Towards Realistic Scene-Aware Human-Object Interaction in 3D World

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering
by Research

by

Kunal Kamalkishor Bhosikar
2022121005
kunal.bhosikar@research.iiit.ac.in

Advisor: Dr. Charu Sharma

International Institute of Information Technology Hyderabad
500 032, India

May 2026

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled ***Towards Realistic Scene-Aware Human-Object Interaction in 3D World*** by *Kunal Kamalkishor Bhosikar*, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Prof. Charu Sharma

To my parents, whose sacrifices made this journey possible.

Acknowledgments

This thesis is the result of the guidance, collaboration, and support of many individuals. I would like to express my sincere gratitude to my advisor, *Dr. Charu Sharma*, for her constant guidance, support, and encouragement throughout my research journey. Her insights, patience, and emphasis on clear thinking have played a crucial role in shaping this work and my approach to research.

I am grateful to the members of the *Machine Learning Lab (MLL)* for fostering a collaborative and intellectually stimulating environment. I would like to especially thank my seniors, *Siddharth* and *Vivek*, with whom I worked on my first research project, for their mentorship and guidance. I also thank *Jaya, Himanshu,* and *Janak* for their support and valuable discussions.

A special thanks to *Sreehari*, whose collaboration and ideas have been invaluable, and with whom I have had the opportunity to work closely on impactful research. I am also thankful to my batchmates and collaborators, *Pranav, Vanshita,* and *Mahika*, for their constant support and teamwork. I appreciate the camaraderie and support from my labmates and friends, including *Sanjay* and *Akash*, as well as my juniors *Vinit, Monish, Kushang, Aditya,* and *Prajas,* for their enthusiasm and collaborative spirit.

I would also like to acknowledge my external collaborators, *Dr. Kai Han, Dr. Nikos Athanasiou, Vanessa Sklyarova, Dr. Lokender Tiwari,* and *Preet Savalia,* for their valuable insights and collaboration, which significantly enriched my research experience.

I am grateful for the opportunities to present my work at international venues such as *BMVC* and *WACV*. I sincerely thank my advisor, my institute, and my parents for supporting these experiences.

I would like to thank my friends–*Ishan, Aaryan, Faizal, Hiya, Aryaman, Prakhar, Aditya, Yashpal, Radhey,* and *Pranav*–for their encouragement, support, and for making this journey enjoyable.

Finally, I am deeply grateful to my parents for their unwavering support, sacrifices, and belief in me, which made this journey possible. I would also like to thank my younger sister, *Shruti*, for her constant encouragement and support.

Abstract

Modeling realistic human-object interaction in three-dimensional environments is a fundamental challenge in computer vision, with broad implications for robotics, virtual reality, and embodied artificial intelligence. Despite significant progress in human motion generation and 3D scene understanding, existing approaches often treat motion, interaction, and environment as separate problems, limiting their ability to generate physically plausible and context-aware behavior.

This thesis addresses this challenge through a unified perspective on *scene-aware human-object interaction*, grounded in both real-world systems and data-driven modeling. We begin by developing a system for real-time human pose understanding and interaction analysis based on joint-level reasoning, formalized in our published patent application. This system demonstrates how structured representations of human motion can enable robust, interpretable, and deployable interaction-aware applications.

Building on this foundation, we introduce a large-scale dataset and benchmark for full-body human motion with object interaction in realistic 3D scenes. This dataset enables the study of scene-aware grasping and exposes key limitations of existing methods in handling interaction under environmental constraints. We then propose a learning-based framework for generating physically plausible, scene-aware human-object interactions that explicitly models the interplay between body motion, object manipulation, and scene geometry.

To enable scalable deployment of such systems, we further develop a fast and robust geometry processing framework based on feature-aware mesh simplification, which preserves interaction-relevant geometric properties while significantly reducing computational complexity.

We also explore the extension of interaction-aware modeling to conditional settings, including multimodal human motion generation. In particular, we highlight the role of object interaction in ensuring consistency between generated motion and environmental constraints, demonstrating that naive integration of motion and interaction leads to physically implausible results.

Through extensive experiments and evaluations, we demonstrate that the proposed approaches significantly improve the realism, physical plausibility, and scalability of human-object interaction in 3D environments. By bridging real-world systems, data-driven modeling, and geometry processing, this thesis contributes towards the development of integrated frameworks for human-centric 3D understanding and interaction.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Understanding how humans interact with objects in three-dimensional (3D) environments lies at the core of intelligent behavior. From grasping everyday objects to navigating cluttered spaces, human interaction seamlessly integrates perception, motion, and physical reasoning. Replicating such capabilities in computational systems remains a fundamental challenge in computer vision, with far-reaching implications for robotics, virtual and augmented reality, animation, and embodied artificial intelligence [46].

Despite rapid advances in 3D scene reconstruction [35, 26], human motion generation [54], and object interaction modeling [18], existing approaches largely treat these components in isolation. Motion generation methods often produce visually plausible body movements without explicitly modeling interaction with objects or the surrounding environment. Conversely, techniques for hand-object interaction or grasp synthesis typically operate in constrained settings, ignoring full-body dynamics and scene context. Similarly, scene-aware approaches often capture coarse environmental constraints but fail to model fine-grained interactions such as grasping and contact.

This fragmentation leads to unrealistic behaviors, including interpenetration between the human body and the scene, physically implausible grasps, and a lack of coordination between motion and interaction. As shown in Figure 1.1, existing approaches fail to jointly model these components, resulting in unstable and physically inconsistent interactions. These failures are not incidental, but symptomatic of a deeper limitation: the lack of a unified framework for interaction-aware modeling.

Figure 1.1 Comparison of human-object interaction generated by existing approaches and our method (blue). Prior methods fail to jointly model motion, interaction, and scene constraints, leading to artifacts such as body-scene interpenetration and unstable grasps. In contrast, our approach produces physically plausible interactions by enforcing contact consistency and respecting scene geometry. This highlights the importance of unified modeling for realistic interaction.

A key challenge lies in the inherently coupled nature of human-object interaction. Realistic interaction requires simultaneously modeling (i) full-body motion, (ii) fine-grained object manipulation, and (iii) the constraints imposed by complex 3D scenes. Errors in any one of these components can propagate and degrade the overall realism of the interaction.

At its core, this thesis is guided by a simple premise: *realistic human-object interaction emerges not from isolated components, but from their principled integration.*

This thesis addresses this challenge through a unified perspective that spans real-world systems, data-driven modeling, and efficient geometric processing.

We begin by grounding this problem in real-world systems through a framework for real-time human pose understanding and interaction analysis. By leveraging joint-level reasoning and geometric

constraints, this system enables robust interpretation of human motion in practical settings, forming a foundation for interaction-aware modeling beyond simulated environments.

Building on this foundation, we address a fundamental bottleneck in existing research: the lack of datasets that jointly capture full-body motion, fine-grained object interaction, and realistic scene context. We introduce a dataset that combines pre-grasp motion with detailed grasping poses in richly annotated 3D scenes, enabling the study of scene-aware interaction and exposing the limitations of existing methods in handling scene constraints.

We then propose a learning-based framework for generating scene-aware grasping motions that explicitly models interactions between the human body, objects, and the surrounding environment. This framework reduces interpenetration, improves contact consistency, and enables physically plausible motion generation.

To support scalable deployment, we further address a critical systems-level challenge: the efficient representation and processing of complex 3D geometry. High-fidelity 3D assets are often dense and computationally expensive, limiting their usability in real-time applications. To overcome this, we propose a feature-aware mesh simplification framework based on quadric error metrics [12], enabling scalable integration of interaction-aware models.

We further explore extensions to conditional motion generation, demonstrating how interaction-aware modeling can be incorporated into generative frameworks guided by high-level inputs. While multimodal signals such as speech and language provide additional context, our focus remains on ensuring that generated motion maintains consistency with object interaction and scene constraints.

Through this combination of real-world systems, dataset creation, model development, and geometry processing, this thesis advances the state of the art in scene-aware human-object interaction.

Together, these contributions demonstrate that realistic human-object interaction is not the result of isolated advances, but of a unified approach that integrates perception, motion, interaction, and geometry across both simulated and real-world settings.

1.1 Motivation and Challenges

Humans interact with the physical world in a seamless and intuitive manner, effortlessly combining full-body motion, object manipulation, and environmental awareness. Consider a simple everyday task such as picking up a cup from a table. This action involves coordinated full-body motion, precise hand-object interaction, and continuous awareness of the surrounding environment.

Figure 1.2 illustrates an example of such scene-aware human-object interaction. Crucially, these components are tightly coupled—failures in contact, motion, or scene awareness can lead to unrealistic behavior.

Figure 1.2 Example of scene-aware human-object interaction from the proposed MOGRAS dataset and model. The generated motion demonstrates coordinated full-body movement, stable hand-object contact, and consistency with surrounding scene constraints. Unlike prior approaches that treat these components independently, our method jointly models motion, interaction, and environment, resulting in physically grounded and realistic behavior.

Despite its importance, achieving realistic human-object interaction remains an open problem. Existing approaches struggle to balance motion realism, interaction fidelity, and scene awareness, particularly in complex environments.

1.2 Problem Statement

We formalize this challenge as the following central problem:

> ***How can we model and generate realistic, scene-aware human-object interactions in 3D environments that jointly account for full-body motion, fine-grained object manipulation, and environmental constraints?***

Addressing this problem requires overcoming challenges in data, modeling, and efficiency.

1.3 Research Landscape

Understanding human-object interaction in 3D environments requires the integration of multiple research domains, including human motion generation, hand-object interaction, scene-aware modeling, multimodal learning, and geometry processing. While each of these areas has seen significant advances in recent years, their development has largely been independent, resulting in fragmented solutions that fail to generalize to realistic interaction scenarios.

In this section, we review these domains and highlight the key limitations that motivate the unified approach proposed in this thesis.

1.3.1 Human Motion Generation

Human motion generation has witnessed rapid progress with the advent of data-driven and generative models. Early approaches relied on motion capture datasets and kinematic modeling, while recent methods leverage deep generative frameworks such as variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion models [54].

These methods are capable of generating diverse and realistic motion sequences conditioned on inputs such as action labels, trajectories, or textual descriptions. However, most motion generation approaches focus primarily on visual realism and temporal coherence, without explicitly modeling interaction with objects or the surrounding environment.

As a result, when these motions are placed in realistic 3D scenes, they often exhibit physically implausible behavior, such as interpenetration with objects or inconsistent contact. This highlights a key limitation: motion generation alone is insufficient for modeling realistic interaction.

1.3.2 Hand-Object Interaction

A parallel line of research focuses on hand-object interaction and grasp synthesis. These methods aim to generate physically plausible hand poses that achieve stable contact with objects [18]. Techniques in this domain often leverage optimization-based formulations or learning-based approaches to model fine-grained contact between the hand and object.

While these methods achieve high accuracy in modeling local interactions, they are typically limited to isolated hand-object configurations. They do not account for full-body motion, balance, or the influence of the surrounding scene.

In real-world scenarios, grasping is not an isolated event but part of a continuous motion sequence involving the entire body. The lack of full-body context limits the applicability of these methods in realistic environments.

1.3.3 Scene-Aware Interaction

Recent work has attempted to incorporate scene context into human motion modeling, leading to scene-aware interaction frameworks [16, 63]. These methods enforce constraints such as collision avoidance, support relationships, and contact consistency, enabling more realistic interaction with the environment.

However, existing approaches often operate at a coarse level, focusing on global constraints rather than fine-grained object interaction. Additionally, many methods rely on optimization-based techniques that can be computationally expensive and difficult to scale.

Furthermore, the lack of datasets that jointly capture full-body motion, object interaction, and scene context limits the ability of learning-based approaches to model such interactions effectively.

1.3.4 Multimodal Interaction

Human behavior is inherently multimodal, driven by high-level intent expressed through language, speech, and context. Recent work has explored generating motion conditioned on such inputs, enabling more expressive and controllable motion synthesis.

While these approaches demonstrate promising results in aligning motion with semantic cues, they often treat object interaction as a secondary component. As a result, generated motions may be semantically meaningful but physically inconsistent when interacting with objects.

This highlights the need for frameworks that integrate multimodal conditioning with interaction-aware modeling, ensuring consistency between high-level intent and low-level physical constraints.

1.3.5 3D Geometry Processing

Efficient representation and processing of 3D geometry is essential for scalable interaction-aware systems. High-resolution meshes are commonly used to represent detailed scenes, but they introduce significant computational overhead in terms of memory, rendering, and collision detection.

Mesh simplification techniques, such as quadric error metrics (QEM) [12], provide a practical solution by reducing geometric complexity while preserving visual fidelity. However, standard simplification methods primarily focus on geometric error and do not account for features that are critical for interaction, such as boundaries, sharp edges, and contact regions.

In the context of human-object interaction, preserving these features is essential for maintaining physical plausibility and accurate contact modeling. This motivates the need for feature-aware geometry processing methods that are tailored to interaction-aware applications.

1.3.6 Summary and Gap

In summary, existing research has made significant progress in individual components of human-object interaction, including motion generation, grasp synthesis, scene-aware modeling, multimodal

learning, and geometry processing. However, these components are largely developed in isolation, resulting in systems that lack coherence and physical realism in complex environments.

In particular, there is a lack of unified approaches that jointly model full-body motion, fine-grained object interaction, scene constraints, and efficient geometry representation, while also enabling deployment in real-world systems.

This thesis addresses this gap by proposing a unified framework that integrates real-world human understanding, interaction-aware motion modeling, and efficient geometry processing, enabling realistic and scalable human-object interaction in 3D environments.

1.4 Thesis Contributions

- **Real-World Interaction System (Patent):** A system for real-time human pose understanding and interaction analysis based on joint-level reasoning, demonstrating practical deployment of interaction-aware modeling.
- **Scene-Aware Dataset (MOGRAS):** A dataset capturing full-body motion and object interaction in 3D scenes.
- **Scene-Aware Motion Generation (GNet++):** A framework for physically plausible interaction.
- **Efficient Geometry Processing (FA-QEM):** Feature-aware mesh simplification.

1.5 Thesis Roadmap

Chapter 2 reviews related work. Chapter 3 presents the real-world interaction system and its methodology. Chapter 4 introduces scene-aware human-object interaction modeling. Chapter 5 presents efficient geometry processing. Chapter 6 explores extensions to conditional interaction modeling. Chapter 7 concludes the thesis.

Chapter 2

Background and Related Work

Understanding human-object interaction in 3D environments requires integrating multiple research domains, including scene representation, human motion modeling, interaction modeling, multimodal learning, and geometry processing. While each of these research directions has independently achieved substantial progress, their integration remains limited. Existing methods often focus on isolated components of interaction, such as motion generation, grasp synthesis, or scene understanding, resulting in systems that struggle to produce physically plausible and context-aware behavior in complex environments.

This chapter reviews key developments across these domains and highlights the limitations that motivate the unified interaction-aware framework proposed in this thesis.

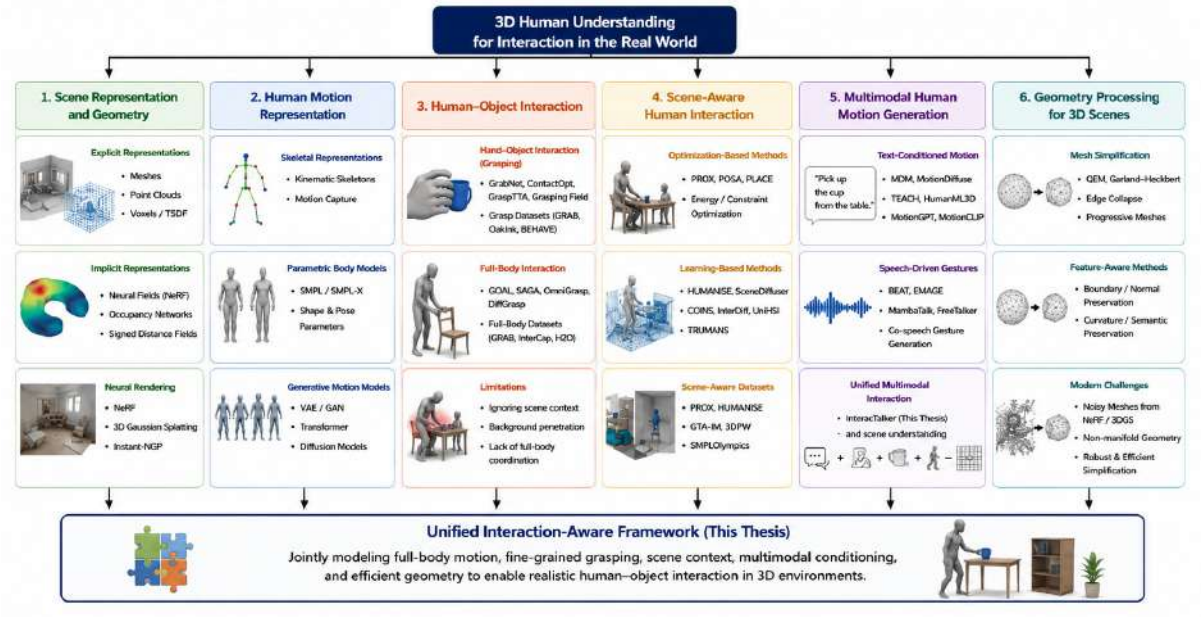

Figure 2.1 Taxonomy of Research Areas Related to Interaction-Aware Human Motion

2.1 3D Scene Representation and Geometry

Accurate representation of three-dimensional (3D) environments forms the foundation of interaction-aware modeling. Traditional scene representations such as polygonal meshes, voxels, and point clouds provide explicit geometric structure, enabling rendering, collision detection, and physical reasoning [25]. Among these, polygonal meshes remain one of the most widely used representations due to their explicit surface connectivity and compatibility with graphics pipelines.

Point cloud representations provide lightweight geometric descriptions and are widely used in large-scale scene capture and reconstruction systems. However, point clouds lack explicit connectivity information, making surface reasoning and interaction modeling challenging. Volumetric and voxel-based representations address this limitation by discretizing space into occupancy grids, enabling spatial reasoning and collision analysis. Nevertheless, voxel representations scale poorly with increasing resolution and scene complexity.

