video content make it well-suited for systems aimed at creating photorealistic videos with detailed human-scene interactions and social dynamics.

Additionally, recent diffusion-based methods like EMO Tian *et al.* (2024) and Champ Zhu *et al.* (2024) showcase promising results for human-centric video generation. Both build on the text-to-image model Stable Diffusion (SD) 1.5 , incorporating extra temporal layers inspired by AnimateDiff to enable the generation of talking or pose-controlled videos, as seen in Figure 6.3. Specifically, EMO generates vocal avatar videos from a single reference image and vocal audio input, achieving synchronized facial expressions and head poses with audio cues. Meanwhile, Champ enhances the shape and pose alignment by integrating 3D depth, normal, and semantic maps derived from the SMPL model within the video generation process.

These diffusion-based methods pave the way for the next generation of video generation engines, offering powerful tools for generating realistic videos that authentically reflect human-scene interactions. By further developing these methods, we can bridge the gap between reconstructed and generated human-scene interactions, enabling seamless realism that aligns closely with the dynamic cues and interactions present in the input signals.

Incorporating Social Ability. Humans are inherently social creatures, interacting not only with our environment but also with each other Freud (1923). To enhance these interactions digitally, the development of social avatars represents a significant step towards creating digital humans capable of realistic interactions. We leverage advancements in modeling the translation from human speech to body motion Yi *et al.* (2023a) and neural rendering Zheng *et al.* (2023), incorporating abilities like communication as depicted in Figure 6.4. By generating expressive 3D motions—from body and hand gestures to facial expressions—based solely on audio cues, we bring digital avatars to life, synchronizing their expressions with speech to enable engaging and believable dialogues and social interactions.

Additionally, as shown Figure 6.5, the "Generative Proxemics" Müller *et al.* (2024) models human-human interactions by generating two individuals in close social proximity, enabling realistic 3D reconstructions from images. This technology is crucial for creating scenarios where digital representations mimic real-life interactions without manual annotations. Complementing this, the "Assistance in Human Interactions" highlighted in the Watch-And-Help Puig *et al.* (2021) challenge involves collaborative efforts where one agent (Bob) observes another (Alice), discerns her objectives and aids her in

Figure 6.3: Illustration of EMO Tian *et al.* (2024) and Champ Zhu *et al.* (2024), two diffusion-based image-to-video generation frameworks. EMO takes a single reference image and vocal audio input (e.g., talking or singing) to produce vocal avatar videos, emphasizing expressive facial expressions and dynamic head poses, synchronized with audio cues. Champ employs a 3D human parametric model (SMPL) within a latent diffusion framework to enhance alignment between body shape and motion, generating 3D human pose-controlled videos.

a new environment to achieve cooperative interaction and goal accomplishment.

For future work, we aim to further refine the integration of social abilities within digital avatars, exploring broader aspects of human interaction such as empathy, conflict resolution, and cooperative tasks. Advancements in models for emotion and gesture recognition, combined with AI-driven behavioral prediction, will enhance the authenticity and responsiveness of digital humans. This evolution will pave the way for avatars that not only interact naturally with their environment but also exhibit complex interpersonal dynamics that mirror human social behavior.

In conclusion, the long-term goals outlined in this thesis chart a path toward advancing digital human interaction modeling and generation in virtual environments. By synthesizing realistic videos that capture the nuances of real-world interactions and enhancing the social capabilities of avatars, this work aims to bridge digital and real-world social experiences. These developments lay the groundwork for immersive applications across multiple domains: in gaming, they can enable more responsive, lifelike NPCs; in virtual reality, they create more authentic, engaging social spaces; and in interactive media, they support storytelling with dynamic, human-centered characters.

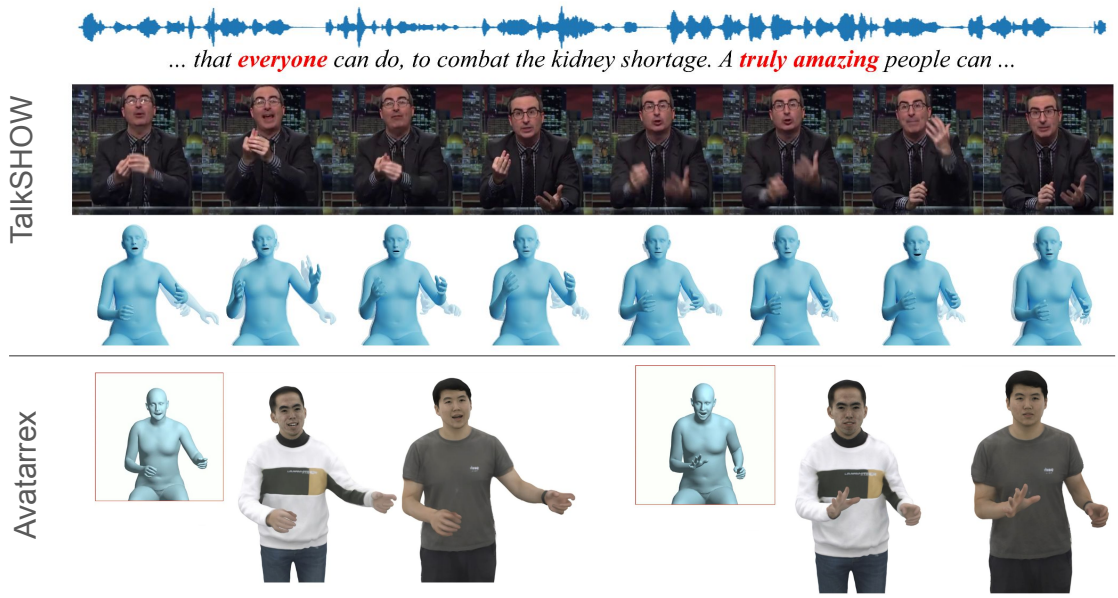


Figure 6.4: Given a human speech, TalkSHOW Yi *et al.* (2023a) first generates realistic sequences of body poses, hand gestures, and facial expressions. Then, AvatarReX Zheng *et al.* (2023) takes the output from TalkSHOW to animate NeRF-based full-body avatars, to produce high-quality images with detailed and realistic appearance.

