HOSIG: Full-Body Human-Object-Scene Interaction Generation with Hierarchical Scene Perception

Wei Yao[1], **Yunlian Sun**[1*], **Hongwen Zhang**[2], **Yebin Liu**[3], **Jinhui Tang**[4*]

[1]Nanjing University of Science and Technology
[2]Beijing Normal University
[3]Tsinghua University
[4]Nanjing Forestry University
{wei.yao, yunlian.sun}@njust.edu.cn, zhanghongwen@bnu.edu.cn, liuyebin@mail.tsinghua.edu.cn, tangjh@njfu.edu.cn

Abstract

Generating high-fidelity full-body human interactions with dynamic objects and static scenes remains a critical challenge in computer graphics and animation. Existing methods for human-object interaction often neglect scene context, leading to implausible penetrations, while human-scene interaction approaches struggle to coordinate fine-grained manipulations with long-range navigation. To address these limitations, we propose **HOSIG**, a novel framework for synthesizing full-body interactions through hierarchical scene perception. Our method decouples the task into three key components: 1) a *scene-aware grasp pose generator* that ensures collision-free whole-body postures with precise hand-object contact by integrating local geometry constraints, 2) a *heuristic navigation algorithm* that autonomously plans obstacle-avoiding paths in complex indoor environments via compressed 2D floor maps and dual-component spatial reasoning, and 3) a *scene-guided motion diffusion model* that generates trajectory-controlled, full-body motions with finger-level accuracy by incorporating spatial anchors and dual-space gradient-based guidance. Extensive experiments on the TRUMANS dataset demonstrate superior performance over state-of-the-art methods. Notably, our framework supports unlimited motion length through autoregressive generation and requires minimal manual intervention. This work bridges the critical gap between scene-aware navigation and dexterous object manipulation, advancing the frontier of embodied interaction synthesis.

Code — https://github.com/yw0208/HOSIG

Introduction

Embodied intelligence represents a pivotal frontier in AI research, aiming to develop agents capable of navigating and manipulating 3D environments. While existing works have extensively explored human-object interaction (HOI) (Taheri et al. 2024; Zhang et al. 2024b; Petrov et al. 2024; Wu et al. 2022; Zheng et al. 2023; Li, Wu, and Liu 2023) and human-scene interaction (HSI) (Karunratanakul et al. 2023; Xiao et al. 2023; Zhao et al. 2023; Zhang and Tang 2022a; Wang et al. 2024b), few address the integrated human-object-scene interaction (HOSI) challenge (Wu et al. 2024; Jiang et al. 2024b,a; Lu et al. 2024). As shown in Figure 1, our goal is to enable characters to move in complex scenes, complete precise object operations, and seamlessly connect them into a long-term motion. The inherent complexity of synthesizing these multimodal interactions presents critical unsolved challenges in spatial reasoning and motion coordination.

While recent advancements in HOI (Yang et al. 2024; Song et al. 2024; Peng et al. 2023; Taheri et al. 2022; Ghosh et al. 2023; Zhang et al. 2024a), they predominantly neglect 3D scene constraints. This oversight leads to implausible human-scene interpenetration, highlighting the necessity for scene-aware reasoning. Besides ignoring the scene in HOI, current HSI approaches exhibit two limitations: 1) Global scene encoding methods (Wang et al. 2024b, 2022b; Huang et al. 2023) lack granularity for precise motion synthesis, and 2) Local context perception strategies (Cen et al. 2024; Mao et al. 2022; Ghosh et al. 2021) struggle with pathfinding in complex environments, resulting in persistent interpenetration during long-range motion generation. These dual challenges underscore the need for unified scene-object-agent coordination.

Inspired by the above observations, we propose **HOSIG**, a hierarchical scene-aware framework that integrates three core components: (1) *Scene-aware grasp pose generator* for precise object manipulation with hand contact in 3D scenes, (2) *Collision-aware navigation planner* enabling obstacle-avoiding pathfinding in complex scenes, and (3) *Trajectory-controlled motion synthesizer* generating unrestricted-length whole-body motions through multi-modal condition integration. Our hierarchical architecture operates through three perception levels: *local* (object grasping positions), *global* (scene navigation topology), and *path-aligned* (continuous spatial guidance). Unlike prior works, HOSIG achieves unified coordination of dynamic object manipulation and static scene interaction while maintaining motion coherence through iterative refinement with spatial constraints.

Our technical implementation features three key innovations. First, the grasp pose generator augments the cVAE framework (Taheri et al. 2022) with local scene geometry constraints, producing physically-plausible hand-object orientations that prevent scene interpenetration. Second, a novel 2D scene abstraction layer enables efficient naviga-

[*] Corresponding author.