Recent advances in neural scene representations have significantly improved the fidelity and realism of reconstructed environments. Neural Radiance Fields (NeRF) [35] introduced continuous volumetric scene representations using neural networks, enabling photorealistic novel view synthesis. Subsequent methods such as Instant-NGP [37] improved rendering efficiency using multiresolution hash encodings, while 3D Gaussian Splatting [26] further enabled real-time rendering of complex scenes with high visual quality.

In parallel, implicit geometric representations such as Occupancy Networks [34], DeepSDF [38], and signed distance field (SDF)-based approaches have enabled continuous surface modeling and high-quality geometry reconstruction. These methods provide smooth and expressive representations of geometry while avoiding discretization artifacts common in voxel-based approaches.

Despite these advances, representing 3D scenes for interaction remains challenging. Neural scene representations are primarily optimized for rendering quality rather than physical reasoning, making tasks such as contact modeling, collision detection, and interaction simulation computationally expensive. Conversely, explicit representations such as meshes are well-suited for interaction reasoning but often struggle to scale efficiently to large and complex environments.

This trade-off between geometric fidelity and computational efficiency is particularly critical for interaction-aware systems, where accurate contact reasoning and real-time processing are both required. These limitations motivate the need for efficient geometry processing methods that preserve interaction-relevant features while enabling scalable computation.

2.2 Human Motion Representation

Human motion modeling is a central problem in computer vision, graphics, robotics, and virtual reality. Human motion is commonly represented using skeletal structures, parametric body models, or learned latent representations.

Early approaches relied on skeletal representations, where the human body is modeled as a hierarchy of joints connected through kinematic chains. Such representations are widely used in motion capture systems and animation pipelines due to their simplicity and efficiency. However, skeletal representations lack surface-level detail and are insufficient for modeling fine-grained interaction and contact.

To address these limitations, parametric human body models such as SMPL [30] and SMPL-X [39] were introduced. These models represent the human body as a deformable mesh controlled through pose and shape parameters, enabling realistic surface deformation and full-body interaction modeling. Large-scale motion datasets such as AMASS [33] further facilitated data-driven learning of realistic human motion using unified body representations.

Recent years have witnessed rapid progress in generative human motion modeling. Early learning-based approaches leveraged recurrent neural networks and variational autoencoders (VAEs) to synthesize temporally coherent motion sequences [21, 19]. Generative adversarial networks (GANs) were subsequently explored for producing diverse and realistic motion distributions [3].

More recently, transformer-based and diffusion-based methods have become dominant due to their ability to model long-range temporal dependencies and complex motion distributions. ACTOR [40] and TEMOS [41] demonstrated the effectiveness of transformer architectures for action-conditioned motion generation. Motion Diffusion Models (MDM) [54] introduced diffusion-based motion synthesis conditioned on textual descriptions, achieving significant improvements in motion realism and semantic controllability. Subsequent methods such as MotionDiffuse [65], MotionGPT [22], and MotionCLIP [53] further improved multimodal alignment and motion diversity.

In parallel, text-to-motion datasets such as HumanML3D [15] enabled large-scale learning of semantically grounded motion representations. Scene-aware motion generation methods such as TeSMo [64], SceneDiffuser [20], and InterDiff [59] further explored integrating environmental context into motion synthesis.

Despite these advances, most motion generation methods primarily optimize for visual realism and temporal smoothness without explicitly modeling physical interaction or environmental constraints. Consequently, generated motions often exhibit artifacts such as object penetration, unstable contacts, and physically implausible body configurations when deployed in realistic environments. These limitations become particularly severe in interaction-heavy scenarios involving fine-grained object manipulation and cluttered scenes.

These challenges motivate the interaction-aware motion modeling framework proposed in this thesis.

2.3 Human-Object Interaction

Human-object interaction (HOI) focuses on modeling the physical relationship between humans and objects, including grasping, manipulation, contact reasoning, and interaction dynamics. HOI plays a fundamental role in robotics, virtual humans, embodied AI, and augmented reality applications.

A significant body of research has focused on hand-object interaction and grasp synthesis. Early methods relied on optimization-based formulations for generating stable grasp configurations [36]. More recent learning-based approaches such as GrabNet [51], Grasping Field [18], ContactOpt [14], and GraspTTA [23] leverage neural networks and contact-aware optimization to generate realistic and physically plausible hand poses.

Large-scale datasets such as GRAB [51], OakInk [61], and BEHAVE [4] enabled substantial progress in data-driven grasp synthesis and interaction modeling. These datasets provide detailed hand-object contact annotations and realistic interaction sequences, enabling models to learn complex contact relationships.

Recent work has extended grasp synthesis from isolated hand interactions to full-body human-object interaction. GOAL [50], SAGA [57], OmniGrasp [32], and DiffGrasp [67] demonstrated the importance of modeling coordinated full-body motion for realistic grasping behavior. These approaches leverage parametric body representations and generative models to synthesize plausible full-body interactions.

However, a major limitation of existing HOI methods is their lack of scene awareness. Most approaches operate in isolated object-centric settings without considering surrounding environmental geometry. While such methods achieve realistic local contact, they often fail when deployed in cluttered scenes, producing severe interpenetration between the human body and the environment.

In realistic settings, grasping is not an isolated event but rather a component of a continuous interaction process involving locomotion, balance, obstacle avoidance, and scene understanding. The lack of joint modeling of full-body motion, object interaction, and environmental constraints remains a major challenge in realistic HOI generation.

This limitation motivates the need for datasets and models that jointly capture full-body motion, fine-grained grasping, and realistic scene context.

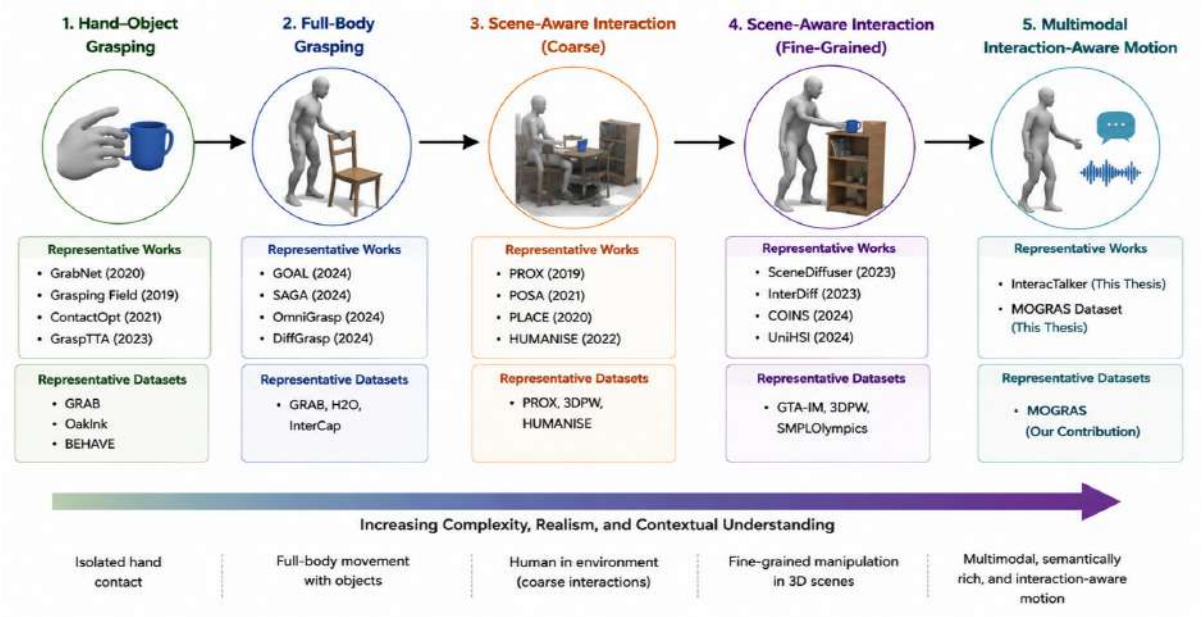

Figure 2.2 Evolution of Research in Human-Object Interaction

2.4 Scene-Aware Human Interaction

Scene-aware human interaction focuses on generating physically plausible human behavior that is consistent with surrounding environmental geometry. This research direction integrates human motion generation with scene understanding, collision avoidance, support constraints, and contact reasoning.

Early approaches relied on optimization-based formulations to place humans realistically within scenes [45, 16]. PROX [16] demonstrated the importance of incorporating scene geometry into human pose estimation and interaction modeling. Subsequent approaches such as POSA [17] and PLACE [66] leveraged learned priors to synthesize semantically meaningful human-scene interactions.

Learning-based scene-aware methods significantly improved scalability and realism. HUMANISE [56], SceneDiffuser [20], InterDiff [59], UniHSI [58], and TRUMANS [24] explored generating scene-consistent human motion using diffusion models, transformer architectures, and interaction priors.

Large-scale scene-aware datasets have also played a critical role in advancing this area. Datasets such as PROX [16], HUMANISE [56], GTA-IM [6], and SMPLOlympics [55] provide diverse examples of human motion in realistic environments.

Despite these advances, existing scene-aware interaction methods primarily focus on coarse interactions such as sitting, walking, navigation, or support relationships. Fine-grained object manipulation and detailed grasping interactions remain comparatively underexplored.

Additionally, optimization-based approaches are often computationally expensive and difficult to scale to long motion sequences or large environments. Learning-based methods, while efficient, remain

limited by the availability of datasets containing detailed contact annotations and realistic interaction behavior.

Consequently, achieving both fine-grained manipulation and full-body scene awareness remains an open challenge. These limitations motivate the need for unified frameworks that jointly model full-body motion, grasping, and environmental constraints.

2.5 Multimodal Human Motion Generation

Human behavior is inherently multimodal, influenced by language, speech, visual context, and environmental interactions. Recent work has explored generating human motion conditioned on such high-level signals, enabling more expressive and controllable motion synthesis.

Text-conditioned motion generation has emerged as a major research direction with the availability of large-scale text-motion datasets such as HumanML3D [15]. Methods such as MDM [54], MotionDiffuse [65], and TEACH [1] demonstrated the ability to synthesize semantically meaningful motion conditioned on textual prompts.

In parallel, speech-driven gesture generation has received significant attention. Early methods relied on recurrent neural networks and audio features to generate upper-body gestures synchronized with speech [13]. Large-scale datasets such as BEAT [28] enabled learning expressive co-speech gestures with realistic timing and style.

Recent methods including EMAGE [27], MambaTalk [60], and FreeTalker [62] leveraged transformer and diffusion architectures to generate highly expressive gesture sequences conditioned on speech, text, and semantic information.

While these approaches generate realistic multimodal motion, most existing methods treat object interaction as a secondary component or ignore it entirely. Consequently, generated motions may align semantically with language or speech while remaining physically inconsistent with the surrounding environment.

A straightforward combination of independently generated co-speech gestures and object interaction motion often leads to unrealistic behavior, including self-penetration, object penetration, and poor coordination between upper-body gestures and lower-body locomotion. This highlights the importance of jointly modeling multimodal conditioning and interaction-aware motion generation.

This challenge is further explored in Chapter 6 through interaction-aware multimodal motion generation.

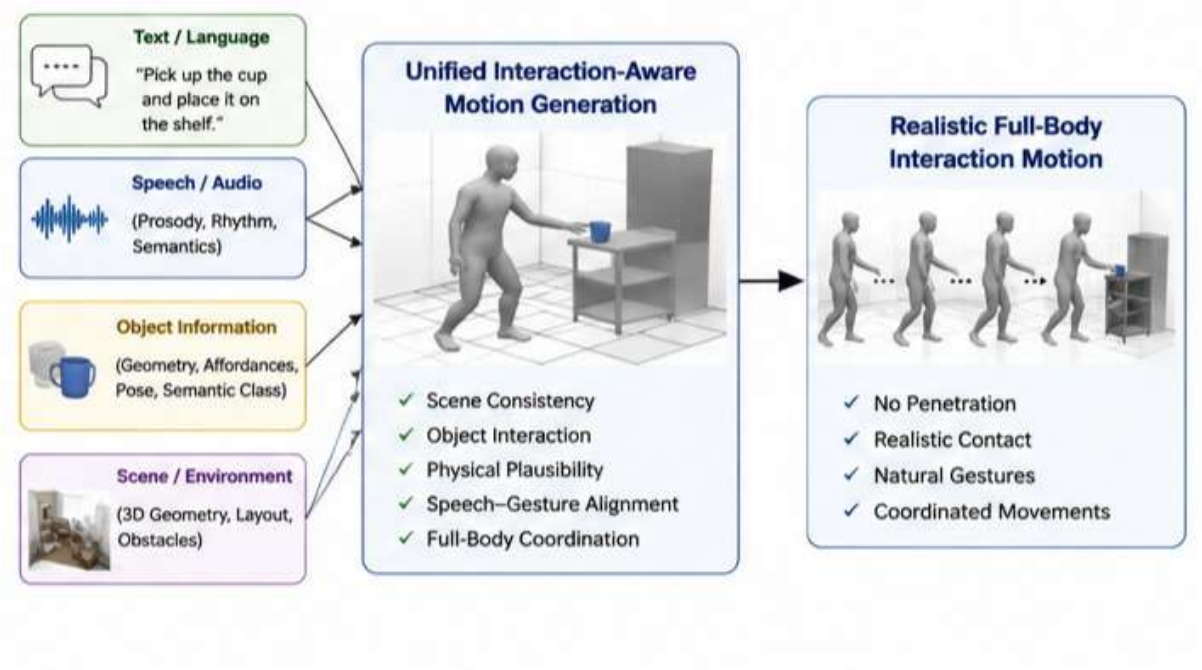

Figure 2.3 Concept of Multimodal Interaction-Aware Motion Generation

2.6 Geometry Processing for 3D Scenes

Efficient geometry processing is essential for scalable interaction-aware systems. High-resolution meshes generated from modern reconstruction and generative pipelines are often computationally expensive, motivating the need for efficient mesh simplification and geometry optimization methods.

Mesh simplification has long been studied in computer graphics. Early approaches explored vertex clustering, remeshing, and decimation techniques [47]. Among these, Quadric Error Metrics (QEM) [12] remains one of the most influential and widely used simplification frameworks due to its balance between efficiency and geometric fidelity.

Subsequent work extended QEM to preserve additional geometric properties such as boundaries, normals, textures, curvature, and semantic features [12, 2]. Feature-preserving simplification methods demonstrated improved retention of sharp structures and interaction-relevant regions under aggressive simplification.

Recent advances in neural rendering and generative 3D modeling, including NeRF and 3D Gaussian Splatting, have resulted in dense and often noisy meshes extracted from implicit scene representations. These meshes frequently contain non-manifold structures, irregular topology, and high-frequency geometric detail, posing significant challenges for traditional simplification methods.

Methods such as Liu et al. [29] explored robust simplification pipelines for real-world meshes. However, many existing methods either sacrifice geometric fidelity for robustness or become computationally expensive under large-scale deployment.

In the context of human-object interaction, preserving interaction-relevant geometric features such as contact surfaces, boundaries, and high-curvature regions is particularly important for accurate collision detection and physical reasoning. These limitations motivate the feature-aware geometry processing framework proposed in this thesis.

2.7 Summary and Research Gap

In summary, existing research has achieved substantial progress in individual components of human-object interaction, including scene representation, motion generation, grasp synthesis, multimodal conditioning, and geometry processing. However, these components are largely developed in isolation, resulting in fragmented systems that lack coherence and physical realism in complex environments.

Human motion generation methods often prioritize perceptual realism without modeling detailed interaction constraints. Human-object interaction approaches achieve realistic local contact but typically ignore surrounding scene geometry. Scene-aware methods focus primarily on coarse interactions while lacking fine-grained manipulation capability. Similarly, multimodal motion generation methods struggle to maintain consistency between semantic conditioning and physical interaction.

In addition, modern interaction-aware systems increasingly rely on large-scale 3D environments and dense geometric assets, introducing substantial computational challenges for real-time deployment and physical simulation.

These limitations reveal a critical gap: the lack of unified frameworks that jointly model full-body motion, fine-grained object interaction, scene constraints, multimodal conditioning, and efficient geometry representation.

This thesis addresses this gap through a unified interaction-aware framework that integrates real-world human understanding, scene-aware motion generation, grasp synthesis, multimodal interaction modeling, and feature-aware geometry processing, enabling realistic and scalable human-object interaction in 3D environments.

Chapter 3

Real-World Interaction System for Human Pose Understanding

While the previous chapter highlighted the challenges in modeling realistic human-object interaction, most existing approaches operate in simulated or offline settings. In contrast, real-world applications require systems that can interpret human motion in real time, reason about interactions, and provide actionable feedback.

In this chapter, we present a practical system for real-time human pose understanding and interaction analysis, developed as part of a system described in a published patent application. The system bridges the gap between theoretical modeling and real-world deployment by enabling robust interpretation of human motion through joint-level reasoning and geometric constraints.

Unlike learning-based approaches that rely heavily on large datasets and complex models, the proposed system adopts a structured pipeline that combines keypoint detection, geometric reasoning, and rule-based interpretation. This design enables efficient and interpretable analysis of human motion, making it suitable for real-time applications such as exercise monitoring, rehabilitation, and human-computer interaction.

The system presented in this chapter forms the basis of a published patent application titled "System and Method for Detecting Exercise Poses and Counting Repetitions Based on Joint Angles" [49].

3.1 Problem Definition

The goal of the system is to analyze human motion from visual input and infer meaningful interaction-level information in real time. Given an input video stream, the system aims to:

- Detect human pose in each frame
- Estimate joint configurations and body posture
- Interpret motion patterns based on joint-level reasoning
- Provide feedback on correctness and repetition of actions

Formally, let $\mathcal{I}_t$ denote the input image at time t. The objective is to compute a structured representation $\mathcal{P}_t$ of the human pose and derive a higher-level interpretation $\mathcal{A}_t$ of the action being performed:

$$\mathcal{I}_t \rightarrow \mathcal{P}_t \rightarrow \mathcal{A}_t \tag{3.1}$$

where $\mathcal{P}_t$ represents joint positions and $\mathcal{A}_t$ encodes action-level semantics such as pose correctness and repetition count.

3.2 System Architecture

The overall system consists of four main components:

1. Keypoint Detection
2. Joint Angle Computation
3. Pose Classification
4. Repetition Counting and Feedback

Each component operates sequentially, forming a pipeline that transforms raw visual input into interpretable interaction signals.