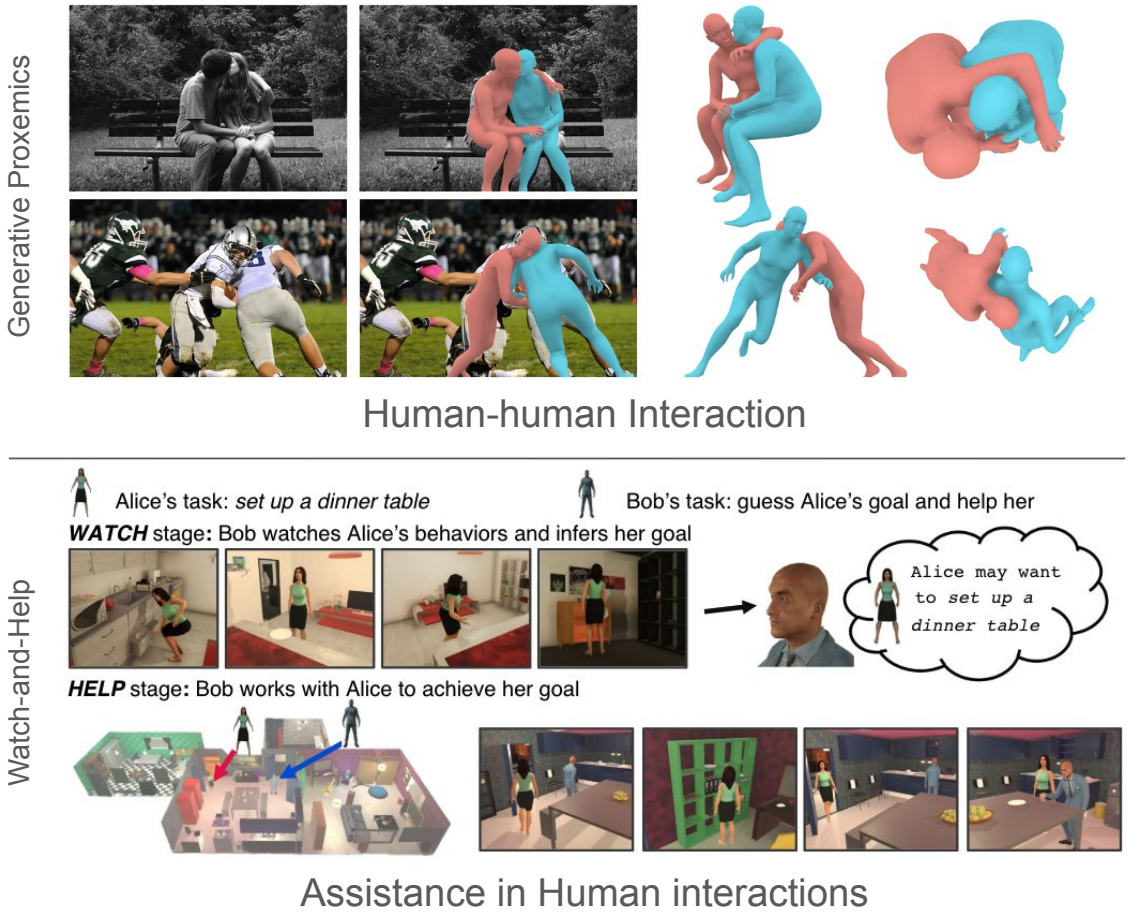


Figure 6.5: "Generative Proxemics" Müller *et al.* (2024) models human-human interactions by simulating two individuals in close proximity, enabling realistic 3D reconstructions from images without manual annotations. Additionally, "Assistance in Human Interactions" Puig *et al.* (2021) within the Watch-And-Help challenge illustrates collaborative efforts, where agent Bob observes Alice to identify her objectives, and then assists in completing the task in a new setting, highlighting teamwork and effective cooperation.

List of Tables

List of Figures

Bibliography

Agrawal, S. and van de Panne, M. (2016). Task-based locomotion. *Transactions on Graphics (TOG)*, **35**(4), 1–11.

Ahn, H., Ha, T., Choi, Y., Yoo, H., and Oh, S. (2018). Text2action: Generative adversarial synthesis from language to action. In *International Conference on Robotics and Automation (ICRA)*, pages 5915–5920. IEEE.

Amazon Web Services, Inc. (Accessed 2024). Amazon mechanical turk.

Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., and Schiele, B. (2018). Posetrack: A benchmark for human pose estimation and tracking. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5167–5176.

Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., and Davis, J. (2005). SCAPE: Shape Completion and Animation of PEople. *Transactions on Graphics (TOG)*, **24**(3), 408–416.

Athanasiou, N., Petrovich, M., Black, M. J., and Varol, G. (2022). TEACH: Temporal Action Compositions for 3D Humans. In *International Conference on 3D Vision (3DV)*, pages 414–423.

Bansal, A., Russell, B., and Gupta, A. (2016). Marr revisited: 2D-3D alignment via surface normal prediction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5965–5974.

Bazavan, E. G., Zanfir, A., Zanfir, M., Freeman, W. T., Sukthankar, R., and Sminchisescu, C. (2021). HSPACE: Synthetic parametric humans animated in complex environments. *arXiv*.

Bhatnagar, B. L., Xie, X., Petrov, I., Sminchisescu, C., Theobalt, C., and Pons-Moll, G. (2022). BEHAVE: Dataset and method for tracking human object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 15935–15946.

Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., and Black, M. J. (2016). Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, volume 9909, pages 561–578.

Božič, A., Palafox, P., Thies, J., Dai, A., and Nießner, M. (2021). TransformerFusion: Monocular RGB scene reconstruction using transformers. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 1403–1414.

Cai, Z., Zhang, M., Ren, J., Wei, C., Ren, D., Lin, Z., Zhao, H., Yang, L., and Liu, Z. (2021). Playing for 3d human recovery. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Cao, Z., Gao, H., Mangalam, K., Cai, Q., Vo, M., and Malik, J. (2020). Long-term human motion prediction with scene context. In *European Conference on Computer Vision (ECCV)*, pages 387–404.