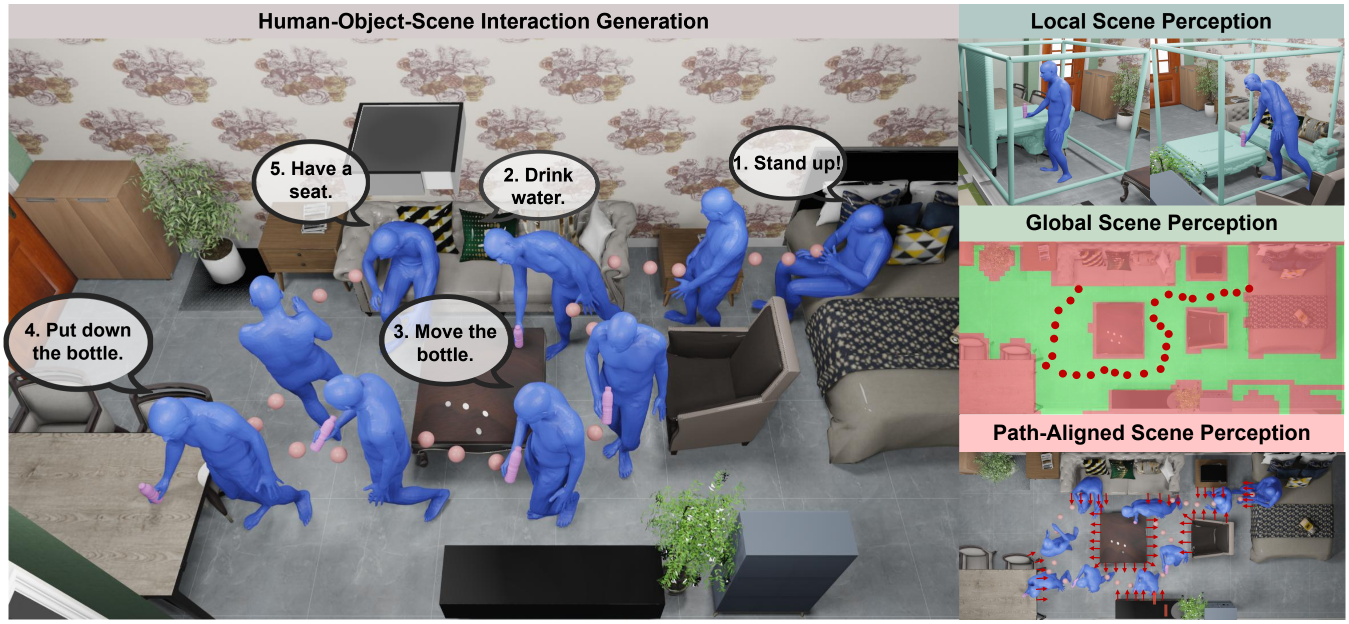

Figure 1: **Human-Object-Scene Interaction Generation**. Our proposed HOSIG can generate high-fidelity full-body human motions. HOSIG can not only generate **interactions with static scenes**, but also generate object manipulation motions with **fine hand-object contact**. Moreover, relying on iterative generation and autonomous navigation, HOISG can generate **long-term** motions in complex indoor scenes.

tion through 3D environments via heuristic pathfinding on compressed obstacle-aware maps, dynamically generating obstacle-aware motions. Third, building upon ControlNet's conditioning paradigm (Zhang, Rao, and Agrawala 2023), we develop a multi-condition diffusion framework that simultaneously integrates: (a) *spatial anchors* from navigation paths, (b) *fine-grained hand control* through grasp poses, and (c) *path-aligned scene priors* for continuous interaction optimization. This unified architecture achieves finger-level motion precision without auxiliary hand modules, surpassing previous trajectory-control methods through iterative spatial constraint.

In summary, our principal contributions are:

- A scene-geometry constrained grasp generator producing interpenetration-free full-body poses via cVAE augmentation.
- A 2D scene abstraction method with heuristic pathfinding for autonomous obstacle-aware navigation.
- A trajectory-language diffusion model integrating spatial anchors and scene guidance for finger-level motion control.
- An autoregressive HOSI pipeline achieving unlimited-length motion synthesis with full automation.

Related Work

Human-Object Interaction Generation

Research on human-object interaction (HOI) motion generation has progressed through two main paradigms. Early approaches primarily utilized reinforcement learning (RL), with initial works like (Peng et al. 2019, 2021) achieving basic interactions such as box touching. Subsequent RL methods developed more complex skills including basketball dribbling (Liu and Hodgins 2018), tennis playing (YUAN and Makoviychuk 2023), multi-object manipulation (Wang et al. 2023), and box carrying (Hassan et al. 2023).

Then generative models are employed for direct motion synthesis. GOAL (Taheri et al. 2022) pioneered this direction using cVAEs for grasping motions. IMoS (Ghosh et al. 2023) enabled action-label conditioned generation (e.g., photography, drinking). Recent works focus on language-conditioned generation: OOD-HOI (Zhang et al. 2024b) synthesizes human-object motions separately with contact optimization, while HOI-Diff (Peng et al. 2023) employs affordance-guided diffusion. Other key innovations include TriDi's unified modeling of bodies/objects/contacts (Petrov et al. 2024) and CHOIS's trajectory-constrained synthesis (Li et al. 2024b). InterDiff (Xu et al. 2023) further enables sequential generation from initial states.