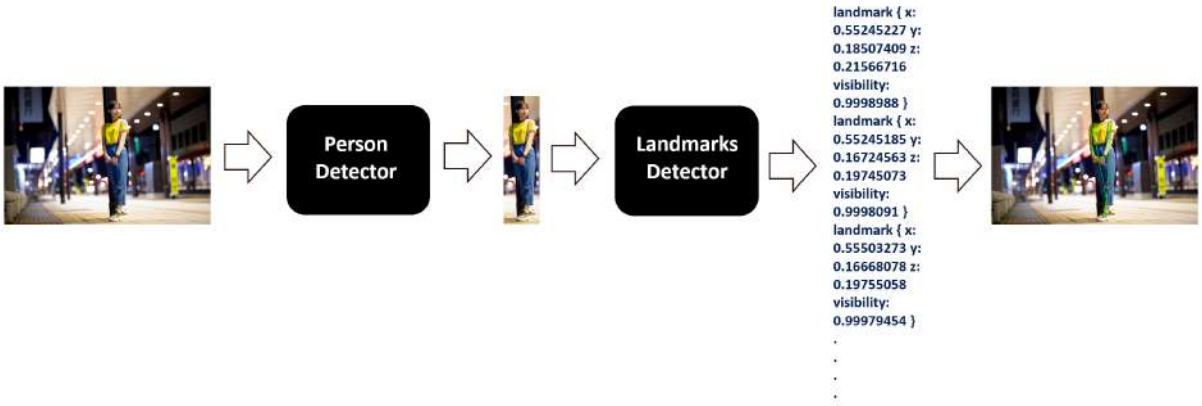

Figure 3.1 Overview of the proposed real-time human pose understanding pipeline. Given an input image, the system first detects the human subject and extracts body landmarks using a pose estimation model. These landmarks are subsequently used for joint angle computation, pose classification, temporal reasoning, and repetition counting. The pipeline enables efficient and interpretable analysis of human motion in real-world environments.

3.2.1 Keypoint Detection

The first stage of the pipeline involves detecting human keypoints from the input image. This is achieved using a pre-trained pose estimation model [7, 31] that predicts the 2D locations of key joints such as shoulders, elbows, hips, and knees.

Let $\mathbf{K}_t = \{k_1, k_2, \ldots, k_n\}$ denote the set of detected keypoints at time t. These keypoints form the basis for subsequent geometric reasoning.

This stage provides a compact representation of human posture while maintaining sufficient information for interaction analysis.

3.2.2 Pose Estimation Network

The system employs a lightweight real-time pose estimation framework based on MediaPipe Pose [31] and OpenPose [7] for extracting 2D human body landmarks.

Given an input RGB frame of resolution $H \times W$, the pose estimation model predicts a set of body keypoints:

$$\mathbf{K} = \{(x_i, y_i)\}_{i=1}^{N} \tag{3.2}$$

where N denotes the number of detected body landmarks.

The detected landmarks include:

- Shoulders
- Elbows
- Wrists
- Hips
- Knees
- Ankles

To improve robustness, Gaussian smoothing is applied to the detected landmark coordinates across consecutive frames.

3.2.3 Joint Angle Computation

Given the detected keypoints, the system computes joint angles to capture the configuration of the human body. For a joint defined by three keypoints (k_i, k_j, k_k), two vectors are computed:

$$\mathbf{v}_1 = k_i - k_j \tag{3.3}$$

$$\mathbf{v}_2 = k_k - k_j \tag{3.4}$$

The angle at the joint is then computed using the cosine similarity:

$$\theta = \cos^{-1}\left(\frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\|\|\mathbf{v}_2\|}\right) \tag{3.5}$$

This representation provides scale-invariant geometric reasoning and enables consistent pose analysis across different camera distances and body proportions.

3.2.4 Pose Classification

Each exercise is represented using a set of predefined angular constraints. The system compares the computed joint angles against exercise-specific thresholds to determine the current pose state and motion phase.

The classification process is based on geometric reasoning, where each exercise is characterized by a distinct configuration of body joints. By monitoring the evolution of joint angles over time, the system can determine whether the user is correctly performing a target exercise.

For example, squat detection is defined using hip and knee angles. During the downward motion phase, the knee angle is expected to decrease below a predefined threshold:

$$\theta_{knee} < 100^\circ \tag{3.6}$$

Similarly, during the standing phase, the knee angle must increase above:

$$\theta_{knee} > 160^\circ \tag{3.7}$$

A complete squat repetition is detected when the system observes a valid transition between these two motion states.

Similarly, push-up detection is based on elbow flexion angles. The downward motion is identified when:

$$\theta_{elbow} < 90^\circ \tag{3.8}$$

while the upward motion is identified when:

$$\theta_{elbow} > 100^\circ \tag{3.9}$$

Exercise	Primary Joint	Lower Threshold	Upper Threshold
Squat	Knee Angle	$100°$	$160°$
Push-up	Elbow Angle	$90°$	$100°$
Lunge	Knee Angle	$70°$	$140°$
Crunch	Shoulder Angle	$80°$	$170°$
V-Sit	Hip Angle	$45°$	$120°$

Table 3.1 Exercise-specific angular thresholds used for pose classification and repetition counting.

The threshold values were empirically selected based on observed joint motion ranges across multiple users and exercise types. The use of angular thresholds enables interpretable and computationally efficient pose classification while maintaining robustness across users with different body proportions and camera distances.

3.2.5 Repetition Counting and Feedback

To analyze motion over time, the system tracks transitions between pose states. A repetition is detected when the system observes a complete cycle of predefined pose transitions.

For instance, a simple exercise may involve transitioning from an initial state to a target state and back. By monitoring these transitions, the system increments a repetition counter.

Algorithm 1 Real-Time Repetition Counting

Initialize repetition counter $C \leftarrow 0$ Initialize motion state $S \leftarrow$ `UP`
for *each incoming frame* **do**
  Detect body keypoints Compute joint angles
  if $\theta < \theta_{down}$ *and* $S =$ *`UP`* **then**
    $S \leftarrow$ `DOWN`
  if $\theta > \theta_{up}$ *and* $S =$ *`DOWN`* **then**
    $C \leftarrow C + 1$ $S \leftarrow$ `UP`

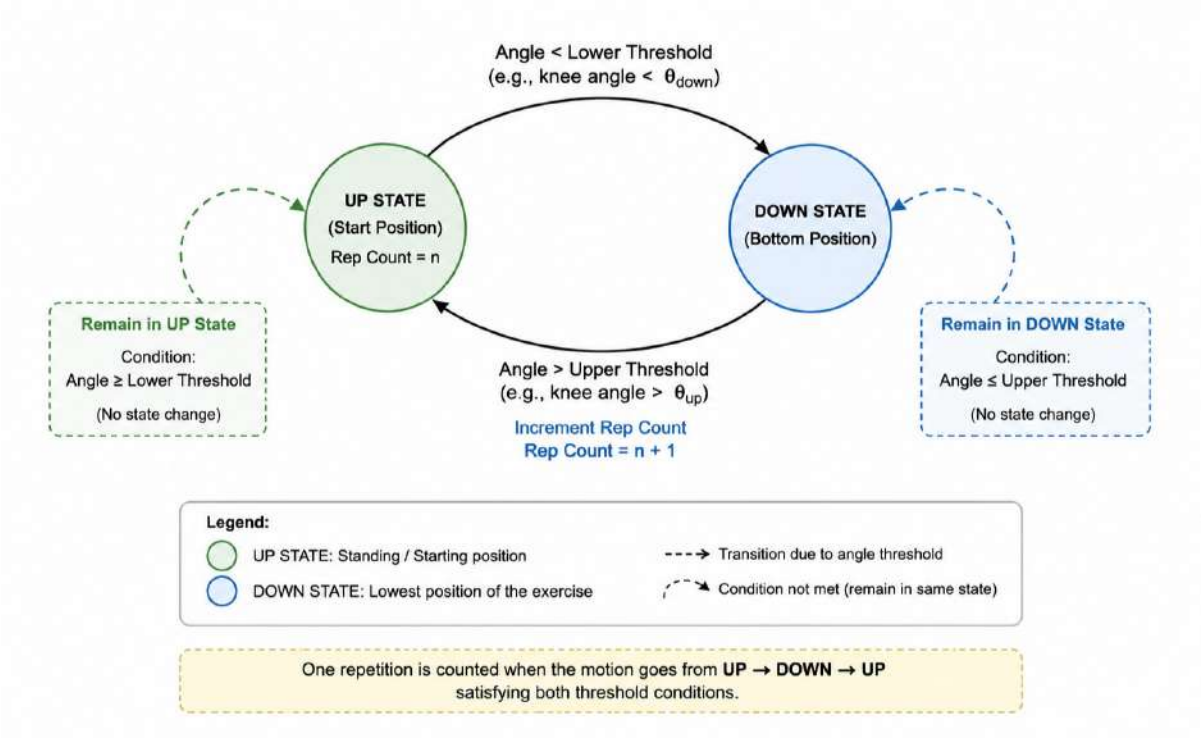

Figure 3.2 State Transition Diagram for Repetition Counting.

Additionally, the system provides real-time feedback by highlighting incorrect joint configurations and guiding the user towards correct posture.

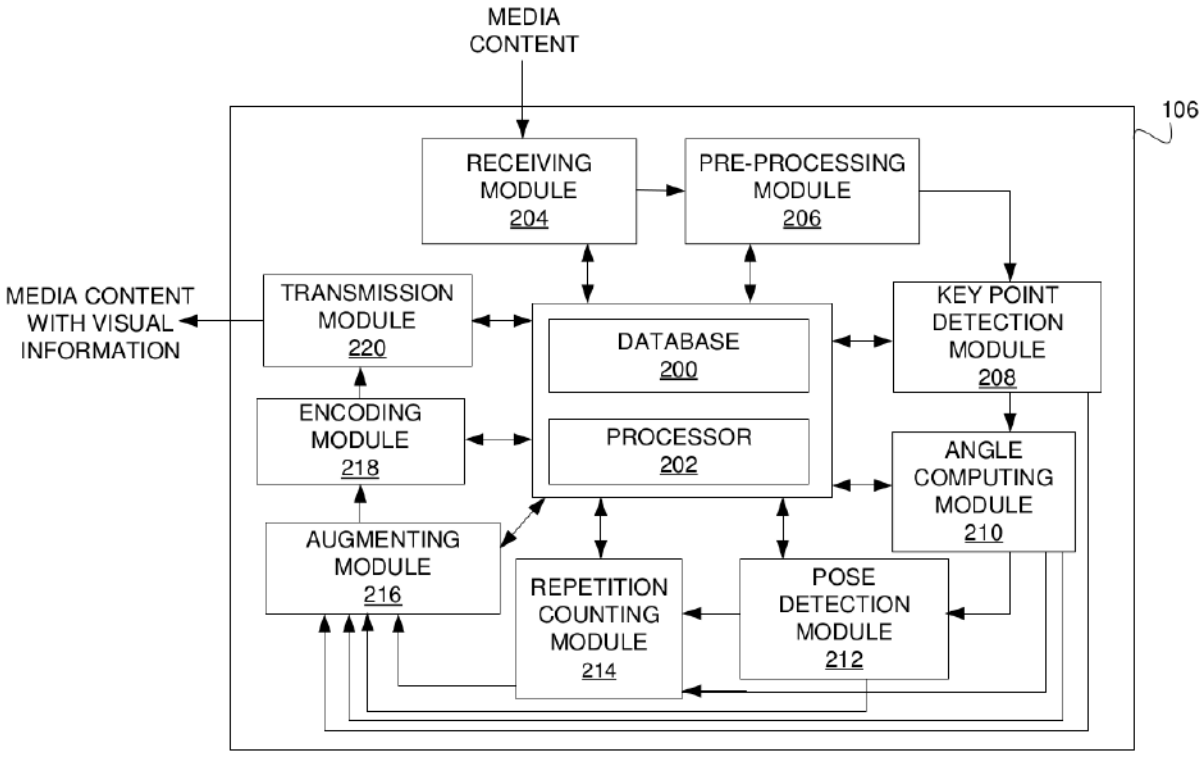

Figure 3.3 Detailed module-level architecture of the proposed exercise tracking server. The system consists of pre-processing, keypoint detection, angle computation, pose classification, repetition counting, augmentation, encoding, and transmission modules.

Component	Function
Keypoint Detection	Extract human joint positions
Joint Angle Computation	Compute geometric pose features
Pose Classification	Identify valid/invalid postures
Temporal Modeling	Track motion consistency
Feedback Module	Provide real-time guidance

Table 3.2 Summary of system components and their roles.

3.3 Temporal Modeling

While individual frames provide information about posture, interaction requires reasoning over time. The system incorporates temporal consistency by tracking joint configurations across consecutive frames.

Let $\mathcal{P}_{t-1}$ and $\mathcal{P}_t$ denote pose representations at consecutive time steps. Temporal smoothing is applied to reduce noise and ensure stable predictions:

$$\mathcal{P}_t^{smooth} = \alpha \mathcal{P}_t + (1 - \alpha)\mathcal{P}_{t-1} \tag{3.10}$$

This improves robustness in real-world scenarios where pose estimation may be noisy.

3.4 Implementation Details

The proposed system was implemented using Python and OpenCV for video processing and visualization. Pose estimation was performed using the MediaPipe framework running on RGB video streams.

The system processes video frames sequentially in real time at approximately 25–30 FPS on a standard consumer CPU.

Input frames are resized to a fixed spatial resolution before pose estimation to ensure stable runtime performance. Temporal smoothing is applied using exponential moving averages to reduce landmark jitter.

The repetition counting logic is implemented as a finite-state machine that tracks transitions between predefined motion states.

Experiments were conducted on a system equipped with an Intel i5 processor, and 8 GB RAM on CPU.

3.5 Applications

The proposed system enables a range of real-world applications:

- **Exercise Monitoring:** Tracking repetitions and ensuring correct posture during workouts
- **Rehabilitation:** Providing feedback for physical therapy exercises
- **Human-Computer Interaction:** Enabling gesture-based control systems
- **Sports Analysis:** Evaluating performance and technique

These applications demonstrate the practical utility of joint-level reasoning for understanding human motion.

3.6 Failure Cases and Limitations

Despite achieving robust real-time performance, the proposed system has several limitations.

Failure cases primarily occur under severe self-occlusion, extreme camera viewpoints, and poor illumination conditions, where the quality of pose estimation deteriorates significantly. In such scenarios, incorrect landmark localization may propagate errors into the joint angle computation and pose classification stages.

The system also assumes that a single human subject is clearly visible within the camera frame. Multi-person interactions or overlapping subjects may introduce ambiguity in keypoint association and temporal tracking.

In addition, the proposed rule-based formulation relies on predefined angular thresholds that are manually selected for each exercise. Although this enables interpretable reasoning and computational efficiency, the approach may not generalize optimally across highly diverse motion styles or unusual body configurations.

Rapid body motion and motion blur can further reduce pose estimation stability, particularly in low-frame-rate video streams. Loose clothing and cluttered backgrounds may also affect landmark detection accuracy.

Future work could explore adaptive threshold learning, 3D pose estimation, and temporal deep learning models to improve robustness under challenging real-world conditions.

3.7 Discussion

The proposed system demonstrates that efficient and interpretable human motion understanding can be achieved using a combination of pose estimation, geometric reasoning, and temporal analysis. Unlike large-scale data-driven approaches that require extensive training data and computational resources, the proposed framework leverages explicit geometric constraints to enable real-time interaction analysis with low computational overhead.

One of the key advantages of the system is its interpretability. Because pose classification and repetition counting are based on explicit angular reasoning, the decision-making process remains transparent and easy to analyze. This is particularly beneficial for applications such as rehabilitation and exercise monitoring, where understandable feedback is often more valuable than black-box predictions.

The modular architecture of the system also enables flexibility and extensibility. Individual components, including pose estimation, temporal smoothing, and repetition counting, can be independently improved or replaced without redesigning the entire pipeline. For instance, more advanced 3D pose estimation methods or learned temporal models could be integrated to improve robustness under challenging conditions.

Compared to purely learning-based systems, the proposed approach offers improved computational efficiency and easier deployment in resource-constrained environments. The use of lightweight geometric operations enables stable real-time performance on standard hardware, making the system suitable for practical deployment in home-based fitness applications, rehabilitation systems, and interactive human-computer interfaces.

At the same time, the chapter highlights the broader importance of structured human motion understanding as a foundational component for realistic human-object interaction systems. The ability to extract meaningful interaction-level semantics from raw visual input serves as an important bridge between low-level perception and higher-level reasoning about human behavior.

Overall, the proposed system demonstrates how interpretable geometric reasoning can complement modern learning-based approaches, providing a practical framework for real-world human motion analysis and interaction-aware applications.

3.8 Summary

In this chapter, we presented a real-world system for human pose understanding and interaction analysis based on joint-level reasoning. By combining keypoint detection, geometric computation, and rule-based interpretation, the system enables efficient and interpretable analysis of human motion.

This system serves as a foundation for interaction-aware modeling, bridging the gap between theoretical methods and practical deployment. In the next chapter, we build upon this foundation to develop data-driven models for scene-aware human-object interaction.

Chapter 4

Scene-Aware Human-Object Interaction

Modeling realistic human-object interaction in 3D environments requires simultaneously reasoning about full-body motion, object manipulation, and scene constraints. As highlighted in earlier chapters, existing approaches typically address these components in isolation, leading to physically implausible behavior when deployed in complex environments.

In this chapter, we address this limitation by introducing a unified framework for scene-aware human-object interaction [16], with a focus on generating realistic grasping motions that are consistent with both object geometry and the surrounding scene.

We begin by introducing MOGRAS [5], a large-scale dataset that captures full-body human motion with object interaction in realistic 3D environments. Unlike existing datasets, MOGRAS explicitly models the coupling between human motion, object interaction, and scene context, enabling the study of interaction in physically grounded settings.

Building on this dataset, we propose a learning-based framework for generating scene-aware grasping motions that explicitly incorporates object geometry and environmental constraints. Together, these contributions provide a principled approach to modeling physically plausible human-object interaction in complex 3D environments.

4.1 Problem Definition

We consider the problem of generating human motion that results in physically plausible interaction with an object in a 3D scene. Given a scene containing a target object, the goal is to generate a sequence of human poses that approaches the object and establishes a stable grasp, while respecting scene constraints.

Formally, let the scene be represented as a 3D environment containing objects and surrounding geometry. Given a target object and an initial human pose, the objective is to generate a sequence of human body configurations that satisfies the following properties:

- **Interaction Consistency:** The generated motion should result in meaningful contact between the human body (particularly the hands) and the object.