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2021). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **43**(1), 172–186.

Cao, Z., Radosavovic, I., Kanazawa, A., and Malik, J. (2021). Reconstructing hand-object interactions in the wild. In *International Conference on Computer Vision (ICCV)*, pages 12417–12426.

Chang, A., Monroe, W., Savva, M., Potts, C., and Manning, C. D. (2015). Text to 3d scene generation with rich lexical grounding. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., and Zhang, Y. (2017a). Matterport3D: Learning from rgb-d data in indoor environments. In *International Conference on 3D Vision (3DV)*.

Chang, A. X., Eric, M., Savva, M., and Manning, C. D. (2017b). Sceneseer: 3d scene design with natural language. *arXiv*.

Chao, Y.-W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y. S., Van Wyk, K., Iqbal, U., Birchfield, S., *et al.* (2021a). Dexycb: A benchmark for capturing hand grasping of objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9044–9053.

Chao, Y.-W., Yang, J., Chen, W., and Deng, J. (2021b). Learning to sit: Synthesizing human-chair interactions via hierarchical control. In *Conference on Artificial Intelligence (AAAI)*, pages 5887–5895.

Chen, Y., Huang, S., Yuan, T., Zhu, Y., Qi, S., and Zhu, S.-C. (2019). Holistic++ scene understanding: Single-view 3D holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *International Conference on Computer Vision (ICCV)*, pages 8647–8656.

Choi, W., Chao, Y.-W., Pantofaru, C., and Savarese, S. (2013). Understanding indoor scenes using 3D geometric phrases. In *Computer Vision and Pattern Recognition (CVPR)*, pages 33–40.

Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., and Black, M. J. (2020). Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, volume 12355, pages 20–40.

Choy, C. B., Xu, D., Gwak, J., Chen, K., and Savarese, S. (2016). 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *European Conference on Computer Vision (ECCV)*, volume 9912, pages 628–644.

CMU Graphics Lab (2000). CMU Graphics Lab Motion Capture Database. `http://mocap.cs.cmu.edu/`.

Comer, M., Bouman, C., and Simmons, J. (2010). Statistical methods for image segmentation and tomography reconstruction. *Microscopy and Microanalysis*, **16**, 1852 – 1853.

Corona, E., Pumarola, A., Alenya, G., and Moreno-Noguer, F. (2020). Context-aware human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6992–7001.

Dabral, R., Shimada, S., Jain, A., Theobalt, C., and Golyanik, V. (2021). Gravity-aware monocular 3d human-object reconstruction. In *International Conference on Computer Vision (ICCV)*, pages 12365–12374.

Dahnert, M., Hou, J., , Nießner, M., and Dai, A. (2021). Panoptic 3D scene reconstruction from a single RGB image. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 8282–8293.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. (2017). ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839.

Dai, Y., Lin, Y., Lin, X., Wen, C., Xu, L., Yi, H., Shen, S., Ma, Y., and Wang, C. (2023). Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. In *Computer Vision and Pattern Recognition (CVPR)*, pages 682–692.

Dasgupta, S., Fang, K., Chen, K., and Savarese, S. (2016). DeLay: Robust spatial layout estimation for cluttered indoor scenes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 616–624.

Denninger, M. and Triebel, R. (2020). 3d scene reconstruction from a single viewport. In *European Conference on Computer Vision (ECCV)*, pages 51–67. Springer.

Devaranjan, J., Kar, A., and Fidler, S. (2020). Meta-sim2: Learning to generate synthetic datasets. In *European Conference on Computer Vision (ECCV)*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*.

Di, X., Yu, P., Zhu, H., Cai, L., Sheng, Q., Sun, C., and Ran, L. (2020). Structural plan of indoor scenes with personalized preferences. In *European Conference on Computer Vision (ECCV)*, pages 455–468. Springer.

Diller, C. and Dai, A. (2024). Cg-hoi: Contact-guided 3d human-object interaction generation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 19888–19901.

Dwivedi, S. K., Schmid, C., Yi, H., Black, M. J., and Tzionas, D. (2024). POCO: 3D pose and shape estimation using confidence. In *International Conference on 3D Vision (3DV)*, pages 85–95.

Eigen, D., Ranzato, M., and Sutskever, I. (2013). Learning factored representations in a deep mixture of experts. *arXiv: Learning*.

Fieraru, M., Zanfir, M., Oneata, E., Popa, A.-I., Olaru, V., and Sminchisescu, C. (2020). Three-dimensional reconstruction of human interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7214–7223.

Fieraru, M., Zanfir, M., Oneata, E., Popa, A.-I., Olaru, V., and Sminchisescu, C. (2021). Learning complex 3D human self-contact. In *Conference on Artificial Intelligence (AAAI)*, pages 1343–1351.

Fisher, M., Ritchie, D., Savva, M., Funkhouser, T., and Hanrahan, P. (2012). Example-based synthesis of 3d object arrangements. *Transactions on Graphics (TOG)*, **31**(6), 1–11.

Fisher, M., Savva, M., Li, Y., Hanrahan, P., and Nießner, M. (2015). Activity-centric scene synthesis for functional 3d scene modeling. *Transactions on Graphics (TOG)*, **34**(6), 1–13.

Freud, S. (1923). *The Ego and the Id*. Hogarth Press, London.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., and Tao, D. (2018). Deep ordinal regression network for monocular depth estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2002–2011.

Fu, H., Cai, B., Gao, L., Zhang, L.-X., Wang, J., Li, C., Zeng, Q., Sun, C., Jia, R., Zhao, B., *et al.* (2021a). 3d-front: 3d furnished rooms with layouts and semantics. In *International Conference on Computer Vision (ICCV)*, pages 10933–10942.

Fu, H., Cai, B., Gao, L., Zhang, L.-X., Wang, J., Li, C., Zeng, Q., Sun, C., Jia, R., Zhao, B., *et al.* (2021b). 3d-front: 3d furnished rooms with layouts and semantics. In *International Conference on Computer Vision (ICCV)*, pages 10933–10942.