Human-Scene Interaction Generation

Human-scene interaction (HSI) generation focuses on interactions with static environments, distinguished from HOI by the immobility of target objects. We classify scene interactions (e.g., with chairs/sofas) as HSI rather than HOI. Current HSI research bifurcates into two technical directions: (1) scene-aware locomotion and (2) semantic interaction synthesis.

For locomotion generation, existing methods adopt either data-driven (Wang et al. 2024a; Zhang and Tang 2022b; Rempe et al. 2023) or algorithmic approaches (Wang et al.

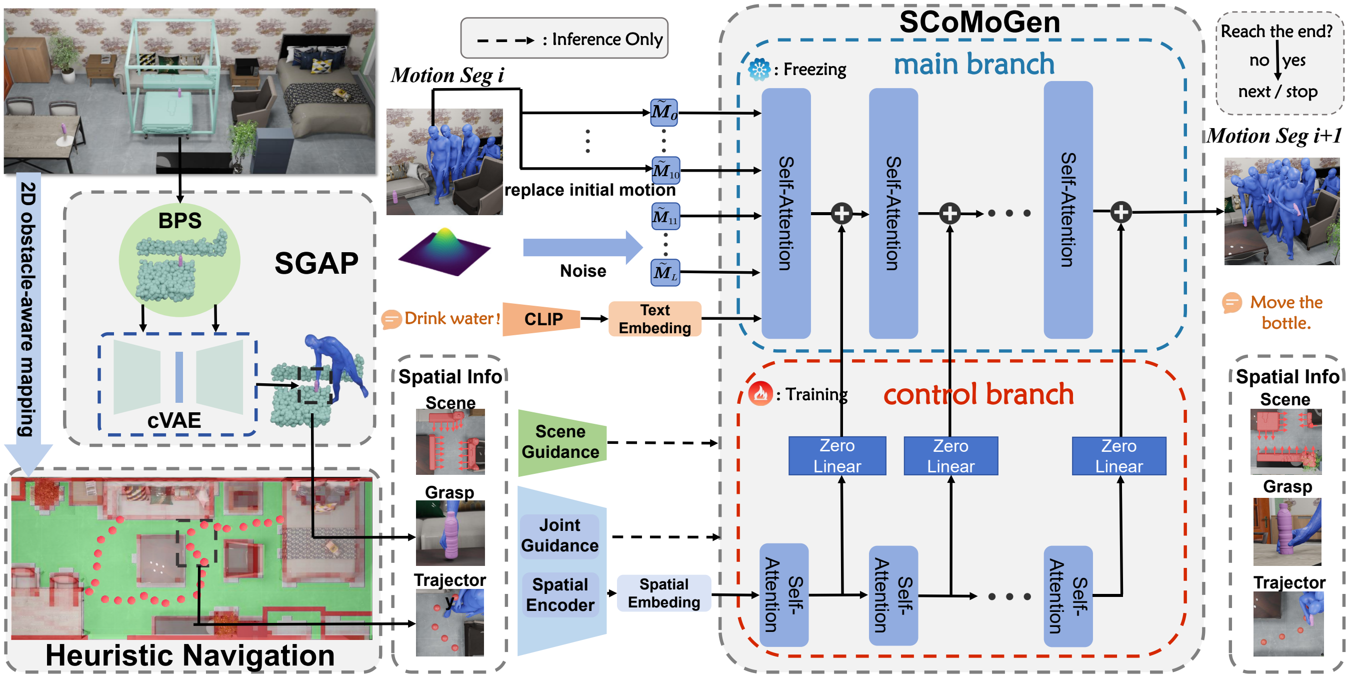

Figure 2: **Overview of Our Pipeline.** HOSIG can iteratively generate long-term motions based on spatial information, text, and the previous motion clip. There are three parts worth noting in the pipeline: (1) **SGAP** generates fine grasping postures to ensure the quality of character interaction. (2) **Heuristic Navigation** generates sparse human root joint trajectories to constrain the subsequent generated motions to be within the traversable area. (3) **SCoMoGen** uses a dual-branch design to achieve spatial control and adds additional joint & scene guidance during inference to achieve high-precision control.

2022a). While data-driven methods struggle in novel complex environments, we propose an algorithmic solution for robust navigation. Interaction synthesis initially produced static poses in scenes (Zhang et al. 2020; Li and Dai 2024; Zhao et al. 2022; Xuan et al. 2023), later evolving into language-guided systems (Wang et al. 2022b, 2024b; Huang et al. 2023) though requiring optimization for scene compliance. Recent advances include navigation-interaction frameworks (Yi et al. 2024) and video-generation-based methods (Li et al. 2024a) with strong generalization but limited contact realism and motion range.