- **Physical Plausibility:** The motion should avoid interpenetration with the scene and maintain realistic body configurations.
- **Full-Body Coordination:** The interaction should involve coherent full-body motion rather than isolated hand movement.

This problem is challenging due to the high dimensionality of human motion, the complexity of object geometry, and the constraints imposed by the surrounding environment.

4.2 MOGRAS Dataset

A key limitation of existing approaches to human-object interaction is the lack of datasets that jointly capture full-body motion, object interaction, and realistic scene context. Most prior datasets focus either on human motion without object interaction [33, 68] or on hand-object interaction without full-body dynamics and scene awareness [18].

To address this gap, we introduce MOGRAS, a dataset specifically designed to capture scene-aware human-object interaction in 3D environments. The dataset consists of motion sequences in which humans approach, interact with, and grasp objects within realistic scenes, capturing the full pipeline of interaction rather than isolated snapshots.

4.2.1 Dataset Motivation

The central motivation behind MOGRAS is that interaction is inherently contextual. Realistic grasping cannot be modeled without considering how the human body moves within a scene, how objects are positioned, and how environmental constraints influence motion.

By explicitly incorporating scene geometry and object context, MOGRAS enables models to learn interaction behaviors that are physically grounded and context-aware, rather than purely kinematic.

More importantly, MOGRAS represents a shift from isolated modeling of motion or grasping toward a unified representation of interaction. Unlike prior datasets that separately study locomotion, grasping, or scene interaction, MOGRAS jointly captures the coupling between full-body motion, object manipulation, and environmental constraints.

This unified representation enables the development of models that reason simultaneously about motion, contact, and scene geometry, providing a more realistic benchmark for scene-aware human-object interaction.

4.2.2 Dataset Generation Pipeline

Generating large-scale datasets for scene-aware human-object interaction is highly challenging due to the complexity of simultaneously capturing full-body motion, fine-grained grasping behavior, and

realistic scene constraints. Manual acquisition of such data would require expensive motion capture systems, dense scene reconstruction, and extensive human annotation.

To address these limitations, we propose an automated synthesis pipeline for generating realistic interaction sequences in 3D indoor environments. The pipeline combines motion alignment, scene refinement, object placement, grasp synthesis, and motion infilling to generate physically plausible interaction sequences at scale.

Figure 4.1 provides an overview of the proposed data generation process.

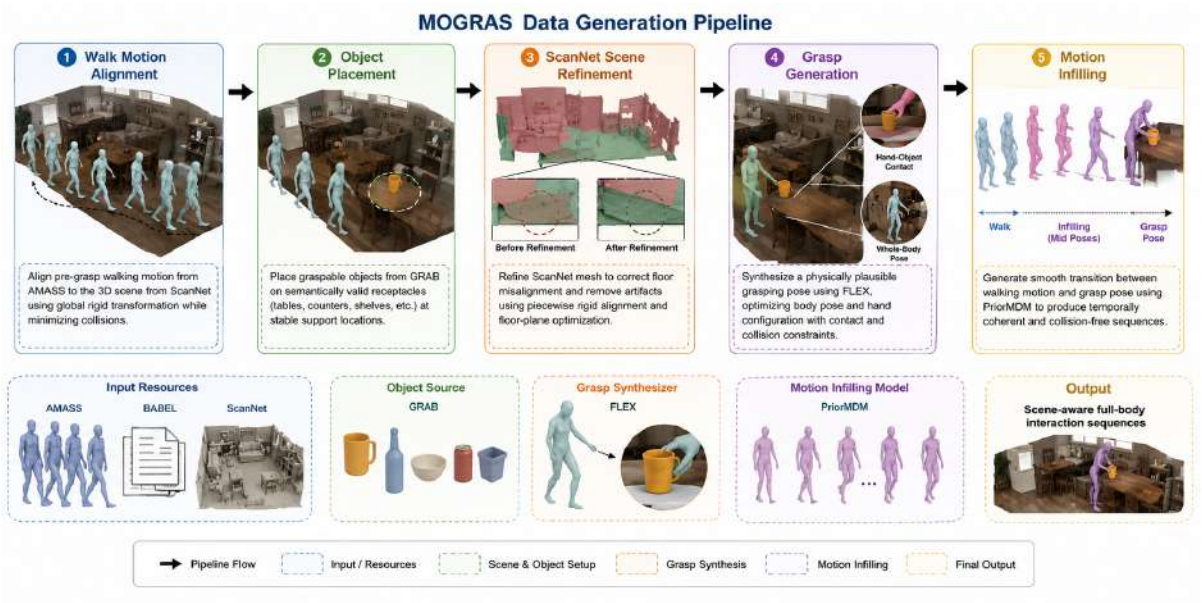

Figure 4.1 Overview of the MOGRAS data generation pipeline. The pipeline consists of walk motion alignment, object placement, ScanNet scene refinement, grasp generation, and motion infilling to produce physically plausible scene-aware interaction sequences.

The generation pipeline consists of five major stages:

1. **Walk Motion Alignment:** We first align walking motion sequences from AMASS [33] to 3D indoor scenes from ScanNet [8]. Motion annotations from BABEL [43] are used to identify walking sequences suitable for interaction. A global rigid transformation is optimized to place the human trajectory within the scene while minimizing collision with surrounding geometry.

2. **Object Placement:** After aligning the motion, graspable objects from GRAB [51] are placed on semantically valid receptacles such as tables, counters, or shelves. Receptacle surfaces are extracted using ScanNet semantic labels, and candidate object placement points are identified using nearest-neighbor search on the surface mesh. Objects are positioned at stable support locations to avoid floating artifacts and ensure reachable grasp configurations.

3. **Scene Refinement:** Raw ScanNet scenes often contain floor misalignment artifacts that lead to unrealistic foot penetration or floating effects. To improve physical plausibility, we refine scene

geometry using a piecewise rigid alignment strategy. Local rigid transformations are estimated over sliding windows on the floor surface to align floor vertices with the global ground plane while preserving scene structure.

4. **Grasp Generation:** Given the aligned motion and placed object, we synthesize a physically plausible full-body grasping pose using FLEX [52]. FLEX optimizes body pose and hand configuration jointly while enforcing contact and collision constraints. To improve scalability, we modify FLEX by reducing optimization iterations and applying post-processing refinements for wrist alignment and contact consistency.

5. **Motion Infilling:** Finally, we generate a smooth transition between the walking sequence and the synthesized grasp pose using PriorMDM [48]. The infilling model generates intermediate poses that preserve trajectory realism and temporal coherence while ensuring collision-free motion.

The resulting dataset contains full-body motion sequences consisting of both pre-grasp approaching motion and final grasping poses within richly annotated 3D indoor environments. The pipeline enables the generation of large-scale interaction data while maintaining physical plausibility and scene consistency.

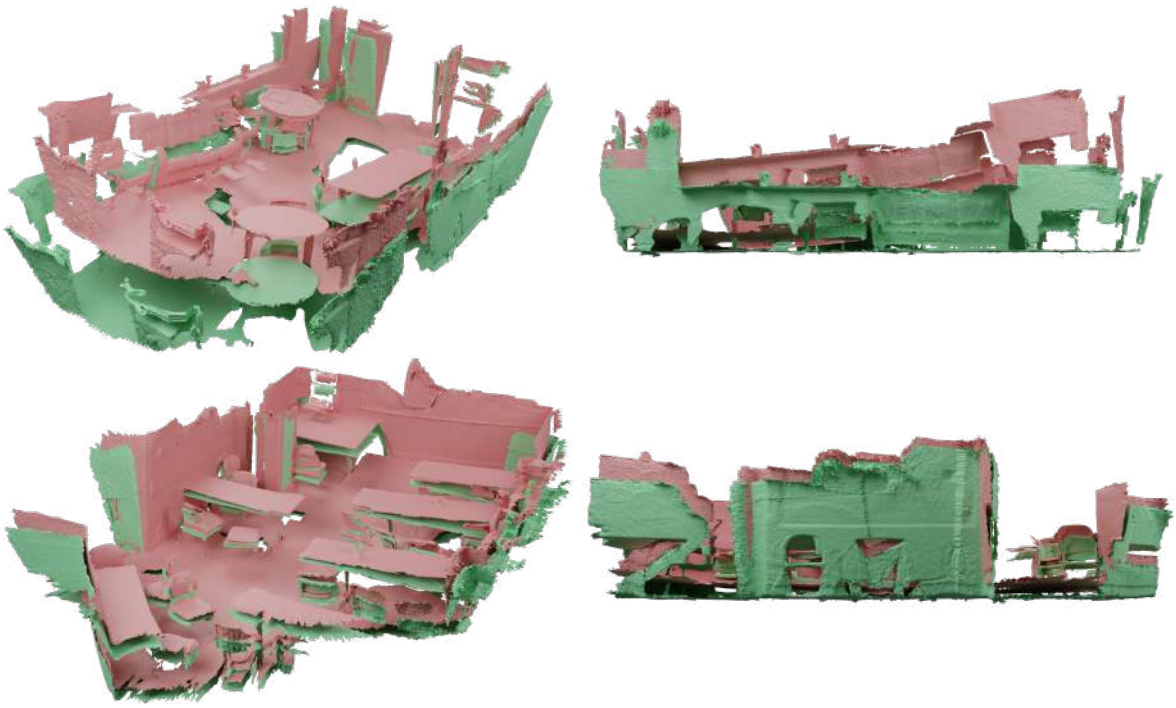

Figure 4.2 Comparison of Original and Refined ScanNet Scenes. Original scenes (pink) often have floor misalignment, causing artifacts like floating feet. Our refined scenes (green) correct this, improving physical plausibility and interaction quality.

4.2.3 Walk Motion Alignment and Object Placement

The first stage of the MOGRAS generation pipeline focuses on aligning human walking motion with realistic indoor environments and placing graspable objects in semantically meaningful locations.

We follow the alignment strategy introduced in HUMANISE [56] to place human motion sequences inside ScanNet [8] scenes. Walking motion clips are obtained from the AMASS [33] motion capture dataset, while motion annotations from BABEL [43] are used to identify sequences corresponding to walking behavior.

The dataset contains 2,988 walking motion clips with durations ranging from 1 to 4 seconds. Each motion sequence is represented using the SMPL-X [39] human body model, enabling consistent full-body pose representation across all scenes.

For each sequence, we first identify a suitable interaction region within the scene. Specifically, semantically meaningful receptacles such as tables, desks, counters, and shelves are extracted from ScanNet semantic annotations. These receptacles serve as candidate surfaces for object placement and interaction.

Given a selected receptacle, we optimize a global rigid transformation consisting of translation $\mathbf{t}$ and rotation $\mathbf{R}$ to align the walking trajectory with the scene while minimizing human-scene collision. The optimization objective ensures that:

- the walking trajectory terminates near the target object,
- the human body remains collision-free during motion,
- the motion preserves its natural trajectory and orientation.

Formally, the transformed motion sequence is defined as:

$$\mathbf{P}' = \mathbf{R}\mathbf{P} + \mathbf{t}, \tag{4.1}$$

where $\mathbf{P}$ denotes the original motion sequence and $\mathbf{P}'$ represents the aligned motion within the scene.

After motion alignment, graspable objects are placed on the selected receptacle surfaces. Objects are sourced from the GRAB [51] dataset, which provides realistic object geometries and grasp annotations.

To determine valid object placement locations, we extract the mesh corresponding to the receptacle surface and identify candidate support points using nearest-neighbor search implemented with a KD-tree. Objects are positioned at stable support locations corresponding to the highest valid surface point to avoid floating artifacts and interpenetration with the environment.

The placement process additionally enforces reachability constraints to ensure that the final walking pose remains physically capable of grasping the target object. This produces realistic interaction configurations where the human naturally approaches the object before initiating the grasp.

Figure 4.3 illustrates several examples of aligned walking trajectories and object placements generated using the proposed pipeline.

Dataset	#Samples	#Frames	#Scenes	#Objects	3D Scene Inclusion	Fine-Grained Grasping	Full-Body Motion
HUMANISE [56]	19.6k	1.2M	643	✗	✓	✗	✓
GRAB [51]	1.3k	1.6M	✗	51	✗	✓	✓
OakInk [61]	50k	0.23M	✗	1800	✗	✓	✗
SceneFun3D [10]	14.8k	✗	710	✗	✓	✗	✗
ScenePlan [58]	1.1k	✗	10	40	✓	✗	✓
MOGRAS	14.2k	0.8M	507	47	✓	✓	✓

Table 4.1 Comparison of Dataset Statistics and Functional Coverage. This table summarizes existing datasets against MOGRAS across key dimensions: the number of samples, frames, scenes, and objects. We also compare support for three critical capabilities: 3D scene inclusion, fine-grained grasping, and full-body motion. MOGRAS is the only dataset to combine all three, addressing a critical gap in the literature.

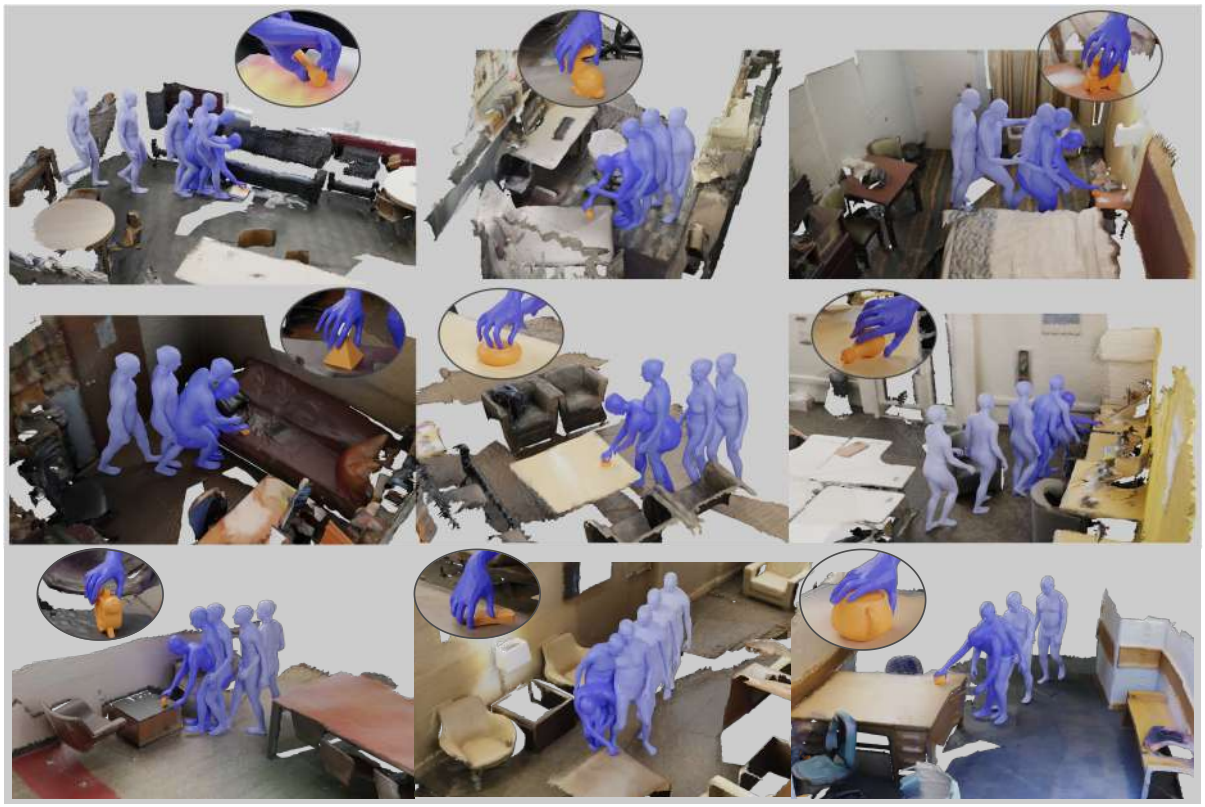

Figure 4.3 Examples from the MOGRAS dataset. MOGRAS provides full-body motion sequences, including a pre-grasp walking phase and a final grasping pose, all within richly annotated 3D scenes. The dataset captures subtle, physically plausible interactions with small objects and their environment, addressing a key gap in existing research.

4.2.4 Dataset Statistics

MOGRAS is a large-scale dataset designed for scene-aware full-body human-object interaction. The dataset contains interaction sequences spanning diverse indoor environments, object categories, and motion patterns, enabling the study of physically plausible interaction generation under realistic scene constraints.

The dataset consists of more than 14,000 motion sequences generated across 507 indoor scenes and 47 graspable object categories. Each sequence contains both a pre-grasp walking trajectory and a final grasping pose, allowing models to reason jointly about approaching behavior and object manipulation.

To ensure robust evaluation and generalization, the dataset is divided into training, validation, and test splits with non-overlapping scene configurations. Table 4.2 summarizes the statistics of each split.

Split	#Motion Samples	#Frames	#Scenes	#Objects
Train	11,479	678,226	427	37
Val	932	54,484	27	4
Test	1,827	106,079	53	6

Table 4.2 Statistics of the MOGRAS dataset splits.

Compared to existing datasets, MOGRAS uniquely combines three critical components necessary for realistic interaction modeling:

- **Full-Body Motion:** The dataset captures complete human body motion rather than isolated hand interaction, enabling realistic coordination between locomotion, posture, and grasping behavior.
- **Fine-Grained Grasping:** MOGRAS includes detailed grasp interaction with small handheld objects, allowing the study of contact-rich human-object interaction.
- **3D Scene Awareness:** Unlike scene-agnostic grasp datasets, all interaction sequences are generated within realistic indoor 3D environments while respecting environmental constraints and collision boundaries.

Table 4.1 compares MOGRAS with existing datasets across dataset scale and functional coverage. Existing datasets typically focus either on large-scale human-scene interaction without fine-grained grasping, or on grasping without realistic scene context. In contrast, MOGRAS jointly models full-body motion, fine-grained grasping, and scene-aware interaction.

To further ensure realism and physical plausibility, the generated interaction sequences undergo a multi-stage quality assurance process consisting of automated penetration filtering and human evaluation. In the user study, participants evaluated generated motions based on naturalness and interaction realism, and low-quality sequences were removed from the final dataset.

The resulting dataset provides a challenging benchmark for scene-aware human-object interaction and supports a wide range of downstream tasks, including motion generation, grasp synthesis, interaction prediction, and physically grounded human-scene reasoning.