Fu, H., Jia, R., Gao, L., Gong, M., Zhao, B., Maybank, S., and Tao, D. (2021c). 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision (IJCV)*, pages 1–25.

Gabeur, V., Franco, J.-S., Martin, X., Schmid, C., and Rogez, G. (2019). Moulding Humans: Non-parametric 3D human shape estimation from single images. In *International Conference on Computer Vision (ICCV)*, pages 2232–2241.

Gibson, J. J. (2014). *The ecological approach to visual perception: classic edition*. Psychology press.

Gkioxari, G., Malik, J., and Johnson, J. (2019). Mesh R-CNN. In *International Conference on Computer Vision (ICCV)*, pages 9785–9795.

Godard, C., Mac Aodha, O., and Brostow, G. J. (2017). Unsupervised monocular depth estimation with left-right consistency. In *Computer Vision and Pattern Recognition (CVPR)*, pages 270–279.

Groueix, T., Fisher, M., Kim, V. G., Russell, B. C., and Aubry, M. (2018). Atlasnet: A papier-mâché approach to learning 3d surface generation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 216–224.

Guler, R. A. and Kokkinos, I. (2019). HoloPose: Holistic 3D human reconstruction in-the-wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10884–10894.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161.

Guzov, V., Mir, A., Sattler, T., and Pons-Moll, G. (2021a). Human POSEitioning system (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors. In *Computer Vision and Pattern Recognition (CVPR)*.

Guzov, V., Mir, A., Sattler, T., and Pons-Moll, G. (2021b). Human poseitioning system (HPS): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4318–4329.

Guzov, V., Jiang, Y., Hong, F., Pons-Moll, G., Newcombe, R., Liu, C. K., Ye, Y., and Ma, L. (2024). HMD^2: Environment-aware motion generation from single egocentric head-mounted device. *arXiv*.

Hart, P. E., Nilsson, N. J., and Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, **4**(2), 100–107.

Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., and Seidel, H.-P. (2009). A statistical model of human pose and body shape. *Computer Graphics Forum*, **28**(2), 337–346.

Hassan, M., Choutas, V., Tzionas, D., and Black, M. J. (2019). Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision (ICCV)*, pages 2282–2292.

Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., and Black, M. J. (2021a). Populating 3D scenes by learning human-scene interaction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 14708–14718.

Hassan, M., Ceylan, D., Villegas, R., Saito, J., Yang, J., Zhou, Y., and Black, M. J. (2021b). Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11374–11384.

Hassan, M., Guo, Y., Wang, T., Black, M., Fidler, S., and Peng, X. B. (2023). Synthesizing physical character-scene interactions. In *ACM SIGGRAPH Conference Proceedings*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask R-CNN. In *International Conference on Computer Vision (ICCV)*, pages 2961–2969.

Hedau, V., Hoiem, D., and Forsyth, D. (2009). Recovering the spatial layout of cluttered rooms. In *International Conference on Computer Vision (ICCV)*, pages 1849–1856.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 30.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. (2022). Video diffusion models. *arXiv preprint*.

Holden, D., Komura, T., and Saito, J. (2017). Phase-functioned neural networks for character control. *ACM Transactions on Graphics*, page 1–13.

Huang, C.-H. P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., and Black, M. J. (2022). Capturing and inferring dense full-body human-scene contact. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285.

Huang, S., Qi, S., Xiao, Y., Zhu, Y., Wu, Y. N., and Zhu, S.-C. (2018a). Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 207–218.

Huang, S., Qi, S., Zhu, Y., Xiao, Y., Xu, Y., and Zhu, S.-C. (2018b). Holistic 3D scene parsing and reconstruction from a single RGB image. In *European Conference on Computer Vision (ECCV)*, volume 11211, pages 194–211.

Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., and Zhu, S.-C. (2023a). Diffusion-based generation, optimization, and planning in 3d scenes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 16750–16761.

Huang, Y., Kaufmann, M., Aksan, E., Black, M. J., Hilliges, O., and Pons-Moll, G. (2018c). Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *Transactions on Graphics (TOG)*, **37**(6), 1–15.

Huang, Y., Yi, H., Liu, W., Wang, H., Wu, B., Wang, W., Lin, B., Zhang, D., and Cai, D. (2023b). One-shot implicit animatable avatars with model-based priors. In *International Conference on Computer Vision (ICCV)*, pages 8974–8985.

Huang, Y., Yi, H., Xiu, Y., Liao, T., Tang, J., Cai, D., and Thies, J. (2024). TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *International Conference on 3D Vision (3DV)*.

Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014). Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **36**(7), 1325–1339.

Izadinia, H., Shan, Q., and Seitz, S. M. (2017). IM2CAD. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5134–5143.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, page 79–87.

Jiang, H., Liu, S., Wang, J., and Wang, X. (2021). Hand-object contact consistency reasoning for human grasps generation. In *International Conference on Computer Vision (ICCV)*, pages 11107–11116.

Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., and Daniilidis, K. (2020). Coherent reconstruction of multiple humans from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588.

Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., and Luo, P. (2020). Whole-body human pose estimation in the wild. In *European Conference on Computer Vision (ECCV)*, volume 12354, pages 196–214.

Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T. S., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., and Sheikh, Y. (2017). Panoptic studio: A massively multiview system for social interaction capture. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Joo, H., Simon, T., and Sheikh, Y. (2018). Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8320–8329.

Joo, H., Neverova, N., and Vedaldi, A. (2020). Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In *International Conference on 3D Vision (3DV)*, pages 42–52.

Kanazawa, A., Black, M. J., Jacobs, D. W., and Malik, J. (2018). End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131.

Kar, A., Prakash, A., Liu, M.-Y., Cameracci, E., Yuan, J., Rusiniak, M., Acuna, D., Torralba, A., and Fidler, S. (2019). Meta-sim: Learning to generate synthetic datasets. In *International Conference on Computer Vision (ICCV)*, pages 4551–4560.

Karunratanakul, K., Preechakul, K., Suwajanakorn, S., and Tang, S. (2023). Guided motion diffusion for controllable human motion synthesis. In *International Conference on Computer Vision (ICCV)*, pages 2151–2162.