Closest to our approach, (Jiang et al. 2024b,a) enable triadic human-object-scene interactions. However, these require frame-level action labels and precise finger positions, whereas our method achieves comparable results using only initial/final object states.

Method

Problem Formulation

We first formalize the task definition with five essential input parameters: (1) scene $\mathcal{S}$, (2) initial human position $\mathbf{p}_0 \in \mathbb{R}^3$, (3) object start pose $\mathbf{T}_s \in SE(3)$, (4) object target pose $\mathbf{T}_t \in SE(3)$, and (5) object mesh $\mathcal{O}$. Given these parameters, our HOSIG produces synchronized motion trajectories $\mathcal{M} = (\mathcal{M}_h, \mathcal{M}_o)$ containing both human motion $\mathcal{M}_h$ and object motion $\mathcal{M}_o$. A standard interaction sequence typically comprises three phases: initial approach (human navigation to $\mathbf{T}_s$), object manipulation (grasping and transportation from $\mathbf{T}_s$ to $\mathbf{T}_t$), and final placement (precise positioning at $\mathbf{T}_t$). Moreover, our framework supports extension through auxiliary interaction nodes to achieve functions such as sitting on a chair. Please refer to the supplementary materials for more applications.

As shown in Figure 2, the HOSIG framework implements its functionality through three interconnected components operating in a collaborative pipeline. The first module, **S**cene-Aware **G**r**a**sp **P**ose Generation (SGAP), computes feasible full-body grasp poses as anchors for other modules. Subsequently, our novel heuristic navigation algorithm with obstacle avoidance constraints computes collision-free 3D trajectories connecting the anchors. The final component, **S**cene-Guided **Co**ntrollable **Mo**tion **Gen**eration (SCoMoGen), synthesizes continuous motion sequences along trajectories. The subsequent sections elaborate on each module's technical formulation.

Scene-Aware Grasp Pose Generation

As shown in Figure 2, the SGAP module is implemented as a conditional variational autoencoder (cVAE) comprising an encoder-decoder architecture. The encoder's input tensor $X \in \mathbb{R}^d$ is formed through the concatenation operation:

$$X = \left[\Theta, \beta, \mathbf{V}, \mathbf{D}^{h\to o}, \hat{\mathbf{h}}, \mathbf{t}^o, \mathbf{B}^o, \mathbf{B}^s\right] \qquad (1)$$

where $\Theta \in \mathbb{R}^{3N_j}$ and $\beta \in \mathbb{R}^{10}$ denote the SMPL-X pose and shape parameters respectively, $\mathbf{V} \in \mathbb{R}^{400\times 3}$ repre-

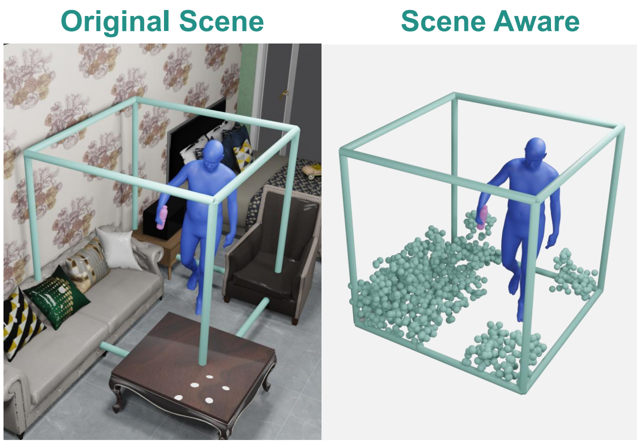

Figure 3: **Visulization of Scene-Aware in SGAP.** The green box is centered on the purple bottle. In SGAP, only the sparse scene point cloud inside the box is used, as shown in the right figure.

sents the sampled body mesh vertices, $\mathbf{D}^{h\rightarrow o} \in \mathbb{R}^{400\times 3}$ encodes vertex-wise directional offsets between the human body mesh and the nearest object surface points, $\hat{\mathbf{h}} \in \mathbb{R}^3$ specifies the head orientation unit vector, $\mathbf{t}^o \in \mathbb{R}^3$ captures the object's translational state, and $\mathbf{B}^o \in \mathbb{R}^{1024}$ denotes the Basis Point Set (BPS) encoding of the object geometry.

The critical component $\mathbf{B}^s$ constitutes the primary scene perception mechanism in SGAP. As illustrated in Figure 3, given a z-up object translation $\mathbf{t}^o = (x, y, z)$, we construct a scene context volume bounded by $[x - 0.8, x + 0.8] \times [y - 0.8, y + 0.8] \times [0.2, 1.8]$ meters, forming a 1.6^3 m^3 cubic region. To capture scene geometry, we employ a volumetric sampling strategy that generates a dense point cloud $\mathcal{P}_s$ containing both surface vertices and interior points from the scene mesh. This sampling follows a voxel grid resolution of 8 cm³, ensuring complete spatial coverage. The BPS transformation processes $\mathcal{P}_s$ through basis projections to produce the scene encoding $\mathbf{B}^s \in \mathbb{R}^{1024}$. These scene features condition the pose generation process, enabling synthesis of scene-adapted full-body poses.