4.3 GNet++: Scene-Aware Grasp Generation

Building on the MOGRAS dataset, we propose a learning-based framework for generating scene-aware grasping motions. The goal is to synthesize human motion that results in a stable grasp of the target object while remaining consistent with the surrounding scene.

Unlike prior approaches that model motion generation and interaction separately, our method explicitly incorporates object geometry and scene constraints into the generation process. This enables the synthesis of motion that is not only visually realistic but also physically grounded.

4.3.1 Method Overview

We formulate scene-aware grasp generation as a conditional generative modeling problem. Given a 3D scene $\mathbf{S}$, a target object $\mathbf{O}$, and an initial human pose $\mathbf{P}_0$, the goal is to generate a physically plausible full-body grasping pose that establishes stable object contact while avoiding collision with the surrounding environment.

Our proposed framework, GNet++, extends the pretrained GNet [50] architecture by explicitly incorporating scene context into the grasp generation process. Figure 4.4 presents an overview of the proposed architecture.

Unlike the original GNet formulation, which generates grasps conditioned only on object geometry, GNet++ additionally conditions the model on 3D scene information. This enables the network to reason about nearby obstacles, free space, and environmental constraints during grasp generation.

The overall framework consists of four major components:

1. **Object Encoding**

   The target object is represented using Basis Point Set (BPS) encoding [42], which converts the object geometry into a compact feature representation suitable for neural processing. The BPS representation captures local geometric structure while remaining invariant to mesh topology and point density variations.

   Given an object mesh $\mathcal{O}$, the object encoder produces an object feature embedding:

   $$\mathbf{f}_{obj} = E_{obj}(\mathcal{O}), \tag{4.2}$$

   where E_{obj} denotes the object encoder network.

2. **Scene Encoding**

   To incorporate environmental awareness, we encode the surrounding 3D scene using a pretrained Vision Transformer (ViT) [11]. Specifically, we project the scene into a bird's-eye-view (BEV) representation and extract a scene embedding that captures spatial layout and obstacle information.

The scene encoder generates a feature representation:

$$\mathbf{f}_{scene} = E_{scene}(\mathcal{S}), \tag{4.3}$$

where E_{scene} represents the ViT-based scene encoder.

The BEV representation provides an efficient mechanism for modeling spatial occupancy and nearby obstacles while maintaining low computational complexity.

3. **Latent Grasp Representation**

   GNet++ adopts a conditional variational autoencoder (cVAE) formulation to model diverse grasping behaviors.

   During training, the encoder maps the human pose, object embedding, and scene embedding into a latent distribution:

$$q(\mathbf{z} \mid \mathbf{P}, \mathbf{f}_{obj}, \mathbf{f}_{scene}), \tag{4.4}$$

   where $\mathbf{z}$ represents a latent grasp code capturing variations in body posture, grasp configuration, and interaction style.

   The latent distribution is parameterized by a mean vector $\boldsymbol{\mu}$ and variance vector $\boldsymbol{\sigma}$:

$$\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}). \tag{4.5}$$

   This latent representation enables the model to generate diverse grasp configurations while preserving interaction realism and physical plausibility.

4. **Conditional Grasp Decoder**

   The decoder generates the final full-body grasping pose conditioned on the latent grasp code, object embedding, and scene embedding.

   Formally, the decoder predicts SMPL-X body parameters:

$$\hat{\mathbf{P}} = D(\mathbf{z}, \mathbf{f}_{obj}, \mathbf{f}_{scene}), \tag{4.6}$$

   where $D(\cdot)$ denotes the conditional decoder network.

   The decoder outputs:

   - body pose parameters,
   - global orientation,
   - hand articulation,
   - contact refinement offsets.

Conditioning the decoder on scene context allows the generated grasp to adapt to surrounding geometry, significantly reducing human-scene interpenetration compared to scene-agnostic baselines.

To further improve physical plausibility, we introduce an explicit penetration-aware loss during training. The surrounding scene is voxelized, and predicted human mesh vertices that intersect occupied voxels are penalized. This encourages the model to generate collision-free grasps that remain consistent with environmental constraints.

Overall, GNet++ extends the original GNet framework into a fully scene-aware grasp generation model capable of synthesizing realistic, physically plausible full-body interactions within complex 3D indoor environments.

4.3.2 Scene and Object Representation

Scene Representation: The scene is represented using a 3D geometric structure capturing spatial layout and surrounding objects. This enables reasoning about collision and free space.

Object Representation: The target object is represented using its geometry and pose, providing necessary information for generating valid grasp configurations.

Human Representation: The human body is modeled using a parametric representation, encoding pose and shape in a compact latent space. This enables consistent full-body motion generation.

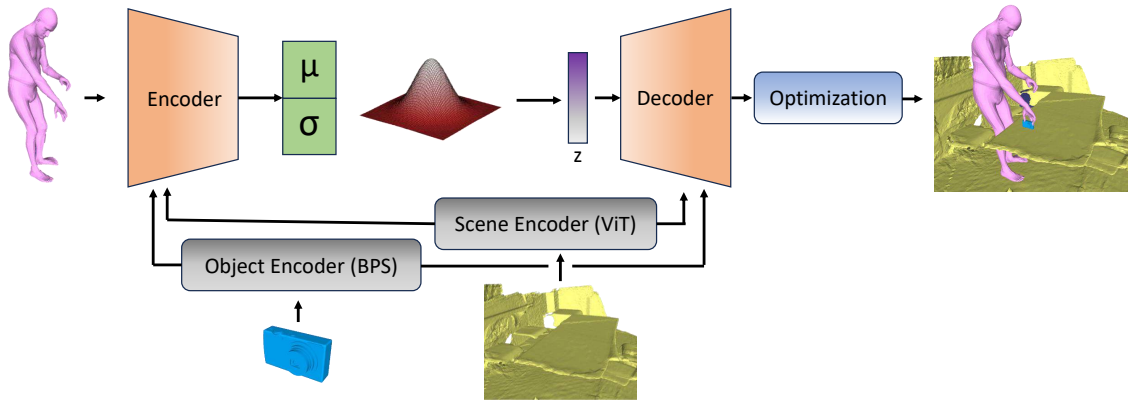

Figure 4.4 Overview of the proposed scene-aware grasp generation framework. The model takes as input scene geometry, object representation, and an initial human pose, and generates interaction-aware motion that respects both contact and environmental constraints.

4.3.3 Latent Grasp Generation

The core of the framework is a conditional generative model that produces motion sequences conditioned on scene and object context. During training, the model learns to reconstruct ground-truth motion while conditioning on $\mathbf{S}$ and $\mathbf{O}$.

At inference time, new motion sequences are generated by sampling from the latent space:

$$\mathbf{z} \sim p(\mathbf{z}), \quad \mathbf{P}_{1:T} \sim p(\mathbf{P}_{1:T} \mid \mathbf{z}, \mathbf{S}, \mathbf{O}). \tag{4.7}$$

This allows the model to generate diverse yet physically consistent interaction behaviors.

4.3.4 Penetration-Aware Optimization

A key component of the framework is the integration of interaction constraints directly into the generation process. The model is trained to minimize interpenetration between the human body and the scene while encouraging stable contact with the object.

A key insight of our approach is that enforcing interaction constraints during generation is essential for realism. Naively generating motion and subsequently applying constraints often leads to unstable or inconsistent behavior. By incorporating these constraints into the learning process, the model produces interaction-aware motion that remains physically plausible throughout the sequence.

4.4 Training and Loss Functions

The proposed framework is trained to generate realistic, physically plausible, and interaction-consistent human motion. The training objective combines reconstruction accuracy with constraints that enforce contact, physical plausibility, and temporal coherence.

4.4.1 Training Objective

Given a training sample consisting of scene $\mathbf{S}$, object $\mathbf{O}$, and ground-truth motion $\mathbf{P}_{1:T}$, the model is trained using the following objective:

$$\mathcal{L}_{total} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{contact}\mathcal{L}_{contact} + \lambda_{pen}\mathcal{L}_{pen} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{temp}\mathcal{L}_{temp} + \lambda_{foot}\mathcal{L}_{foot}. \tag{4.8}$$

4.4.2 Reconstruction Loss

The reconstruction loss ensures that generated motion matches ground-truth motion:

$$\mathcal{L}_{rec} = \|\mathbf{P}_{pred} - \mathbf{P}_{gt}\|_2^2. \tag{4.9}$$

4.4.3 Contact Loss

To encourage meaningful interaction, we enforce contact between hand vertices $\mathcal{V}_{hand}$ and object surface $\mathcal{S}_{obj}$:

$$\mathcal{L}_{contact} = \sum_{v \in \mathcal{V}_{hand}} \min_{s \in \mathcal{S}_{obj}} \|v - s\|_2^2. \tag{4.10}$$

4.4.4 Penetration Loss

To ensure physical plausibility, we penalize interpenetration with the scene:

$$\mathcal{L}_{pen} = \sum_{v \in \mathcal{V}_{body}} \phi(v, \mathcal{S}_{scene}), \tag{4.11}$$

where $\phi(\cdot)$ measures penetration using signed distance functions.

4.4.5 KL Divergence Loss

To regularize the latent space, we include a KL divergence term:

$$\mathcal{L}_{KL} = D_{KL}\big(q(\mathbf{z} \mid \mathbf{P}, \mathbf{S}, \mathbf{O}) \parallel p(\mathbf{z})\big). \tag{4.12}$$

This encourages the learned latent distribution to remain close to a prior distribution, enabling smooth and diverse motion generation.

4.4.6 Temporal Smoothness Loss

To ensure smooth transitions between frames, we introduce a temporal consistency term:

$$\mathcal{L}_{temp} = \sum_{t=2}^{T} \|\mathbf{P}_t - \mathbf{P}_{t-1}\|_2^2. \tag{4.13}$$

This reduces jitter and improves motion coherence.

4.4.7 Foot Contact / Stability Loss

To ensure physically stable motion, we enforce consistency of foot contact with the ground:

$$\mathcal{L}_{foot} = \sum_{t \in \mathcal{T}_{contact}} \|\mathbf{v}_{foot}^{t} - \mathbf{v}_{foot}^{t-1}\|_2^2, \tag{4.14}$$

where $\mathcal{T}_{contact}$ denotes time steps where foot contact occurs.

This term prevents unrealistic sliding and improves global motion stability.

4.5 Implementation Details

This section describes the implementation details of the proposed MOGRAS generation pipeline and the GNet++ framework, including dataset preprocessing, training configuration, scene representation, and modifications introduced to the baseline FLEX [52] framework.

4.5.1 Dataset Preprocessing

All human motion sequences are represented using the SMPL-X [39] parametric body model, which provides a unified representation for full-body pose, hand articulation, and global body orientation.

Walking motion sequences are obtained from the AMASS [33] dataset and aligned to ScanNet [8] indoor scenes using global rigid transformations. Before alignment, all scene meshes are normalized to a common coordinate system with the floor aligned to the global ground plane.

To improve physical realism, ScanNet scenes undergo a refinement stage that corrects floor misalignment artifacts using local rigid alignment. The refined scenes are subsequently used for motion alignment and collision checking.

For scene-aware grasp generation, each scene is voxelized around the interaction region to construct an occupancy representation. Occupied voxels correspond to scene geometry and are used for computing the penetration-aware loss during training.

Additionally, bird's-eye-view (BEV) projections of the scene are generated and used as input to the Vision Transformer (ViT)-based scene encoder. The BEV representation provides an efficient encoding of scene layout and nearby obstacles while reducing computational complexity.

4.5.2 Training Configuration

GNet++ is initialized from pretrained GNet [50] weights and subsequently fine-tuned on the MOGRAS dataset.

The model is implemented using PyTorch and trained using the Adam optimizer with default momentum parameters:

$$\beta_1 = 0.9, \quad \beta_2 = 0.999. \tag{4.15}$$

Training is performed using a batch size of 32 and an initial learning rate of:

$$\eta = 1 \times 10^{-4}. \tag{4.16}$$

The learning rate is decayed during training using a cosine annealing schedule to improve convergence stability.

The network is trained for 100 epochs on NVIDIA RTX-series GPUs. Gradient clipping is applied during optimization to stabilize training and prevent exploding gradients.

During training, latent grasp codes are sampled from a Gaussian prior, and reconstruction, contact, penetration, KL-divergence, and temporal losses are jointly optimized.

4.5.3 Scene Representation and Voxelization

To enforce scene-aware physical constraints, the surrounding 3D scene is converted into a voxel occupancy grid centered around the target interaction region.

Given a scene mesh $\mathcal{S}$, voxel occupancy is computed by discretizing the scene into a uniform grid:

$$V(i, j, k) \in \{0, 1\}, \tag{4.17}$$

where occupied voxels correspond to regions containing scene geometry.

To account for unreachable regions beneath surfaces such as tables or shelves, voxels below occupied cells are also marked as occupied. This produces a conservative occupancy representation that discourages human-scene penetration during training.

During optimization, predicted human body vertices that intersect occupied voxels are penalized using the proposed penetration-aware loss.

4.5.4 FLEX Modifications

The original FLEX [52] framework is computationally expensive for large-scale dataset generation due to its iterative optimization procedure. To improve scalability while preserving interaction quality, we introduce several modifications to the optimization process.

Reduced Optimization Iterations: The number of optimization iterations is reduced to improve generation speed while maintaining stable grasp quality. This significantly accelerates dataset synthesis and enables large-scale generation of interaction sequences.

Scene-Aware Grasp Selection: Unlike the original FLEX formulation, our modified pipeline explicitly incorporates scene constraints during grasp generation. Candidate grasp poses that produce excessive human-scene collision are rejected, ensuring physically plausible interactions.

Wrist Alignment Correction: After optimization, a post-processing refinement step is applied to align wrist orientation with the target object. This prevents unnatural hand articulation and improves contact realism.

Contact Consistency Refinement: We additionally enforce contact consistency by checking the distance between hand vertices and object surfaces. Minor contact deviations are corrected using local pose refinement to improve grasp stability and reduce unrealistic body compensation.

These modifications significantly improve the efficiency and physical plausibility of generated interaction sequences, making the proposed pipeline suitable for large-scale scene-aware human-object interaction generation.

4.6 Experiments and Results

We evaluate the proposed scene-aware human-object interaction framework on the MOGRAS dataset, with the goal of assessing realism, physical plausibility, and interaction quality of the generated motion.

Our evaluation focuses on both geometric correctness (e.g., penetration) and interaction fidelity (e.g., contact consistency), enabling a comprehensive assessment of the generated behavior.

4.6.1 Experimental Setup

All experiments are conducted on the MOGRAS dataset using the train, validation, and test splits summarized in Table 4.2. The training split consists of 11,479 motion sequences across 427 indoor scenes, while the validation and test splits contain 932 and 1,827 sequences, respectively. The dataset covers diverse scene layouts, object categories, and interaction configurations, enabling comprehensive evaluation of scene-aware grasp generation.

Hardware Configuration: Experiments are performed on a workstation equipped with NVIDIA RTX-series GPUs, Intel i7/i9 processors, and 32 GB RAM. Training and inference are implemented using PyTorch with CUDA acceleration.

Training Protocol: GNet++ is initialized using pretrained GNet [50] weights and fine-tuned on the MOGRAS dataset. During training, motion sequences are normalized to a common coordinate system and represented using SMPL-X [39] body parameters.

The network is trained using the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 32. Training is performed for 100 epochs with cosine learning-rate decay.

Scene Representation: For scene-aware reasoning, each ScanNet [8] scene is voxelized around the interaction region to construct an occupancy representation. The voxel grid is centered near the target object and encodes occupied regions corresponding to scene geometry.

In addition, bird's-eye-view (BEV) projections of the scene are generated and encoded using a Vision Transformer (ViT)-based scene encoder. This representation enables efficient modeling of nearby obstacles and spatial layout.

Evaluation Metrics: We evaluate the generated grasps using four metrics that measure physical plausibility and interaction quality.

- **Human-Scene Penetration** ($\downarrow$)**:** Measures the percentage of human body vertices that intersect occupied scene voxels. Lower values indicate fewer collisions with surrounding geometry.

- **Human-Object Penetration** ($\downarrow$)**:** Measures penetration between the generated human mesh and the target object using signed distance field (SDF) values computed at contact vertices.

- **Human-Floor Penetration** ($\downarrow$)**:** Measures the fraction of foot vertices located below the floor plane, penalizing unrealistic floating or sinking artifacts.

- **Contact Precision/Recall/F1** (↑)**:** Evaluates the quality of predicted contact regions between the human body and the object or floor. Precision measures the correctness of predicted contacts, recall measures contact coverage, and the F1 score provides a balanced evaluation of interaction quality.

Evaluation Protocol: For penetration evaluation, the generated human mesh is tested against the voxelized scene occupancy representation. Human vertices located inside occupied voxels are considered penetrating vertices.

Contact metrics are computed by comparing predicted contact vertices against ground-truth contact regions derived from the MOGRAS dataset. A contact prediction is considered correct if the distance between predicted and ground-truth contact regions falls below a predefined threshold.

We compare GNet++ against two state-of-the-art baselines: GOAL [50] and SAGA [57]. In addition, we evaluate ablated variants of GNet++ to analyze the importance of scene conditioning and penetration-aware training.

4.6.2 Quantitative Results

Table 4.3 summarizes the quantitative comparison between our method and baseline approaches. Our method consistently achieves lower penetration errors and improved contact consistency, indicating more physically plausible and interaction-aware motion.

In particular, we observe a significant reduction in human-scene and human-object interpenetration compared to methods that do not explicitly model scene constraints. This demonstrates the effectiveness of incorporating scene awareness into the generation process.

	Penetration (↓)			**Object Contact** (↑)			**Floor Contact** (↑)		
Method	Scene	Object	Floor	Precision	Recall	F1	Precision	Recall	F1
GOAL [50]	4.35%	3.49%	3.62%	0.85	0.76	0.80	0.70	0.83	0.76
SAGA [57]	5.80%	2.67%	1.14%	0.77	0.71	0.74	0.81	0.83	0.82
GNet++ (w/o MOGRAS ft)	4.47%	2.48%	0.58%	0.85	0.77	0.81	0.72	0.84	0.78
GNet++ (w/o pen. loss)	3.91%	2.07%	0.098%	0.74	0.70	0.72	0.81	0.88	0.84
GNet++ (Ours)	**3.13%**	**1.45%**	**0.029%**	**0.86**	**0.79**	**0.82**	**0.90**	**0.95**	**0.92**

Table 4.3 Quantitative comparison with baseline methods. Lower is better.