Keshavarzi, M., Parikh, A., Zhai, X., Mao, M., Caldas, L., and Yang, A. Y. (2020). SceneGen: Generative contextual scene augmentation using scene graph priors. *arXiv*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv*.

Kirillov, A., Wu, Y., He, K., and Girshick, R. (2020). PointRend: Image segmentation as rendering. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9799–9808.

Kocabas, M., Athanasiou, N., and Black, M. J. (2020). VIBE: Video inference for human body pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5253–5263.

Kocabas, M., Huang, C.-H. P., Hilliges, O., and Black, M. J. (2021). PARE: Part attention regressor for 3D human body estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11127–11137.

Kolotouros, N., Pavlakos, G., and Daniilidis, K. (2019a). Convolutional mesh regression for single-image human shape reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4496–4505.

Kolotouros, N., Pavlakos, G., Black, M. J., and Daniilidis, K. (2019b). Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision (ICCV)*, pages 2252–2261.

Kovar, L., Gleicher, M., and Pighin, F. (2023). Motion graphs. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 723–732.

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 *Computer Vision and Pattern Recognition (CVPR)*, pages 947–957.

Kuo, W., Angelova, A., Lin, T.-Y., and Dai, A. (2020). Mask2CAD: 3D shape prediction by learning to segment and retrieve. In *European Conference on Computer Vision (ECCV)*, volume 12348, pages 260–277.

Kwon, T., Tekin, B., Stuhmer, J., Bogo, F., and Pollefeys, M. (2021). H2o: Two hands manipulating objects for first person interaction recognition. In *International Conference on Computer Vision (ICCV)*, pages 10138–10148.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., and Navab, N. (2016). Deeper depth prediction with fully convolutional residual networks. In *International Conference on 3D Vision (3DV)*, pages 239–248. IEEE.

Lee, D. C., Hebert, M., and Kanade, T. (2009). Geometric reasoning for single image structure recovery. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2136–2143.

Lee, J. and Joo, H. (2023). Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments. In *International Conference on Computer Vision (ICCV)*, pages 9663–9674.

Lee, J., Chai, J., Reitsma, P. S. A., Hodgins, J. K., and Pollard, N. S. (2002). Interactive control of avatars animated with human motion data. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 491–500.

Lee, K. H., Choi, M. G., and Lee, J. (2006). Motion patches. *Transactions on Graphics (TOG)*, page 898–906.

Li, C., Chibane, J., He, Y., Pearl, N., Geiger, A., and Pons-Moll, G. (2024a). Unimotion: Unifying 3d human motion synthesis and understanding. *arXiv preprint arXiv:2409.15904*.

Li, J., Wu, J., and Liu, C. K. (2023). Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, **42**(6), 1–11.

Li, J., Clegg, A., Mottaghi, R., Wu, J., Puig, X., and Liu, C. K. (2024b). Controllable human-object interaction synthesis. In *European Conference on Computer Vision (ECCV)*.

Li, M., Patil, A. G., Xu, K., Chaudhuri, S., Khan, O., Shamir, A., Tu, C., Chen, B., Cohen-Or, D., and Zhang, H. (2019a). Grains: Generative recursive autoencoders for indoor scenes. *Transactions on Graphics (TOG)*, **38**(2), 1–16.

Li, X., Liu, S., Kim, K., Wang, X., Yang, M., and Kautz, J. (2019b). Putting humans in a scene: Learning affordance in 3D indoor environments. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12368–12376.

Liao, T., Yi, H., Xiu, Y., Tang, J., Huang, Y., Thies, J., and Black, M. J. (2024). Tada! text to animatable digital avatars. In *International Conference on 3D Vision (3DV)*, pages 1508–1519.

Liu, S., Zhang, Y., Peng, S., Shi, B., Pollefeys, M., and Cui, Z. (2020). Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2019–2028.

Liu, S., Jiang, H., Xu, J., Liu, S., and Wang, X. (2021). Semi-supervised 3d hand-object poses estimation with interactions in time. In *Computer Vision and Pattern Recognition (CVPR)*, pages 14687–14697.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. (2015). SMPL: A skinned multi-person linear model. *Transactions on Graphics (TOG)*, **34**(6), 248:1–248:16.

Lorensen, W. E. and Cline, H. E. (1987). Marching cubes: A high resolution 3D surface construction algorithm. *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 347–353.

Luo, A., Zhang, Z., Wu, J., and Tenenbaum, J. B. (2020). End-to-end optimization of scene layout. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3754–3763.

Luo, Z., Hachiuma, R., Yuan, Y., and Kitani, K. (2021). Dynamics-regulated kinematic policy for egocentric pose estimation. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, pages 25019–25032.

Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., and Black, M. J. (2019). AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, pages 5442–5451.

Mallya, A. and Lazebnik, S. (2015). Learning informative edge maps for indoor scene layout prediction. In *International Conference on Computer Vision (ICCV)*, pages 936–944.

Martinez, J., Hossain, R., Romero, J., and Little, J. J. (2017). A simple yet effective baseline for 3D human pose estimation. In *International Conference on Computer Vision (ICCV)*, pages 2659–2668.

Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., and Theobalt, C. (2018). Single-shot multi-person 3D pose estimation from monocular rgb. In *International Conference on 3D Vision (3DV)*, pages 120–130. IEEE.

Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., and Geiger, A. (2019). Occupancy networks: Learning 3D reconstruction in function space. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4460–4470.

Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding (CVIU)*, **104**(2), 90–126.

Monszpart, A., Guerrero, P., Ceylan, D., Yumer, E., and Mitra, N. J. (2019). iMapper: interaction-guided scene mapping from monocular videos. *Transactions on Graphics (TOG)*, **38**(4), 92:1–92:15.

Müller, L., Osman, A. A. A., Tang, S., Huang, C.-H. P., and Black, M. J. (2021). On self-contact and human pose. In *Computer Vision and Pattern Recognition (CVPR)*.

Müller, L., Ye, V., Pavlakos, G., Black, M. J., and Kanazawa, A. (2024). Generative proxemics: A prior for 3D social interaction from images. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9687–9697.