To enforce geometric compatibility between human motions and environmental structures, we implement a physics-informed scene distance loss. The scene distance loss $\mathcal{L}_{sd}$ operates on the joint-space representation:

$$\mathcal{L}_{sd} = -\frac{1}{N}\sum_{j=1}^{22}\sum_{k=1}^{N}\|\mathbf{J}_j - \mathbf{S}_k\|_2^2 \tag{2}$$

where $\mathbf{J}_j \in \mathbb{R}^3$ denotes the j-th body joint position from the SMPL-X kinematic tree. $\mathbf{S}_k \in \mathbb{R}^3$ represents the k-th point in the localized scene point cloud $\mathcal{P}_s$, as shown in Figure 3. Point numbers N is variable, depending on the number of point clouds in the current localized scene. This formulation imposes stronger penalties as joints approach scene surfaces. During training, $\mathcal{L}_{sd}$ backpropagates collision-avoidance constraints through the cVAE's latent space, encouraging the generator to produce poses maintaining clearance from scene geometry while preserving natural motion kinematics.

During inference, SGAP's output parameters undergo selecting to extract three critical control signals: root joint translation $\mathbf{a}_r \in \mathbb{R}^3$, wrist joint position $\mathbf{a}_w \in \mathbb{R}^3$, and hand joint rotations $\mathbf{a}_h \in \mathbb{R}^{15\times 6}$. These elements constitute the spatial anchor tuple $\mathcal{A} = (\mathbf{a}_r, \mathbf{a}_w, \mathbf{a}_h)$, which serves as key constraints for subsequent motion synthesis.

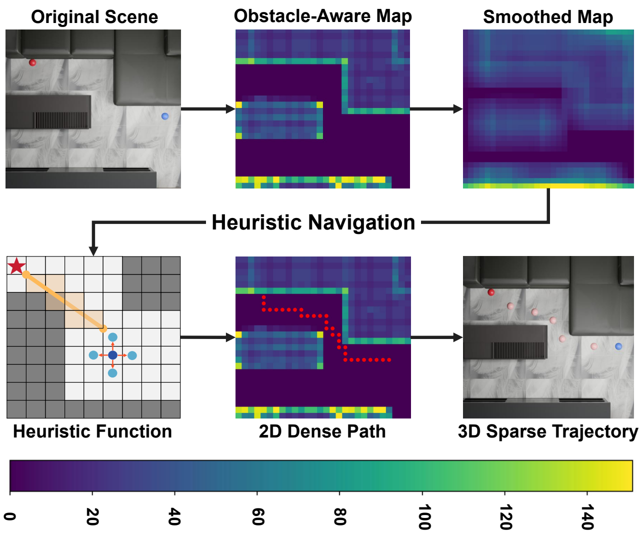

Figure 4: **Pipeline of Heuristic Navigation.** The blue ball in the original scene is the starting point, and the red ball is the end point. The obstacle-aware map is presented in the form of a heat map, and the values correspond to the axis at the bottom. In *Heuristic Function*, the dark blue dot represents the current node, the light blue dot represents the candidate node, and the red star represents the end point.

Heuristic Navigation on 2D Obstacle-Aware Map

Our Heuristic Navigation framework constitutes a customized A* variant engineered for 3D point cloud environments, designed as a plug-and-play module compatible with other models. Crucially, the algorithm maintains full replaceability. Any trajectory synthesis method satisfying interface requirements can be integrated into our pipeline. This design validates our hypothesis: global scene perception substantially outperforms local perception strategies for long-range navigation in complex 3D environments.

As illustrated in Figure 4, the pipeline executes four cohesive stages. Initially, the 3D point cloud is partitioned into volumetric blocks and compressed vertically into an obstacle-aware map, where each grid cell encodes the point numbers along the vertical axis. Then, this representation undergoes convolution smoothing to deliberately blur boundaries between obstacles and walkable regions, mitigating the algorithm's tendency to generate trajectories adhering close to obstacle surfaces. Subsequently, a dense 2D path is computed via our implementation, employing a dual-component heuristic function. As shown in the lower left corner of the Figure 4, one term evaluates Euclidean distance to the goal, like the orange line. While the other term utilizes the Bresenham line algorithm to compute cumulative traversal costs by summing values along the direct path between candidate points and the destination, like the light orange area. Finally, the resulting 2D dense path is adaptively downsampled, added height, and transformed back into the native 3D coordinate system to yield a 3D sparse trajectory. This trajectory will be added to the $\mathbf{a}_r$ of the spatial anchor tuple $\mathcal{A}$ as a root joint control signal to serve the subsequent controllable motion generation.