4.6.3 Qualitative Results

Figure 4.5 presents qualitative examples of generated human-object interactions. Our method produces realistic full-body motion that smoothly approaches and grasps objects while respecting scene constraints.

Compared to baseline methods, which often result in unrealistic artifacts such as object penetration or unstable grasps, our approach generates interaction sequences that are visually coherent and physically plausible.

Figure 4.5 Qualitative comparison of generated interactions.

4.6.4 Ablation Study

To analyze the contribution of individual components of GNet++, we perform an ablation study focusing on two key aspects of the proposed framework:

- fine-tuning on the MOGRAS dataset,
- the proposed penetration-aware loss.

The quantitative results are summarized in Table 4.3.

4.6.4.1 Effect of MOGRAS Fine-Tuning

To evaluate the importance of the proposed dataset, we compare the full GNet++ model against a variant that does not use MOGRAS fine-tuning.

Specifically, comparing *GNet++ (w/o MOGRAS ft)* and *GNet++ (Ours)* demonstrates the significant impact of training on scene-aware interaction data.

Without fine-tuning on MOGRAS, the model achieves a scene penetration of 4.47%, whereas the full model reduces this to 3.13%. Similarly, object penetration decreases from 2.48% to 1.45%, while floor penetration is reduced from 0.58% to 0.029%.

We also observe improvements in contact quality. Floor-contact F1 score improves from 0.78 to 0.92 after fine-tuning on MOGRAS, indicating substantially more stable and physically plausible interaction with the environment.

These results demonstrate that MOGRAS provides critical examples of full-body interaction within realistic scenes, enabling the model to learn scene-aware body configurations and collision avoidance behavior.

4.6.4.2 Effect of Penetration-Aware Loss

We further evaluate the contribution of the proposed penetration-aware loss by comparing the full GNet++ model against the variant trained without penetration supervision.

Comparing *GNet++ (w/o pen. loss)* and *GNet++ (Ours)* shows that the penetration-aware loss plays a crucial role in enforcing physical plausibility.

Without the penetration loss, scene penetration increases from 3.13% to 3.91%, while object penetration increases from 1.45% to 2.07%.

Although the model without penetration loss achieves reasonable contact scores, the generated poses frequently intersect nearby scene geometry, particularly around tables, shelves, and receptacle surfaces.

The explicit penetration-aware supervision encourages the network to avoid occupied scene regions during grasp generation, resulting in collision-free and physically consistent interaction behavior.

4.6.4.3 Discussion

The ablation results demonstrate that both scene-aware conditioning and penetration-aware optimization are essential for realistic grasp generation in cluttered 3D environments.

Fine-tuning on MOGRAS enables the model to learn the relationship between human motion, object interaction, and surrounding scene geometry. In contrast, scene-agnostic training often produces grasps that appear plausible in isolation but fail under environmental constraints.

Similarly, the proposed penetration-aware loss provides direct geometric supervision that discourages unrealistic human-scene intersections. The combination of scene conditioning and explicit physical constraints enables GNet++ to generate physically plausible full-body grasps that substantially outperform existing baselines.

4.6.5 Analysis of Results

The experimental results provide strong evidence that our approach effectively addresses the limitations of existing methods by jointly modeling motion, interaction, and scene constraints. The combination of quantitative improvements and qualitative realism highlights the importance of a unified framework for scene-aware human-object interaction.

Furthermore, the ablation study confirms that incorporating contact modeling and physical constraints is essential for generating realistic and physically plausible motion.

4.7 Discussion

The results presented in this chapter demonstrate the importance of jointly modeling full-body motion, object interaction, and scene constraints. By integrating these components within a unified framework, the proposed approach is able to generate motion that is both visually coherent and physically plausible.

A key contribution of the proposed framework lies in the combination of reconstruction, contact, penetration, temporal consistency, and latent regularization objectives. Together, these loss functions enable the model to balance realism, interaction fidelity, motion diversity, and physical plausibility. In particular, the contact and penetration losses enforce interaction consistency with both the target object and surrounding environment, while KL regularization and temporal smoothing encourage stable and diverse motion generation.

A major strength of the method lies in its ability to capture the coupling between motion and interaction. Unlike prior approaches that treat these aspects independently, our framework models their interdependence, leading to more stable and realistic behavior.

The experimental results further highlight the importance of explicit scene-aware conditioning. By incorporating environmental constraints directly into the generation process, the model produces grasping motions that significantly reduce human-scene penetration while maintaining stable object contact.

However, several limitations remain. The method relies on high-quality scene geometry and object representations, which may not always be available in real-world settings. Additionally, while the model captures a wide range of grasping behaviors, it does not explicitly model physical forces or long-horizon planning.

Future work could explore integrating physics-based interaction simulation, multi-step task planning, and dynamic scene understanding to further improve realism and interaction generalization.

Despite these limitations, the proposed framework represents a significant step towards realistic human-object interaction, providing a strong foundation for future work in both simulation and real-world applications.

4.8 Summary

In this chapter, we presented a unified approach to modeling scene-aware human-object interaction in 3D environments. We first introduced the MOGRAS dataset, which captures full-body human motion with object interaction in realistic scenes, addressing a key gap in existing datasets.

Building on this dataset, we proposed a learning-based framework for generating grasping motions that jointly account for object geometry and scene constraints. By incorporating interaction-aware objectives such as contact consistency and penetration avoidance, the model is able to produce physically plausible and realistic motion.

Through extensive experiments, we demonstrated that the proposed method outperforms existing approaches in both quantitative metrics and qualitative realism. The results highlight the importance of jointly modeling motion, interaction, and scene context.

Overall, this chapter establishes a foundation for realistic human-object interaction in 3D environments. In the next chapter, we focus on efficient geometry processing through feature-aware mesh simplification, enabling scalable deployment of interaction-aware models in complex 3D environments.

Chapter 5

Efficient Geometry Processing for Interaction-Aware Systems

The methods presented in previous chapters enable realistic human-object interaction through data-driven modeling. However, these approaches rely on high-resolution 3D geometry, which introduces significant computational overhead in real-world applications.

In this chapter, we address this challenge by developing an efficient geometry processing framework that preserves interaction-relevant features while enabling scalable computation. Specifically, we propose a feature-aware extension of quadric error metrics (QEM) [12] that incorporates geometric and interaction-based cues into the simplification process.

This chapter complements the learning-based methods by providing a systems-level contribution that bridges the gap between high-fidelity modeling and real-world applicability.

5.1 Problem Formulation

We consider the problem of simplifying a high-resolution 3D mesh while preserving features that are critical for human-object interaction. Given an input mesh $\mathcal{M}$ consisting of vertices, edges, and faces, the goal is to produce a simplified mesh $\mathcal{M}'$ with significantly fewer elements while maintaining geometric and structural fidelity.

Formally, the objective is to minimize a geometric error metric subject to constraints that preserve important features. Let $\mathcal{E}(\mathcal{M}, \mathcal{M}')$ denote the approximation error between the original and simplified mesh. The problem can be expressed as:

$$\min_{\mathcal{M}'} \mathcal{E}(\mathcal{M}, \mathcal{M}') \tag{5.1}$$

subject to:

- **Feature Preservation:** Important geometric features such as sharp edges, boundaries, and high-curvature regions should be retained.
- **Topological Consistency:** The simplified mesh should preserve the topology of the original mesh.

- **Interaction Fidelity:** Regions relevant for human-object interaction, such as contact surfaces, should remain accurate.

This problem introduces additional challenges beyond traditional mesh simplification. Standard approaches focus primarily on minimizing geometric error, often at the expense of feature preservation. However, in the context of interaction-aware systems, preserving features that influence contact and collision is essential.

To address these challenges, we propose a feature-aware extension of the QEM framework that incorporates geometric and interaction-based cues into the simplification process.

$$\mathcal{E}_{total} = \mathcal{E}_{geom} + \lambda_f \mathcal{E}_{feature}, \tag{5.2}$$

where $\mathcal{E}_{geom}$ represents geometric approximation error and $\mathcal{E}_{feature}$ captures feature-aware penalties for preserving curvature, boundaries, and interaction regions.

5.2 Background: Mesh Simplification

Mesh simplification is a fundamental problem in geometry processing, aiming to reduce the complexity of a 3D mesh while preserving its visual and structural properties. Given the high resolution of modern 3D assets, simplification is essential for enabling efficient storage, rendering, and interaction.

5.2.1 Overview of Mesh Simplification

A typical mesh consists of vertices, edges, and faces that define the surface geometry of an object or scene. Mesh simplification reduces the number of elements in this representation by iteratively removing or collapsing components while minimizing the resulting approximation error.

Various simplification strategies have been proposed [2, 29], including vertex decimation, edge collapse, and clustering-based methods. Among these, edge collapse has emerged as one of the most widely used approaches due to its flexibility and efficiency. In this paradigm, edges are iteratively collapsed to reduce mesh complexity while attempting to preserve geometric fidelity.

5.2.2 Quadric Error Metrics (QEM)

One of the most influential methods for mesh simplification is the Quadric Error Metrics (QEM) framework. QEM assigns an error metric to each vertex that quantifies the cost of approximating the surrounding geometry.

For a given vertex, a quadric matrix is constructed by accumulating the squared distances to the planes of adjacent faces. Let a plane be defined by the equation $ax + by + cz + d = 0$. This plane can be represented as a vector $\mathbf{p} = [a, b, c, d]^T$, and the corresponding quadric matrix is given by:

$$\mathbf{Q} = \mathbf{p}\mathbf{p}^T. \tag{5.3}$$

The error associated with placing a vertex at position $\mathbf{v}$ is then:

$$\mathcal{E}(\mathbf{v}) = \mathbf{v}^T \mathbf{Q} \mathbf{v}. \tag{5.4}$$

During simplification, when an edge connecting two vertices is collapsed, their corresponding quadrics are combined, and the optimal new vertex position is computed by minimizing the resulting error.

This formulation allows efficient evaluation of edge collapse costs and enables fast simplification of large meshes.

5.2.3 Advantages of QEM

The QEM framework offers several advantages:

- **Efficiency:** The error metric can be computed and updated efficiently, enabling scalable simplification.
- **Simplicity:** The formulation is mathematically elegant and easy to implement.
- **Good Approximation Quality:** QEM produces visually acceptable results for many applications.

These properties have made QEM a standard choice for mesh simplification in graphics and geometry processing.

5.2.4 Limitations of Standard QEM

Despite its effectiveness, standard QEM has several limitations, particularly in the context of interaction-aware systems.

First, QEM primarily focuses on minimizing geometric error and does not explicitly account for important features such as sharp edges, boundaries, or high-curvature regions. As a result, these features may be smoothed out during simplification.

Second, the method does not consider semantic or functional importance. In interaction scenarios, certain regions of the mesh—such as contact surfaces—are more critical than others. Standard QEM treats all regions uniformly, which can lead to degradation of interaction-relevant features.

Third, QEM does not incorporate appearance-related information such as textures or materials. This can result in visual artifacts, particularly when simplifying textured meshes.

Finally, in complex scenes with multiple objects and boundaries, QEM may introduce inconsistencies or artifacts near boundaries and intersections.

5.2.5 Towards Feature-Aware Simplification

These limitations motivate the need for extensions to the QEM framework that incorporate additional information beyond purely geometric error. In particular, preserving features relevant to human-object interaction requires incorporating cues such as curvature, boundaries, and contact regions into the simplification process.

In the next section, we introduce a feature-aware extension of QEM that addresses these challenges and enables efficient yet interaction-preserving mesh simplification.

5.3 Feature-Aware Quadric Error Metrics

To address the limitations of standard QEM discussed in the previous section, we propose a feature-aware extension of the quadric error metrics framework. The key idea is to augment the geometric error formulation with additional constraints and weighting mechanisms that preserve features critical for human-object interaction.

A key distinction of our approach is that feature importance is explicitly encoded into the error metric itself, rather than applied as a post-processing constraint. This allows the simplification process to naturally prioritize interaction-relevant regions during optimization, rather than correcting errors after they occur.

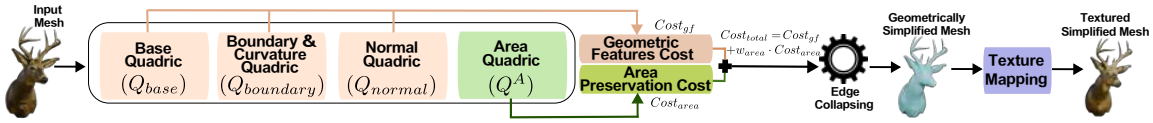

Figure 5.1 The FA-QEM pipeline. We form a composite geometric-feature quadric Q_{gf} by combining the base, boundary–curvature, and normal-alignment quadrics, yielding the geometric-feature cost $Cost_{gf}$. In parallel, we compute the area-preservation quadric Q^A and its cost $Cost_{area}$. The final collapse priority uses a weighted sum of these costs to obtain $Cost_{total}$. During simplification, successive mapping ensures consistent texture transfer. See the main text for full details.

5.3.1 Overview

Standard QEM treats all regions of the mesh uniformly, minimizing geometric error without considering the relative importance of different surface regions. However, in interaction-aware systems, certain regions—such as contact surfaces, boundaries, and high-curvature areas—play a more significant role.

Our approach introduces feature-aware weighting and constraints into the QEM formulation, enabling the simplification process to prioritize the preservation of interaction-relevant features.

5.3.2 Feature-Aware Quadric Formulation

In the standard QEM framework, each vertex is associated with a quadric matrix $\mathbf{Q}$ that accumulates geometric error. We extend this formulation by introducing a weighting function that modulates the contribution of each face based on its importance.

Let $\mathbf{Q}_i$ denote the quadric associated with a face or vertex. We define a weighted quadric as:

$$\mathbf{Q}_i^{'} = w_i \mathbf{Q}_i, \tag{5.5}$$

where w_i is a feature-aware weight that reflects the importance of the corresponding region.

The combined quadric for a vertex is then given by:

$$\mathbf{Q} = \sum_i \mathbf{Q}_i^{'}. \tag{5.6}$$

By adjusting the weights w_i, the simplification process can prioritize certain regions over others.

5.3.3 Feature Weighting Strategy

We design the weighting function to capture multiple geometric and interaction-related cues:

- **Curvature-Based Weighting:** Regions with high curvature are assigned higher weights to preserve sharp features and edges.
- **Boundary Preservation:** Boundary vertices are assigned increased weights to prevent distortion of mesh boundaries.
- **Interaction Regions:** Regions that are likely to be involved in contact, such as object surfaces and interaction zones, are assigned higher importance.

These weights can be computed based on local geometric properties or predefined interaction annotations.

5.3.4 Edge Collapse Cost

Given two vertices with quadrics $\mathbf{Q}_1$ and $\mathbf{Q}_2$, the cost of collapsing the edge is computed using the combined quadric:

$$\mathbf{Q}_{combined} = \mathbf{Q}_1 + \mathbf{Q}_2. \tag{5.7}$$

The optimal position $\mathbf{v}^*$ of the new vertex is obtained by minimizing the error:

$$\mathbf{v}^* = \arg\min_{\mathbf{v}} \mathbf{v}^T \mathbf{Q}_{combined} \mathbf{v}. \tag{5.8}$$

The introduction of feature-aware weights influences both the cost computation and the optimal vertex placement, ensuring that important regions are preserved.

5.3.5 Constraint Handling

In addition to weighting, we incorporate constraints to further preserve critical features:

- **Boundary Constraints:** Edge collapses that would alter mesh boundaries are restricted or penalized.
- **Normal Consistency:** Changes in surface normals are monitored to avoid significant distortion of local geometry.
- **Topology Preservation:** Invalid collapses that would alter mesh topology are prevented.

These constraints ensure that the simplification process maintains structural and geometric integrity.

5.3.6 Algorithm Overview

The overall simplification process follows an iterative edge collapse strategy:

1. Initialize quadrics for all vertices using the feature-aware formulation.
2. Compute edge collapse costs for all edges.
3. Insert edges into a priority queue based on cost.
4. Iteratively collapse the edge with the lowest cost while satisfying constraints.
5. Update quadrics and costs for affected vertices and edges.

This process continues until the desired level of simplification is achieved.

Algorithm 2 Feature-Aware QEM Mesh Simplification

Input: Mesh $\mathcal{M}$, target reduction ratio r
Output: Simplified mesh $\mathcal{M}'$
Initialize quadrics $\mathbf{Q}_v$ for all vertices using feature-aware weighting
Compute edge collapse costs for all edges
Insert edges into priority queue $\mathcal{Q}$
while *mesh size* $> r \cdot |\mathcal{M}|$ **do**
    Extract edge e with minimum cost from $\mathcal{Q}$
    Compute optimal collapse position $\mathbf{v}^*$
    if *collapse satisfies constraints* **then**
        Collapse edge e to $\mathbf{v}^*$ Update local mesh connectivity Recompute quadrics for affected vertices
        Update edge costs in $\mathcal{Q}$
return $\mathcal{M}'$

5.3.7 Discussion

The proposed feature-aware QEM framework extends the classical approach by incorporating geometric and interaction-based cues into the simplification process. By prioritizing important regions and enforcing constraints, the method preserves features that are critical for realistic human-object interaction.

Compared to standard QEM, our approach produces simplified meshes that maintain both visual fidelity and interaction quality, enabling efficient yet accurate representation of 3D scenes.

5.4 Implementation Details

We implement the proposed feature-aware mesh simplification framework using an edge-collapse based pipeline built upon the classical QEM formulation. This section describes the key components of the implementation, including data structures, quadric computation, and optimization strategies.

5.4.1 Mesh Representation

The input mesh is represented as a triangular mesh consisting of vertices, edges, and faces. We maintain adjacency information for efficient traversal and updates during the simplification process. Each vertex stores its associated quadric matrix, while edges maintain collapse costs and connectivity information.

To efficiently support edge collapses and neighborhood updates, we use a half-edge or adjacency-list based representation, which allows constant-time access to neighboring vertices and faces.

5.4.2 Quadric Initialization

For each face in the mesh, we compute the plane equation and construct the corresponding quadric matrix. These face quadrics are accumulated at each vertex to initialize the vertex quadrics.

The feature-aware weighting described in Section 5.4 is incorporated during this stage. Specifically, weights based on curvature, boundary information, and interaction relevance are applied when accumulating face quadrics.