Müller, P., Wonka, P., Haegler, S., Ulmer, A., and Van Gool, L. (2006). Procedural modeling of buildings. In *Proceedings of the international conference on Computer graphics and interactive techniques (SIGGRAPH)*, pages 614–623.

Mura, C., Pajarola, R., Schindler, K., and Mitra, N. (2021). Walk2map: Extracting floor plans from indoor walk trajectories. *Computer Graphics Forum (CGF)*, **40**(2), 375–388.

Narasimhaswamy, S., Nguyen, T., and Nguyen, M. (2020). Detecting hands and recognizing physical contact in the wild. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 7841–7851.

Nie, Y., Han, X., Guo, S., Zheng, Y., Chang, J., and Zhang, J. J. (2020). Total3DUnderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 55–64.

Nie, Y., Dai, A., Han, X., and Nießner, M. (2022). Pose2room: understanding 3d scenes from human activities. In *European Conference on Computer Vision (ECCV)*, pages 425–443. Springer.

OpenAI (2024). Sora: Creating video from text. `https://openai.com/index/sora`. Accessed: 27 May 2024.

Para, W. R., Guerrero, P., Mitra, N., and Wonka, P. (2023). COFS: Controllable furniture layout synthesis. In *ACM SIGGRAPH Asia Conference Proceedings*, pages 1–11.

Parish, Y. I. and Müller, P. (2001). Procedural modeling of cities. In *Proceedings of the international conference on Computer graphics and interactive techniques (SIGGRAPH)*, pages 301–308.

Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S. (2019). DeepSDF: Learning continuous signed distance functions for shape representation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 165–174.

Paschalidou, D., Kar, A., Shugrina, M., Kreis, K., Geiger, A., and Fidler, S. (2021). ATISS: Autoregressive transformers for indoor scene synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, pages 12013–12026.

Patel, P., Huang, C.-H. P., Tesch, J., Hoffmann, D. T., Tripathi, S., and Black, M. J. (2021). AGORA: Avatars in geography optimized for regression analysis. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D., and Black, M. J. (2019a). Expressive body capture: 3d hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D., and Black, M. J. (2019b). https://github.com/vchoutas/smplx/tree/master/transfer_model.

Pavlakos, G., Kolotouros, N., and Daniilidis, K. (2019c). TexturePose: Supervising human mesh estimation with texture consistency. In *International Conference on Computer Vision (ICCV)*, pages 803–812.

Peng, X., Xie, Y., Wu, Z., Jampani, V., Sun, D., and Jiang, H. (2023). Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*.

Peng, X. B., Guo, Y., Halper, L., Levine, S., and Fidler, S. (2022). Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *Transactions on Graphics (TOG)*, **41**(4).

Petrovich, M., Black, M. J., and Varol, G. (2022). TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, pages 480–497.

Petrovich, M., Litany, O., Iqbal, U., Black, M. J., Varol, G., Peng, X. B., and Rempe, D. (2024). Multi-track timeline control for text-driven 3d human motion generation. In *CVPR Workshop on Human Motion Generation*, pages 1911–1921.

Pi, H., Peng, S., Yang, M., Zhou, X., and Bao, H. (2023). Hierarchical generation of human-object interactions with diffusion probabilistic models. In *International Conference on Computer Vision (ICCV)*, pages 15061–15073.

Prakash, A., Boochoon, S., Brophy, M., Acuna, D., Cameracci, E., State, G., Shapira, O., and Birchfield, S. (2019). Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In *International Conference on Robotics and Automation (ICRA)*, pages 7249–7255. IEEE.

Prokudin, S., Lassner, C., and Romero, J. (2019). Efficient learning on point clouds with basis point sets. In *International Conference on Computer Vision (ICCV)*, pages 4332–4341.

Puig, X., Shu, T., Li, S., Wang, Z., Liao, Y.-H., Tenenbaum, J. B., Fidler, S., and Torralba, A. (2021). Watch-and-help: A challenge for social perception and human-ai collaboration. In *International Conference on Learning Representations (ICLR)*.

Purkait, P., Zach, C., and Reid, I. (2020). Sg-vae: Scene grammar variational autoencoder to generate new indoor scenes. In *European Conference on Computer Vision (ECCV)*, pages 155–171. Springer.

Qi, S., Zhu, Y., Huang, S., Jiang, C., and Zhu, S.-C. (2018). Human-centric indoor scene synthesis using stochastic grammar. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5899–5908.

Ranftl, R., Bochkovskiy, A., and Koltun, V. (2021). Vision transformers for dense prediction. In *International Conference on Computer Vision (ICCV)*, pages 12179–12188.

Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., and Guibas, L. J. (2021). HuMoR: 3D human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, pages 11488–11499.

Rempe, D., Luo, Z., Peng, X. B., Yuan, Y., Kitani, K., Kreis, K., Fidler, S., and Litany, O. (2023). Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13756–13766.

Ritchie, D., Wang, K., and Lin, Y.-a. (2019). Fast and flexible indoor scene synthesis via deep convolutional generative models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6182–6190.

Rogez, G. and Schmid, C. (2016). MoCap-guided data augmentation for 3D pose estimation in the wild. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 3108–3116.

Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., and Li, H. (2019). PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision (ICCV)*, pages 2304–2314.

Saito, S., Simon, T., Saragih, J., and Joo, H. (2020). PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Computer Vision and Pattern Recognition (CVPR)*, pages 84–93.

Sarafianos, N., Boteanu, B., Ionescu, B., and Kakadiaris, I. A. (2016). 3D human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding (CVIU)*, **152**, 1–20.

Savva, M., Chang, A. X., Hanrahan, P., Fisher, M., and Nießner, M. (2016). PiGraphs: Learning Interaction Snapshots from Observations. *ACM Transactions on Graphics (TOG)*, **35**(4).

Shafir, Y., Tevet, G., Kapon, R., and Bermano, A. H. (2023). Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*.

Shin, D., Ren, Z., Sudderth, E. B., and Fowlkes, C. C. (2019). 3d scene reconstruction with multi-layer depth and epipolar transformers. In *International Conference on Computer Vision (ICCV)*, pages 2172–2182.