Due to space limitations, if you are curious about the details of this algorithm, please refer to the detailed description in the supplementary materials.

Scene-Guided Controllable Motion Generation

To enable the generation of high-fidelity interactions, two key challenges must be addressed. First, effective control over human body movement along the specified trajectory is required, with particular emphasis on enabling precise hand movements for object grasping. Second, in narrow and complex scenes, relying solely on the trajectory is insufficient to guarantee scene avoidance. Thus, a mechanism for deeply integrating scene information into the motion generation process is necessary. To address these two challenges, we respectively propose two corresponding approaches.

Controllable Generation A ControlNet-like architecture is employed, where spatial control signals serve as inputs to the control branch for generating motions aligned with trajectories, as illustrated in Figure 2. Specifically, the main branch of SCoMoGen takes the motion noises $\left\{\tilde{M}_t\right\}_{t=0}^{L}$ and the text embedding encoded by CLIP as inputs, and outputs next motion segmentation. Among them, for motion coherence, $\tilde{M}_0$ to $\tilde{M}_{10}$ are the last 11 frames of the previous motion segment. The control branch of SCoMoGen takes control signals as inputs. The control signals are converted from spatial anchor $\mathcal{A}$, where uncontrolled parts are set to 0. Key designs consist of initializing the control branch with pretrained weights from the main branch and implementing zero-initialized connection layers. These designs enable the incorporation of additional control capabilities while preserving the original model's generative performance. Notably, for the first time, SCoMoGen achieves full-body control encompassing hand control. This advancement stems from optimized hand pose representation: wrist and 15 finger joints are parameterized using 6D global rotations within the world coordinate. Deviations in trunk joint rotations result in cumulative error propagation through SMPL-X's kinematic chains in conventional methods. The proposed representation determines hand poses through 16 global joint rotations combined with wrist positioning, effectively eliminating error accumulation pathways.

Joint & Scene Guidance Despite the implemented spatial control mechanisms, generating motions that satisfy precision requirements remains challenging, since minor deviations in interactive tasks can produce noticeable motion artifacts. To address this limitation, gradient-based guidance (Guo et al. 2024) is incorporated to facilitate human-scene interaction generation without requiring additional motion post-processing. As depicted in Figure 5, two complementary constraints joint-level and scene-level guidance are implemented within the framework.

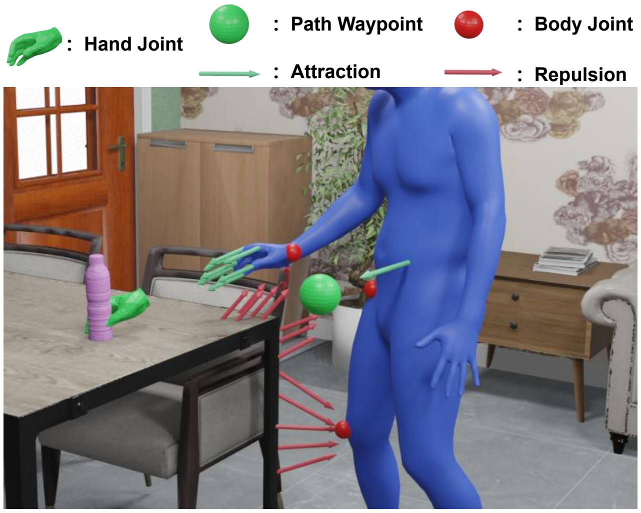

Figure 5: **Visulization of Guidance in SCoMoGen.** Precise control of motions is achieved through gradient-based guidance. Green arrows represent joints being attracted by green anchors (hand joints, path waypoints). Red arrows represent the repulsive force on the joint (red body joint) close to the scene.

The joint constraint $\mathcal{L}_j$ establishes attraction forces between body joints $\mathcal{J}$ and target anchors $\mathcal{A} = (\mathbf{a}_r, \mathbf{a}_w, \mathbf{a}_h)$ in 3D space. For root joint j_r and root anchors $\mathbf{a}_r$, the loss is formulated as:

$$\mathcal{L}_j^r = \sum_{i=1}^{L_r} \|j_r^{(i)} - \mathbf{a}_r^{(i)}\|_2^2 \tag{3}$$

where L_r is the number of path waypoints. For hand-object interactions, the constraint is implemented using:

$$\mathcal{L}_j^{hand} = \sum_{i=1}^{L_h} \parallel j_w^{(i)} - \mathbf{a}_w^{(i)} \parallel_2^2 + \sum_{i=1}^{L_h} \parallel r_h^{(i)} - \mathbf{a}_h^{(i)} \parallel_2^2 \tag{4}$$

where $j_w^{(i)}$ denotes wrist positions and $r_h^{(i)}$ represents 16 hand rotations. The parameter L_h typically equals 2, corresponding to two key poses generated by SGAP for object pickup and placement. This parameter can be extended depending on application requirements.