Curvature is estimated using local geometric properties, such as dihedral angles between adjacent faces. Boundary vertices are detected based on mesh connectivity, and interaction-relevant regions are identified using object annotations or proximity-based heuristics.

5.4.3 Edge Priority Queue

All edges in the mesh are inserted into a priority queue based on their collapse cost. The cost of collapsing an edge is computed using the combined quadric of its endpoints, as described in Section 5.4.

We use a min-heap data structure to efficiently extract the edge with the lowest collapse cost at each iteration. After each collapse, the costs of affected edges are updated and reinserted into the queue.

5.4.4 Edge Collapse Operation

During each iteration, the edge with the lowest cost is selected for collapse. The new vertex position is computed by minimizing the quadric error. If a valid solution cannot be obtained (e.g., due to a non-invertible matrix), a fallback strategy such as selecting one of the original vertices or the midpoint of the edge is used.

Before performing the collapse, we verify that it satisfies the constraints described in Section 5.4, including boundary preservation, normal consistency, and topology constraints. Invalid collapses are skipped to maintain mesh integrity.

5.4.5 Update Strategy

After an edge collapse, the affected region of the mesh is updated. This includes:

- Updating vertex positions and connectivity
- Recomputing quadrics for affected vertices
- Updating edge collapse costs in the priority queue

To improve efficiency, updates are restricted to the local neighborhood of the collapsed edge rather than the entire mesh.

5.4.6 Stopping Criteria

The simplification process continues until a predefined target is reached. This target can be specified in terms of:

- A desired number of vertices or faces
- A target reduction ratio
- A maximum allowable error threshold

In our experiments, we primarily use a target face count or reduction ratio to control the level of simplification.

5.4.7 Implementation Details and Efficiency

The algorithm is implemented in a modular manner, allowing integration with existing geometry processing pipelines. Efficient data structures and localized updates ensure that the method scales to large meshes.

In practice, the feature-aware weighting introduces minimal overhead compared to standard QEM, while providing significant improvements in preserving interaction-relevant features.

5.4.8 Discussion

The implementation is designed to balance efficiency and robustness. By combining feature-aware weighting with an optimized edge-collapse pipeline, the method achieves scalable mesh simplification while preserving the geometric properties necessary for interaction-aware systems.

5.5 Experiments and Results

We evaluate the proposed feature-aware mesh simplification framework on a variety of 3D scenes and object meshes. The goal of these experiments is to assess the ability of the method to preserve geometric features and interaction-relevant regions while achieving significant reduction in mesh complexity.

5.5.1 Experimental Setup

Datasets: We conduct experiments on meshes derived from 3D scenes and object datasets used in previous chapters. These include complex indoor environments and object models with varying geometric detail.

Baselines: We compare our method with standard quadric error metrics (QEM) based simplification, which serves as a widely used baseline for mesh decimation.

Evaluation Metrics: To evaluate performance, we consider the following metrics:

- **Geometric Error:** Measures the deviation between the original and simplified mesh.
- **Normal Consistency:** Evaluates preservation of surface orientation.
- **Feature Preservation Score:** Quantifies how well sharp edges and boundaries are retained.
- **Reduction Ratio:** Indicates the level of simplification achieved.
- **Runtime:** Measures computational efficiency.

5.5.2 Quantitative Results

Table 5.1 summarizes the quantitative comparison between standard QEM and the proposed feature-aware method. Our approach achieves comparable geometric error while significantly improving feature preservation and normal consistency.

In particular, we observe that the proposed method maintains higher fidelity in high-curvature regions and along boundaries, which are critical for interaction-aware applications. The runtime overhead introduced by feature-aware weighting is minimal, demonstrating the efficiency of the approach.

Method	**10% Resolution**			**1% Resolution**	
	$Hausdorff(\times 10^{-1}) \downarrow$	$Chamfer(\times 10^{-1}) \downarrow$	$Time(s) \downarrow$	$Hausdorff(\times 10^{-1}) \downarrow$	$Chamfer(\times 10^{-1}) \downarrow$
QEM [7]	0.13	0.4600	**4.25**	0.57	1.3700
QEM4VR [1]	0.78	<u>0.0136</u>	29.01	2.75	<u>0.1386</u>
ICE [19]	2.71	0.2815	16.55	4.25	0.8254
CWF [35]	0.66	0.0189	21.77	-	-
Liu et. al [20]	<u>0.10</u>	0.2900	37.70	<u>0.44</u>	1.1100
FA-QEM	**0.08**	**0.0072**	<u>10.60</u>	**0.25**	**0.0173**

Table 5.1 Comparison of mesh simplification methods.

5.5.3 Qualitative Results

Figure 5.2 presents visual comparisons between meshes simplified using standard QEM and our method. While both methods achieve similar levels of reduction, the standard QEM approach tends to smooth out sharp features and distort boundaries.

In contrast, our method preserves important geometric details, including edges and interaction-relevant surfaces. This results in meshes that are more suitable for downstream tasks such as collision detection and human-object interaction modeling.

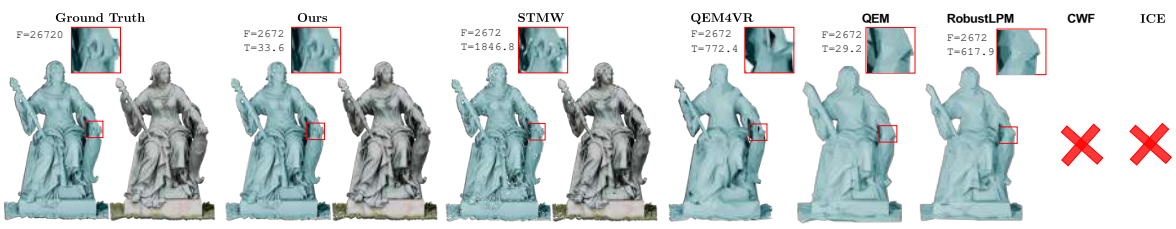

Figure 5.2 Comparison of standard QEM and feature-aware simplification.

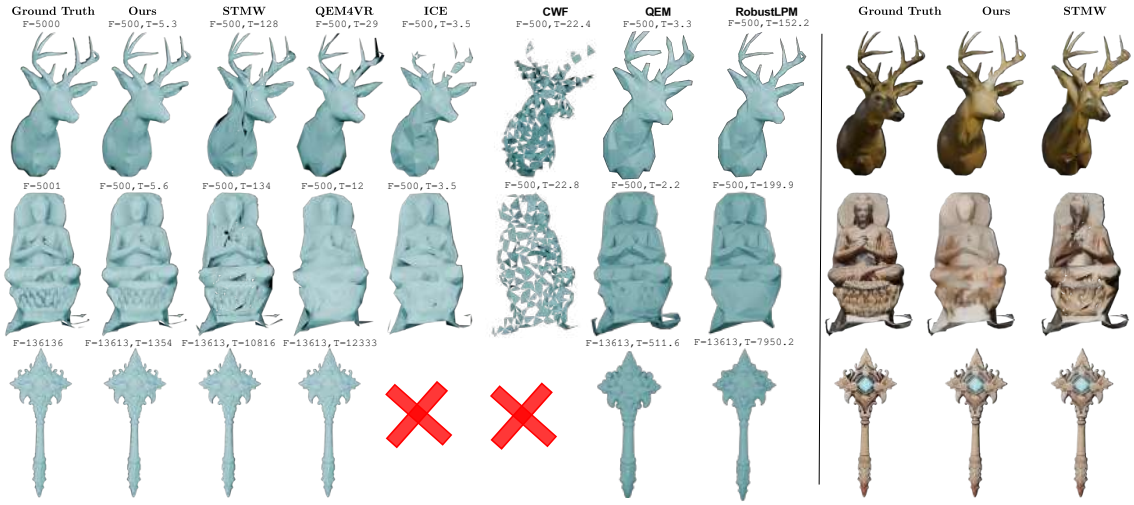

Figure 5.3 Comparison between standard QEM and the proposed feature-aware QEM. Standard QEM smooths out sharp features and distorts boundaries, whereas our method preserves high-curvature regions and interaction-relevant surfaces, resulting in improved geometric fidelity.

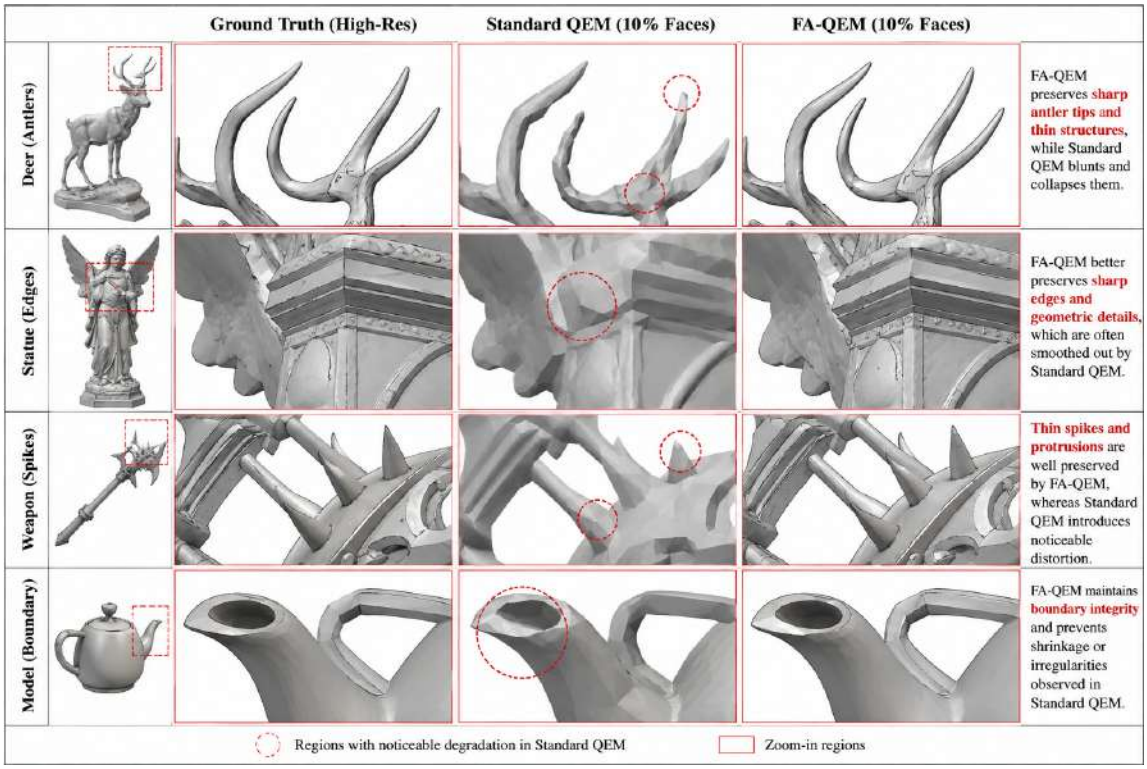

Figure 5.4 Close-up comparison of feature preservation under aggressive mesh simplification. Standard QEM tends to smooth sharp edges, collapse thin structures, and distort boundary regions. In contrast, the proposed FA-QEM preserves high-curvature features, fine geometric structures, and interaction-relevant regions due to the incorporation of curvature-aware, boundary-preserving, and normal-consistency constraints into the simplification process.

The close-up comparisons in Figure 5.4 highlight the importance of preserving high-curvature and boundary regions during simplification. Standard QEM often produces excessive smoothing around sharp geometric structures, leading to loss of fine details and distortion of interaction-relevant surfaces. In contrast, the proposed feature-aware formulation maintains edge sharpness, thin structures, and boundary integrity even under aggressive simplification ratios.

These improvements are particularly important for downstream applications involving collision detection, contact reasoning, and physically grounded human-object interaction, where geometric accuracy near contact regions directly influences interaction quality.

5.5.4 Ablation Study

To evaluate the contribution of individual components, we perform an ablation study as shown in Table 5.2:

Without Curvature Weighting: Removing curvature-based weights leads to loss of sharp features and increased smoothing.

Without Boundary Constraints: Disabling boundary preservation results in distortion along mesh boundaries.

Without Interaction-Aware Weighting: Ignoring interaction-relevant regions reduces the quality of contact surfaces, which can negatively impact interaction modeling.

These results highlight the importance of incorporating feature-aware cues into the simplification process.

Configuration	$Geometric\ Hausdorff$ $(\times 10^{-1})(\downarrow)$	$Geometric\ Chamfer$ $(\times 10^{-1})(\downarrow)$	$Texture\ Chamfer$ $(\times 10^{-1})(\downarrow)$
(1) FA-QEM (w/ all w's=0)	0.1970	0.0267	0.1141
(2) + w_{area}	0.1404	0.0173	0.1115
(3) + w_{plane_area}	0.1230	0.0135	0.1109
(4) + $w_{boundary}$	0.0910	0.0084	0.1035
(5) Full FA-QEM (+ w_{normal})	**0.0800**	**0.0072**	**0.0990**

Table 5.2 Ablation study of FA-QEM components. Each row adds one term of our composite metric to the previous configuration. We report geometric (Hausdorff, Chamfer) and texture-space Chamfer errors at 10% resolution. Each component yields measurable gains, highlighting the effectiveness of the full FA-QEM formulation.

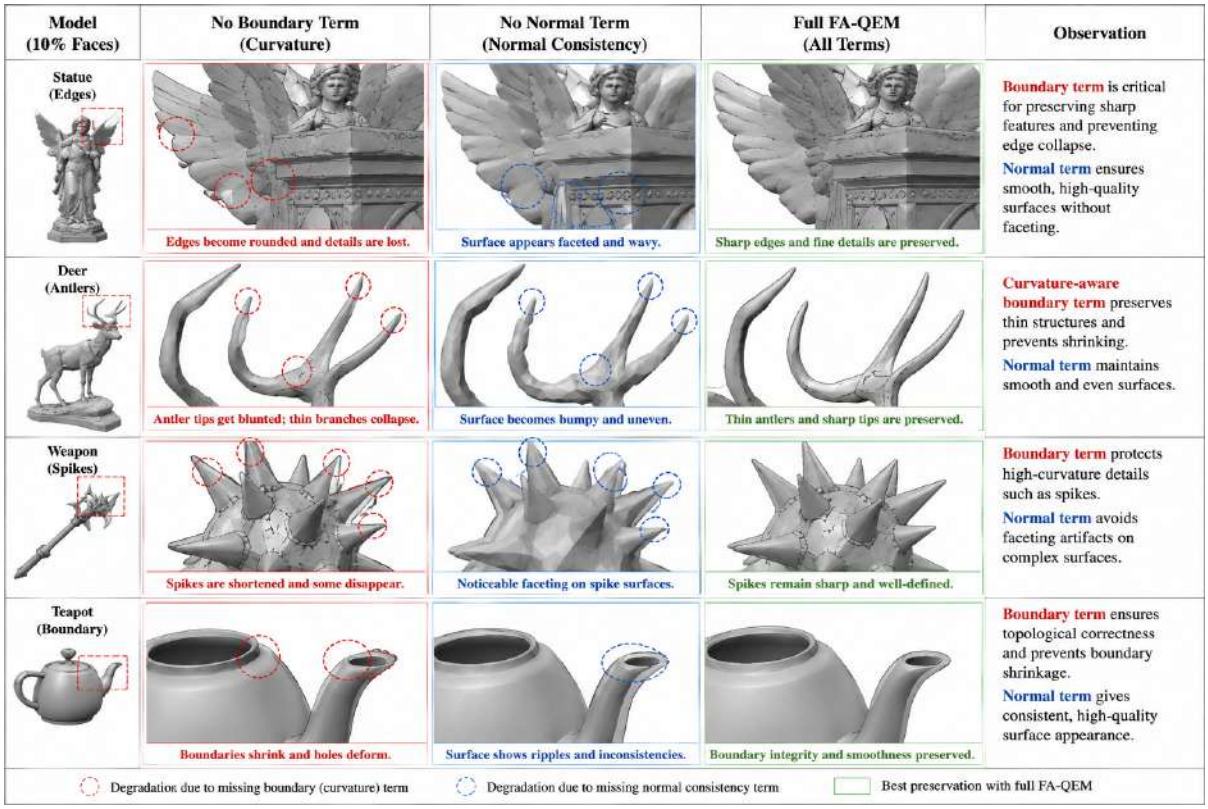

Figure 5.5 Visual ablation study of FA-QEM components. Removing curvature-aware boundary preservation causes feature flattening near sharp edges, while removing normal preservation introduces surface smoothing artifacts. The full FA-QEM formulation preserves both geometric detail and boundary structure.

5.5.5 Efficiency Analysis

We analyze the computational efficiency of the proposed method by measuring runtime across meshes of varying sizes. The results show that the method scales well with mesh complexity and introduces only a modest overhead compared to standard QEM.

This demonstrates that feature-aware simplification can be applied in practical settings without significantly compromising performance.

5.5.6 Analysis of Results

The experimental results provide strong evidence that the proposed feature-aware QEM framework effectively balances geometric fidelity and computational efficiency. By preserving features critical for interaction, the method enables the use of simplified meshes in interaction-aware systems without degrading performance.

Compared to standard QEM, our approach provides improved visual quality and better preservation of interaction-relevant regions, making it well-suited for applications involving human-object interaction.

5.6 Discussion

The results presented in this chapter demonstrate that incorporating feature-aware cues into the mesh simplification process significantly improves the preservation of geometry relevant for human-object interaction. By extending the classical QEM framework with feature weighting and constraint mechanisms, the proposed method achieves a balance between geometric fidelity and computational efficiency.

A key strength of the approach lies in its ability to preserve interaction-critical regions, such as boundaries, high-curvature areas, and contact surfaces. These regions play an important role in tasks such as collision detection and interaction modeling, and their preservation directly impacts the quality of downstream applications.

In contrast to standard QEM, which treats all regions uniformly, the proposed method introduces a notion of importance that allows the simplification process to prioritize structurally and functionally significant regions. This leads to improved visual quality and better suitability for interaction-aware systems.

Another advantage of the approach is its compatibility with existing geometry processing pipelines. The method builds upon the well-established QEM framework and introduces minimal additional computational overhead, making it practical for large-scale applications.

However, the method also has certain limitations. First, the effectiveness of feature-aware weighting depends on the accurate estimation of geometric properties such as curvature and boundary information. Errors in these estimates can affect the quality of the simplified mesh.