Sigal, L., Balan, A. O., and Black, M. J. (2010). Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV)*, **87**(1), 4–27.

Sitzmann, V., Zollhöfer, M., and Wetzstein, G. (2019). Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 32.

Smith, D., Loper, M., Hu, X., Mavroidis, P., and Romero, J. (2019). FACSIMILE: Fast and accurate scans from an image in less than a second. In *International Conference on Computer Vision (ICCV)*, pages 5329–5338.

Starke, S., Zhang, H., Komura, T., and Saito, J. (2019). Neural state machine for character-scene interactions. *Transactions on Graphics (TOG)*, **38**(6).

Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J. J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H. M., Nardi, R. D., Goesele, M., Lovegrove, S., and Newcombe, R. (2019). The Replica dataset: A digital replica of indoor spaces. *arXiv*.

Taheri, O., Ghorbani, N., Black, M. J., and Tzionas, D. (2020). GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, pages 581–600.

Taheri, O., Choutas, V., Black, M. J., and Tzionas, D. (2022). GOAL: Generating 4D whole-body motion for hand-object grasping. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13263–13273.

Talton, J. O., Lou, Y., Lesser, S., Duke, J., Měch, R., and Koltun, V. (2011). Metropolis procedural modeling. *Transactions on Graphics (TOG)*, **30**(2), 1–14.

Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V., and Fua, P. (2016). Structured prediction of 3D human pose with deep neural networks. In *British Machine Vision Conference (BMVC)*.

Tevet, G., Raab, S., Gordon, B., Shafir, Y., Bermano, A. H., and Cohen-Or, D. (2023). Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*.

Tian, L., Wang, Q., Zhang, B., and Bo, L. (2024). Emo: Emote portrait alive - generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision (ECCV)*.

Tome, D., Russell, C., and Agapito, L. (2017). Lifting from the deep: Convolutional 3D pose estimation from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5689–5698.

Tripathi, S., Chatterjee, A., Passy, J.-C., Yi, H., Tzionas, D., and Black, M. J. (2023). DECO: Dense estimation of 3D human-scene contact in the wild. In *International Conference on Computer Vision (ICCV)*, pages 8001–8013.

Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., and Schmid, C. (2018). BodyNet: Volumetric inference of 3D human body shapes. In *European Conference on Computer Vision (ECCV)*, pages 20–38.

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 *Conference on Neural Information Processing Systems (NeurIPS)*.

von Marcard, T., Henschel, R., Black, M. J., Rosenhahn, B., and Pons-Moll, G. (2018). Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, pages 614–631.

Wang, J., Yan, S., Dai, B., and Lin, D. (2021a). Scene-aware generative network for human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12206–12215.

Wang, J., Xu, H., Xu, J., Liu, S., and Wang, X. (2021b). Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9401–9411.

Wang, K., Savva, M., Chang, A. X., and Ritchie, D. (2018a). Deep convolutional priors for indoor scene synthesis. *Transactions on Graphics (TOG)*, **37**(4), 1–14.

Wang, K., Lin, Y.-A., Weissmann, B., Savva, M., Chang, A. X., and Ritchie, D. (2019a). Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *Transactions on Graphics (TOG)*, **38**(4), 1–15.

Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., and Jiang, Y.-G. (2018b). Pixel2mesh: Generating 3d mesh models from single rgb images. In *European Conference on Computer Vision (ECCV)*, pages 52–67.

Wang, X., Yeshwanth, C., and Nießner, M. (2021c). Sceneformer: Indoor scene generation with transformers. In *International Conference on 3D Vision (3DV)*, pages 106–115.

Wang, Z., Chen, L., Rathore, S., Shin, D., and Fowlkes, C. (2019b). Geometric pose affordance: 3D human pose with scene constraints. *arXiv*.

Wang, Z., Chen, Y., Liu, T., Zhu, Y., Liang, W., and Huang, S. (2022). Humanise: Language-conditioned human motion generation in 3d scenes. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Weinzaepfel, P., Brégier, R., Combaluzier, H., Leroy, V., and Rogez, G. (2020). DOPE: distillation of part experts for whole-body 3d pose estimation in the wild. In *European Conference on Computer Vision (ECCV)*, volume 12371, pages 380–397.

Weng, Z. and Yeung, S. (2020). Holistic 3D human and scene mesh estimation from single view images. In *Computer Vision and Pattern Recognition (CVPR)*, pages 334–343.

Xiang, D., Joo, H., and Sheikh, Y. (2019). Monocular total capture: Posing face, body, and hands in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10957–10966.

Xiao, Z., Wang, T., Wang, J., Cao, J., Dai, B., Lin, D., and Pang, J. (2024). Unified human-scene interaction via prompted chain-of-contacts. In *International Conference on Learning Representations (ICLR)*.

Xie, X., Lenssen, J. E., and Pons-Moll, G. (2025). Intertrack: Tracking human object interaction without object templates. In *International Conference on 3D Vision (3DV)*.

Xie, Y., Jampani, V., Zhong, L., Sun, D., and Jiang, H. (2024). Omnicontrol: Control any joint at any time for human motion generation. In *International Conference on Learning Representations (ICLR)*.

Xiu, Y., Yang, J., Tzionas, D., and Black, M. J. (2022). ICON: Implicit Clothed humans Obtained from Normals. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296.

Xu, H., Bazavan, E. G., Zanfir, A., Freeman, W. T., Sukthankar, R., and Sminchisescu, C. (2020). GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6183–6192.

Yang, L., Zhan, X., Li, K., Xu, W., Li, J., and Lu, C. (2021). Cpf: Learning a contact potential field to model the hand-object interaction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11097–11106.

Ye, S., Wang, Y., Li, J., Park, D., Liu, C. K., Xu, H., and Wu, J. (2022). Scene synthesis from human motion. In *ACM SIGGRAPH Asia Conference Proceedings*, SA '22.

Yi, H., Huang, C.-H. P., Tzionas, D., Kocabas, M., Hassan, M., Tang, S., Thies, J., and Black, M. J. (2022). Human-aware object placement for visual environment reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3959–3970.

Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., and Black, M. J. (2023a). Generating holistic 3D human motion from speech. In *Computer Vision and Pattern Recognition (CVPR)*, pages 469–480.

Yi, H., Huang, C.-H. P., Tripathi, S., Hering, L., Thies, J., and Black, M. J. (2023b). MIME: Human-aware 3D scene generation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12965–12976.

Yi, H., Thies, J., Black, M. J., Peng, X. B., and Rempe, D. (2024). Generating human interaction motions in scenes with text control. In *European Conference on Computer Vision (ECCV)*, pages 246–263.

Yu, Z., Yoon, J. S., Lee, I. K., Venkatesh, P., Park, J., Yu, J., and Park, H. S. (2020). HUMBI: A large multiview dataset of human body expressions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2990–3000.

Yuan, Y., Wei, S.-E., Simon, T., Kitani, K., and Saragih, J. (2021). Simpoe: Simulated character control for 3d human pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7159–7169.

Yuan, Y., Iqbal, U., Molchanov, P., Kitani, K., and Kautz, J. (2022). Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11038–11049.

Yuksel, S. E., Wilson, J. N., and Gader, P. D. (2012). Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, page 1177–1193.

Zanfir, A., Marinoiu, E., and Sminchisescu, C. (2018). Monocular 3D pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2148–2157.

Zhang, C., Cui, Z., Zhang, Y., Zeng, B., Pollefeys, M., and Liu, S. (2021a). Holistic 3D scene understanding from a single image with implicit representation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8833–8842.

Zhang, J. Y., Pepose, S., Joo, H., Ramanan, D., Malik, J., and Kanazawa, A. (2020a). Perceiving 3D human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, volume 12357, pages 34–51.

Zhang, L., Rao, A., and Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision (ICCV)*, pages 3836–3847.

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 preprint arXiv:2208.15001*.

Zhang, S., Zhang, Y., Ma, Q., Black, M. J., and Tang, S. (2020b). PLACE: Proximity learning of articulation and contact in 3D environments. In *International Conference on 3D Vision (3DV)*.

Zhang, S., Zhang, Y., Bogo, F., Marc, P., and Tang, S. (2021b). Learning motion priors for 4d human body capture in 3d scenes. In *International Conference on Computer Vision (ICCV)*.

Zhang, S.-H., Zhang, S.-K., Xie, W.-Y., Luo, C.-Y., and Fu, H.-B. (2020c). Fast 3d indoor scene synthesis with discrete and exact layout pattern extraction. *arXiv*.

Zhang, W., Dabral, R., Leimkühler, T., Golyanik, V., Habermann, M., and Theobalt, C. (2024a). Roam: Robust and object-aware motion generation using neural pose descriptors. *International Conference on 3D Vision (3DV)*, pages 1392–1402.

Zhang, X., Bhatnagar, B. L., Starke, S., Guzov, V., and Pons-Moll, G. (2022b). Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision (ECCV)*, pages 518–535. Springer.

Zhang, X., Bhatnagar, B. L., Starke, S., Petrov, I., Guzov, V., Dhamo, H., Pérez Pellitero, E., and Pons-Moll, G. (2024b). Force: Dataset and method for intuitive physics guided human-object interaction. *Arxiv*.

Zhang, Y. and Tang, S. (2022). The wanderings of odysseus in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20481–20491.

Zhang, Y., Song, S., Yumer, E., Savva, M., Lee, J.-Y., Jin, H., and Funkhouser, T. (2017). Physically-based rendering for indoor scene understanding using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5287–5295.

Zhang, Y., Hassan, M., Neumann, H., Black, M. J., and Tang, S. (2020d). Generating 3D people in scenes without people. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6194–6204.

Zhang, Y., Zhang, H., Hu, L., Yi, H., Zhang, S., and Liu, Y. (2024c). Real-time monocular full-body capture in world space via sequential proxy-to-motion learning. In *Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Z., Yang, Z., Ma, C., Luo, L., Huth, A., Vouga, E., and Huang, Q. (2020e). Deep generative modeling for scene synthesis via hybrid representations. *Transactions on Graphics (TOG)*, **39**(2), 1–21.

Zhao, K., Wang, S., Zhang, Y., Beeler, T., , and Tang, S. (2022). Compositional human-scene interaction synthesis with semantic control. In *European Conference on Computer Vision (ECCV)*.

Zhao, K., Zhang, Y., Wang, S., Beeler, T., , and Tang, S. (2023). Synthesizing diverse human motions in 3d indoor scenes. In *International Conference on Computer Vision (ICCV)*, pages 14738–14749.

Zhao, Y. and Zhu, S.-C. (2013). Scene parsing by integrating function, geometry and appearance models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3119–3126.

Zheng, Z., Yu, T., Wei, Y., Dai, Q., and Liu, Y. (2019). DeepHuman: 3D human reconstruction from a single image. In *International Conference on Computer Vision (ICCV)*, pages 7738–7748.

Zheng, Z., Zhao, X., Zhang, H., Liu, B., and Liu, Y. (2023). Avatarrex: Real-time expressive full-body avatars. *ACM Transactions on Graphics (TOG)*, **42**(4), 1–19.

Zhou, Y., While, Z., and Kalogerakis, E. (2019). Scenegraphnet: Neural message passing for 3d indoor scene augmentation. In *International Conference on Computer Vision (ICCV)*, pages 7384–7392.

Zhu, S., Chen, J. L., Dai, Z., Xu, Y., Cao, X., Yao, Y., Zhu, H., and Zhu, S. (2024). Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision (ECCV)*.

Zhu, W., Ma, X., Ro, D., Ci, H., Zhang, J., Shi, J., Gao, F., Tian, Q., and Wang, Y. (2023). Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zollhöfer, M., Stotko, P., Görlitz, A., Theobalt, C., Nießner, M., Klein, R., and Kolb, A. (2018). State of the art on 3D reconstruction with RGB-D cameras. *Computer Graphics Forum (CGF)*, **37**(2), 625–652.

Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., Li, B., Zhang, C., Zhang, C., Wei, Y., *et al.* (2021). End-to-end human object interaction detection with hoi transformer. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11825–11834.