The scene constraint $\mathcal{L}_s$ generates adaptive repulsion forces through path-aligned point cloud. For each joint j, the repulsion loss is defined as:

$$\mathcal{L}_s = -\sum_{j\in\mathcal{J}} \sum_{p\in\mathcal{N}(j,\ 0.3\mathrm{m})} \|j - p\|_2^2 \tag{5}$$

where $\mathcal{N}(j, r)$ represents scene points located within a radius r of joint j. By restricting attention to the local $\mathcal{N}(j, r)$, irrelevant scene elements are excluded, sharply reducing

Methods	Object Locomotion			Scene Interaction			Object Interaction		
	Dist. ↓	Time ↓	Rate ↑	Pene. Rate ↓	Pene. Mean ↓	Pene. Max ↓	Contact Rate ↑	Pene. Mean ↓	Pene. Max ↓
LINGO	0.4169	16.6100	0.2333	0.3175	0.3745	89.5247	0.2418	0.0001	0.0070
CHOIS	0.3602	11.5033	0.0333	0.4134	26.3719	4271.7520	0.2809	0.0001	0.0056
Ours	**0.0270**	13.0367	**0.9333**	**0.1851**	0.6562	201.8264	**0.9800**	0.0007	0.0113

Table 1: **Quantitative results of human-object-scene interaction generation.** This mainly involves the interaction between human and scenes, human and objects when characters operate objects in the scene. At the same time, this also evaluates the efficiency and accuracy of the character in carrying objects.

computational overhead and steering optimization toward a more precise direction.

The described losses are not directly applied for result optimization, but rather integrated with the diffusion framework to refine the predicted mean μ_t at designated timesteps t through:

$$\mu_t = \mu_t - \tau \nabla_{\mu_t} \left(\lambda_1 \mathcal{L}_j^{root} + \lambda_2 \mathcal{L}_j^{hand} + \lambda_3 \mathcal{L}_s \right) \quad (6)$$

where τ controls the optimization magnitude and $\lambda_{1,2,3}$ balance constraint contributions. This approach demonstrates superior motion naturalness compared to direct iterative result optimization by preventing artifact generation from over-constrained objectives.

Experiments

Implementation Details

Our evaluation protocol comprehensively assesses both absolute performance and component effectiveness. All experiments are conducted on the TRUMANS dataset (Jiang et al. 2024b), containing 100 indoor scenes with annotated human-object interactions. We establish two evaluation axes: 1) comparison against SOTA methods in HOI and HSI generation, and 2) ablation studies isolating our core technical innovations.

Metrics We employ two complementary metric categories: Object Locomotion Assessment evaluates spatial-temporal performance through: (1) terminal positioning accuracy (Dist: Euclidean distance to target), (2) temporal efficiency (Time: task duration), and (3) reliability (Rate: success proportion within 0.05m threshold). Human Interaction Analysis quantifies physical plausibility via SDF-based penetration metrics: collision frequency (Penetration Rate), hand-object contact quality (Contact Rate), average severity (Mean Penetration Volume), and worst-case failures (Max Penetration Depth).

SOTA Comparisons While no existing method achieves identical functionality, we select two representative baselines LINGO (Jiang et al. 2024a) and CHOIS (Shi et al. 2023) for fair comparison. Quantitative evaluations cover three axes: object locomotion, human-scene interaction, hand-object interaction. The user study employs 30 sets of comparison videos assessing generation performance. Detailed metric explanation, visual comparisons, and more implementation details like baseline reproduction and training details, are provided in the supplementary material.

Ablation Study We validate our three key designs through controlled experiments: (1) **Scene-Aware**: Removes local scene information B^s and relative scene distance loss function $\mathcal{L}_{sd}$. (2) **Heuristic Navigation**: No additional experiments were set up, and its efficiency can be reflected in Table 1. (3) **Scene-Guided**: Disables gradient-based guidance (Eq. 5), relying solely on joint-based optimization.

Results

Comparison Quantitative Experiments Our method demonstrates **three key advantages** through comprehensive benchmarking. As shown in Table 1, HOSIG achieves a 93.3% success rate in object locomotion tasks, outperforming LINGO (23.3%) and CHOIS (3.3%) by factors of 4.0× and 28.3× respectively. This quantifies our method's enhanced controllability in full-body motion generation, particularly in complex navigation-objective coordination scenarios. The hierarchical scene perception mechanism yields a penetration rate of 42% - 55% lower than baselines. While LINGO shows lower mean penetration volume (0.37 vs 0.66), our method prevents catastrophic failures as evidenced by maximum penetration values: 201.8 vs CHOIS' 4,271.8. This confirms our layered constraint system effectively balances micro/macro scene interactions. As comparable contact rates (98% vs 24-28%), quantitative analysis reveals our full-body generation enables *functional* hand-object interactions, not just torching as baselines. Although baselines have low metrics on the penetration volume, this is mainly due to their extremely low hand-object contact rate. Overall, ours achieves better full-body HOSI generation.