Second, the identification of interaction-relevant regions may require additional annotations or heuristics, which may not always be readily available. Developing more robust and automated methods for identifying such regions is an important direction for future work.

Additionally, while the method improves feature preservation, it does not explicitly model appearance attributes such as textures or materials. Incorporating appearance-aware constraints could further enhance visual fidelity.

Finally, although the computational overhead is modest, further optimization may be required for real-time applications involving extremely large scenes.

Despite these limitations, the proposed feature-aware QEM framework provides a practical and effective solution for geometry simplification in interaction-aware systems. It bridges the gap between high-fidelity geometric modeling and efficient computation, enabling scalable deployment of the methods developed in earlier chapters.

5.7 Summary

In this chapter, we addressed the challenge of efficiently processing high-resolution 3D geometry for interaction-aware systems. We highlighted the limitations of standard mesh simplification techniques, particularly in preserving features that are critical for human-object interaction.

To overcome these limitations, we proposed a feature-aware extension of the quadric error metrics framework. By incorporating geometric and interaction-based cues into the simplification process, the method preserves important features such as boundaries, sharp edges, and contact regions while achieving significant reduction in mesh complexity.

Through quantitative and qualitative evaluations, we demonstrated that the proposed approach outperforms standard QEM in terms of feature preservation and interaction fidelity, while maintaining comparable efficiency.

Overall, this chapter complements the interaction modeling contributions of the thesis by providing a systems-level solution for scalable geometry processing. Together with the methods presented in previous chapters, it contributes to a unified framework for realistic and efficient human-object interaction in 3D environments.

Chapter 6

Conditional Object Interaction Modeling

Recent advances in generative modeling have enabled the synthesis of human motion conditioned on high-level inputs such as text and speech [54, 9]. In collaborative work presented in InteracTalker [44], a unified framework was proposed for generating full-body human motion conditioned on multimodal inputs, including speech, text, and object interaction.

While the full framework addresses multimodal motion generation, this chapter focuses specifically on the object interaction component, which aligns with the broader theme of this thesis.

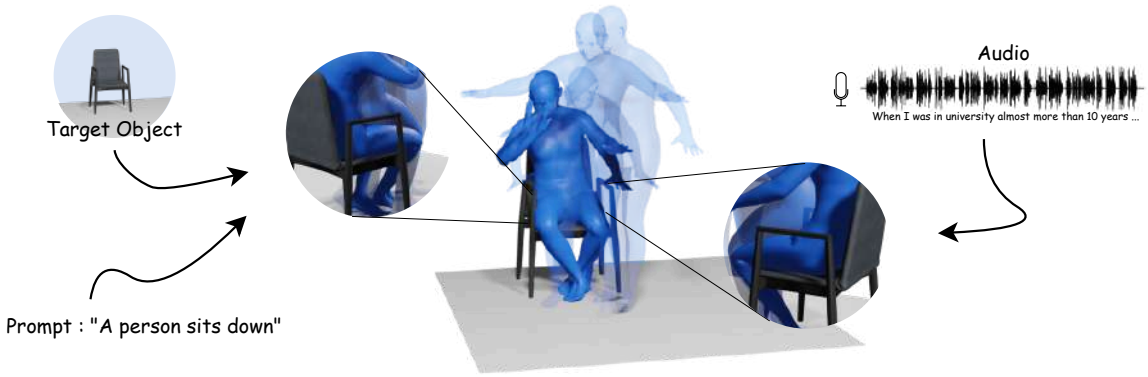

Figure 6.1 InteracTalker generates realistic, full-body human motion, precisely integrating object interaction with expressive co-speech gestures, conditioned by speech audio, text prompts, and a target object.

6.1 Object Interaction in Conditional Motion Generation

A key challenge in conditional motion generation is ensuring that generated motion remains consistent with object interaction. Naively combining motion generation with object conditioning often leads to physically implausible results, such as interpenetration, unstable grasps, or inconsistent contact.

To address this, the framework incorporates an interaction-aware modeling component that explicitly conditions motion generation on object geometry and interaction constraints. This enables the generation of motion sequences that are not only responsive to high-level inputs but also physically grounded.

A central insight is that object interaction must be integrated into the generation process itself, rather than treated as a post-processing step. This parallels the design principles adopted in Chapter 4, where interaction constraints are incorporated directly into the learning framework.

6.2 Interaction-Aware Modeling

The object interaction component introduces mechanisms to ensure consistency between the human body and the manipulated object. This includes:

- **Object-Conditioned Motion:** Motion generation is conditioned on object geometry and spatial configuration, enabling context-aware interaction.
- **Contact Consistency:** The model encourages stable and continuous contact between the human hand and the object throughout the motion sequence.
- **Physical Plausibility:** Constraints are imposed to reduce interpenetration and ensure realistic interaction with the environment.

By integrating these components, the model generates motion that maintains coherence between high-level intent and low-level physical interaction.

6.3 Discussion

This work highlights the importance of incorporating interaction constraints into conditional motion generation frameworks. While multimodal inputs provide rich context for motion synthesis, ensuring consistency with object interaction remains critical for realism.

The insights from this work reinforce the central theme of this thesis: realistic human-object interaction requires joint modeling of motion, interaction, and environmental constraints. Even in the presence of high-level conditioning signals, interaction-aware modeling remains essential.

6.4 Summary

In this chapter, we briefly discussed the object interaction component of a multimodal motion generation framework. By explicitly incorporating object-aware constraints into the generation process, the approach improves the physical plausibility and consistency of generated motion.

This serves as an extension of the interaction modeling framework developed in earlier chapters, demonstrating its applicability in conditional generation settings.

Chapter 7

Conclusions

This thesis addressed the fundamental challenge of modeling realistic human-object interaction in complex 3D environments. While significant progress has been made in individual domains such as human motion generation, object interaction, and scene understanding, existing approaches largely treat these components in isolation. This fragmentation limits their ability to generate physically plausible, semantically meaningful, and context-aware behavior.

To overcome these limitations, this thesis adopted a unified perspective on interaction-aware modeling that spans real-world systems, data-driven learning, and efficient geometry processing. By integrating these components, the work moves towards a more holistic and practically deployable understanding of human behavior in 3D environments.

Summary of Contributions.

- **Real-World Interaction System (Patent):** We introduced a system for real-time human pose understanding and interaction analysis based on joint-level reasoning. This system demonstrates how interaction-aware modeling can be translated into practical, deployable solutions, bridging the gap between theoretical methods and real-world applications.

- **Scene-Aware Human-Object Interaction:** We introduced a framework for generating human motion that jointly models full-body dynamics, object interaction, and scene constraints. Supported by the MOGRAS dataset, this approach enables the synthesis of physically plausible and interaction-consistent motion in realistic environments.

- **Efficient Geometry Processing:** We proposed a feature-aware extension of quadric error metrics for mesh simplification, enabling efficient processing of high-resolution 3D scenes while preserving features critical for interaction. This contribution bridges the gap between high-fidelity modeling and scalable deployment.

- **Extension to Conditional Interaction:** We explored extensions to interaction-aware motion generation in conditional settings, demonstrating how object interaction can be incorporated into generative frameworks guided by high-level inputs, while maintaining physical consistency.

- **Unified Perspective:** Beyond individual components, this thesis provides a unified perspective on human-object interaction that integrates real-world understanding, motion generation, and geometry processing within a single framework.

Limitations and Future Work. While the proposed framework makes significant progress towards realistic interaction modeling, several challenges remain. Current methods rely on high-quality 3D scene representations and annotated data, which may not always be available in real-world settings. Additionally, the modeling of long-horizon interactions and complex manipulation tasks remains an open problem.

Future work can explore the integration of physics-based reasoning, improved interaction modeling under partial observations, and more scalable data-driven approaches. Bridging the gap between real-world perception systems and interaction-aware generative models also represents a promising direction. Extending these methods to robotics and embodied agents can further enable physically grounded interaction in real-world environments.

Final Remarks. This thesis takes a step towards bridging the gap between perception, action, and interaction in 3D environments. By combining real-world systems, interaction-aware learning, and efficient geometry processing, it contributes to the development of models capable of generating realistic and context-aware human behavior.

At its core, this thesis demonstrates that realistic human-object interaction emerges not from isolated advances, but from the principled integration of perception, motion, interaction, and geometry across both simulated and real-world settings.

We believe that the ideas presented in this work will serve as a foundation for future research in human-centric AI, enabling applications in virtual reality, robotics, and human-computer interaction. Ultimately, achieving seamless interaction between humans and digital environments requires systems that are not only accurate and efficient but also grounded in the physical and semantic structure of the real world.

Related Publications

Thesis Publications

- Charu Sharma, **Kunal Bhosikar**, Mahika Jain, Vanshita Mahajan; *System and Method for Detecting Exercise Poses and Counting Repetitions Based on Joint Angles*; **Indian Patent Published, 2025**.

- **Kunal Bhosikar**, Siddharth Katageri, Vivek Madhavaram, Kai Han, Charu Sharma; *MOGRAS: Human Motion with Grasping in 3D Scenes*; **From Scene Understanding to Human Modeling (SU2HM) Workshop at The British Machine Vision Conference (BMVC), 2025 (Oral Presentation)**.

- **Kunal Bhosikar**, Preet Savalia, Lokender Tiwari, Brojeshwar Bhowmick; *FA-QEM: Fast and Robust Mesh Simplification with Geometry and Appearance Preservation*; **The IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop on 3D Geometry Generation for Scientific Computing (3D4S) 2026 (Best Paper Award Runner-up)**.

Other Publications

- Sreehari Rajan*, **Kunal Bhosikar***, Charu Sharma; *InteracTalker: Prompt-Based Human-Object Interaction with Co-Speech Gesture Generation*; **The IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2026 (Oral Presentation)**.

- Prajas Wadekar, Pranav Bachina*, **Kunal Bhosikar***, Ankit Gangwal, Charu Sharma; *PatchPoison: Poisoning Multi-View Datasets to Degrade 3D Reconstruction*; **The IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop on Security, Privacy, and Adversarial Robustness in 3D Generative Vision Models (SPAR-3D), 2026**.

Under Review Manuscripts

- Sreehari Rajan, Kunal Bhosikar, Charu Sharma, Nikos Athanasiou; *ENACT: Single-Image Human-Scene Interaction Motion from Language via Foundation-Model Orchestration*; **Under Review NeurIPS, 2026.**

Bibliography

[1] N. Athanasiou, M. Petrovich, M. J. Black, and G. Varol. Teach: Temporal action compositions for 3d humans. In *International Conference on 3D Vision (3DV)*, September 2022.

[2] K. Bahirat, C. Lai, R. McMahan, and B. Prabhakaran. Designing and evaluating a mesh simplification algorithm for virtual reality. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14:1–26, 06 2018.

[3] E. Barsoum, J. Kender, and Z. Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[4] B. L. Bhatnagar, X. Xie, I. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2022.

[5] K. Bhosikar, S. Katageri, V. Madhavaram, K. Han, and C. Sharma. Mogras: Human motion with grasping in 3d scenes. In *Proceedings of the 36th British Machine Vision Conference (BMVC) Workshops*. British Machine Vision Association, 2025.

[6] Z. Cao, H. Gao, K. Mangalam, Q.-Z. Cai, M. Vo, and J. Malik. Long-term human motion prediction with scene context. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 387–404, Berlin, Heidelberg, 2020. Springer-Verlag.

[7] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43(01):172–186, Jan. 2021.

[8] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[9] R. Daněček, K. Chhatre, S. Tripathi, Y. Wen, M. Black, and T. Bolkart. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*, page 1–13. ACM, Dec. 2023.

[10] A. Delitzas, A. Takmaz, F. Tombari, R. Sumner, M. Pollefeys, and F. Engelmann. Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14531–14542, June 2024.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[12] M. Garland and P. S. Heckbert. Surface simplification using quadric error metrics. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '97, page 209–216, USA, 1997. ACM Press/Addison-Wesley Publishing Co.

[13] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. Learning individual styles of conversational gesture. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019.

[14] P. Grady, C. Tang, C. D. Twigg, M. Vo, S. Brahmbhatt, and C. C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[15] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng. 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, June 2022.

[16] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision*, Oct. 2019.

[17] M. Hassan, P. Ghosh, J. Tesch, D. Tzionas, and M. J. Black. Populating 3D scenes by learning human-scene interaction. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[18] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019.

[19] D. Holden, J. Saito, and T. Komura. A deep learning framework for character motion synthesis and editing. *ACM Trans. Graph.*, 35(4), July 2016.

[20] S. Huang, Z. Wang, P. Li, B. Jia, T. Liu, Y. Zhu, W. Liang, and S.-C. Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[21] J. S. J. Y. Ikhsanul Habibie, Daniel Holden and T. Komura. A recurrent variational autoencoder for human motion synthesis. In G. B. Tae-Kyun Kim, Stefanos Zafeiriou and K. Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 119.1–119.12. BMVA Press, September 2017.

[22] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] H. Jiang, S. Liu, J. Wang, and X. Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the International Conference on Computer Vision*, 2021.

[24] N. Jiang, Z. Zhang, H. Li, X. Ma, Z. Wang, Y. Chen, T. Liu, Y. Zhu, and S. Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024.

[25] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, SGP '06, page 61–70, Goslar, DEU, 2006. Eurographics Association.

[26] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023.

[27] H. Liu, Z. Zhu, G. Becherini, Y. Peng, M. Su, Y. Zhou, X. Zhe, N. Iwamoto, B. Zheng, and M. J. Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1144–1154, June 2024.

[28] H. Liu, Z. Zhu, N. Iwamoto, Y. Peng, Z. Li, Y. Zhou, E. Bozkurt, and B. Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, page 612–630, Berlin, Heidelberg, 2022. Springer-Verlag.

[29] H.-T. D. Liu, X. Zhang, and C. Yuksel. Simplifying textured triangle meshes in the wild. *ACM Trans. Graph.*, 44(6), Dec. 2025.

[30] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), Nov. 2015.

[31] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*, 2019.

[32] Z. Luo, J. Cao, S. Christen, A. Winkler, K. M. Kitani, and W. Xu. Omnigrasp: Simulated humanoid grasping on diverse objects. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[33] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. Amass: Archive of motion capture as surface shapes. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019.

[34] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[35] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, Dec. 2021.

[36] A. T. Miller and P. K. Allen. Graspit!: A versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004.

[37] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4), July 2022.

[38] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

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

[40] M. Petrovich, M. J. Black, and G. Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *International Conference on Computer Vision (ICCV)*, 2021.

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

[42] S. Prokudin, C. Lassner, and J. Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[43] A. R. Punnakkal, A. Chandrasekaran, N. Athanasiou, A. Quiros-Ramirez, and M. J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, June 2021.

[44] S. Rajan, K. Bhosikar, and C. Sharma. Interactalker: Prompt-based human-object interaction with co-speech gesture generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.

[45] M. Savva, A. X. Chang, P. Hanrahan, M. Fisher, and M. Nießner. Scenegrok: inferring action maps in 3d environments. *ACM Trans. Graph.*, 33(6), Nov. 2014.

[46] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra. Habitat: A platform for embodied AI research. *CoRR*, abs/1904.01201, 2019.

[47] W. J. Schroeder, J. A. Zarge, and W. E. Lorensen. Decimation of triangle meshes. In *Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '92, page 65–70, New York, NY, USA, 1992. Association for Computing Machinery.

[48] Y. Shafir, G. Tevet, R. Kapon, and A. H. Bermano. Human motion diffusion as a generative prior. In *The Twelfth International Conference on Learning Representations*, 2024.

[49] C. Sharma, K. Bhosikar, M. Jain, and V. Mahajan. System and method for detecting exercise poses and counting repetitions based on joint angles. Indian Patent Published, Application No. 202541010885, 2025.

[50] O. Taheri, V. Choutas, M. J. Black, and D. Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[51] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas. *GRAB: A Dataset of Whole-Body Human Grasping of Objects*, page 581–600. Springer International Publishing, 2020.

[52] P. Tendulkar, D. Surís, and C. Vondrick. Flex: Full-body grasping without full-body grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21179–21189, June 2023.

[53] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or. Motionclip: Exposing human motion generation tonbsp;clip space. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, page 358–374, Berlin, Heidelberg, 2022. Springer-Verlag.

[54] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

[55] J. Wang, J. Hodgins, and J. Won. Strategy and skill learning for physics-based table tennis animation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

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

[57] Y. Wu, J. Wang, Y. Zhang, S. Zhang, O. Hilliges, F. Yu, and S. Tang. Saga: Stochastic whole-body grasping with contact. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[58] Z. Xiao, T. Wang, J. Wang, J. Cao, W. Zhang, B. Dai, D. Lin, and J. Pang. Unified human-scene interaction via prompted chain-of-contacts. In *The Twelfth International Conference on Learning Representations*, 2024.

[59] S. Xu, Z. Li, Y.-X. Wang, and L.-Y. Gui. InterDiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023.

[60] Z. Xu, Y. Lin, H. Han, S. Yang, R. Li, Y. Zhang, and X. Li. Mambatalk: Efficient holistic gesture synthesis with selective state space models. *Advances in Neural Information Processing Systems*, 37:20055–20080, 2024.

[61] L. Yang, K. Li, X. Zhan, F. Wu, A. Xu, L. Liu, and C. Lu. OakInk: A large-scale knowledge repository for understanding hand-object interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[62] S. Yang, Z. Xu, H. Xue, Y. Cheng, S. Huang, M. Gong, and Z. Wu. Freetalker: Controllable speech and text-driven gesture generation based on diffusion models for enhanced speaker naturalness. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[63] S. Ye, Y. Wang, J. Li, D. Park, C. K. Liu, H. Xu, and J. Wu. Scene synthesis from human motion. In *SIGGRAPH Asia 2022 Conference Papers*, SA '22, New York, NY, USA, 2022. Association for Computing Machinery.

[64] H. Yi, J. Thies, M. J. Black, X. B. Peng, and D. Rempe. Generating human interaction motions in scenes with text control. In *ECCV*, 2024.

[65] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(6):4115–4128, June 2024.

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

[67] Y. Zhang, Q. He, Y. Wan, Y. Zhang, X. Deng, C. Ma, and H. Wang. Diffgrasp: Whole-body grasping synthesis guided by object motion using a diffusion model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[68] Y. Zhu, N. Samet, and D. Picard. H3wb: Human3.6m 3d wholebody dataset and benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20166–20177, October 2023.