User Study The user study involved participants evaluating motion generation outcomes through four quality metrics: Motion Quality, Trajectory Quality, Manipulation Quality, and Overall Quality. Performance assessment is conducted by measuring the relative contribution of each method to the aggregated score, with metric definitions provided in supplementary materials. As depicted in Figure 6, our HOSIG method outperforms both baselines across all metrics, achieving values approaching 50% – the theoretical maximum for each indicator. This demonstrates that our method is consistently preferred over alternatives in pairwise comparisons. The subjective evaluations align with quantitative experimental results, closely matching the observed performance hierarchy HOSIG > LINGO > CHOIS. These findings collectively demonstrate that hierarchical scene perception enables HOSIG to achieve robust human-object-scene interaction synthesis.

Ablation Study

Scene-Aware The Scene-Aware variant (removing local point cloud constraints) reveals critical insights into our spatial mechanism. As shown in Table 2, the full SGAP

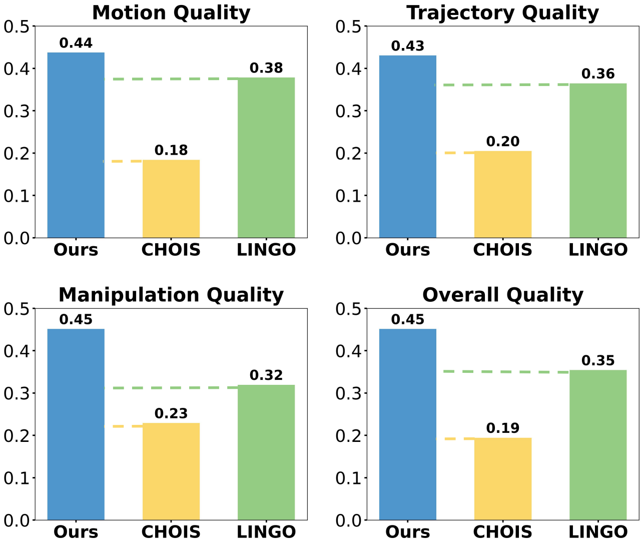

Figure 6: **User Study.** Users are asked to rate generated motions based on four different indicators, with the best being 3 points, the middle being 2 points, and the worst being 1 point. The scores are then tallied and the proportion of each method is calculated. The best result that could be obtained is 0.50. The closer it is to 0.50, the better the performance.

model improves scene interaction metrics by 16.3% (penetration rate), 69.9% (mean penetration volume), and 59.4% (max penetration volume) compared to the ablated version. Crucially, object interaction precision remained stable with $< 2\%$ variation in grasp success rate, confirming that scene awareness enhances environment adaptation without compromising manipulation capabilities. The models trained without the scene distance loss (Eq. 2) suffer from higher penetration in scene avoidance, as unstructured scene features introduce noise in the latent space. In conclusion, SGAP not only needs additional scene information B^s to perceive the scene, but also needs corresponding loss $\mathcal{L}_{sd}$ to learn how to use the scene information. Both scene information and scene distance loss are indispensable.

Heuristic Navigation The efficacy of our Heuristic Navigation framework manifests most prominently in two critical metrics: temporal efficiency (**Time** in Object Locomotion) and collision integrity (**Penetration Rate** in Scene Interaction). As quantified in Table 1, our approach achieves a 3.6 second average improvement in task completion time over the LINGO baseline. This significant acceleration stems directly from global scene comprehension, which enables anticipatory obstacle avoidance and optimal route planning. Conversely, local perception methods frequently encounter pathological scenarios, particularly in geometrically complex environments. Myopic decision making leads to navigation dead-ends and recovery behaviors, as qualitatively demonstrated in our supplementary visualizations. Furthermore, our method's superior *Penetration Rate* (↓ 23% versus LINGO) also confirms enhanced path precision.

Methods	Scene Interaction		
	Pene. Rate ↓	Pene. Mean ↓	Pene. Max ↓
SGAP	**0.5533**	**0.4484**	**12.0956**
w/o sd	0.7037	3.4615	134.6292
w/o sa	0.6611	1.4918	29.8254
Methods	Object Interaction		
	Contact. Rate ↑	Pene. Mean ↓	Pene. Max ↓
SGAP	0.9833	0.0011	0.0071
w/o sd	0.9815	**0.0007**	**0.0053**
w/o sa	**1.0000**	**0.0007**	0.0061

Table 2: **Ablation studies of SGAP.** The w/o sd means training SGAP without scene distance loss. The w/o sa means completely removing scene perception, including scene distance loss constraints and scene information B^s input to the model.

Methods	Object Locomotion		
	Dist ↓	Time ↓	Rate ↑
SCoMoGen	**0.0270**	13.0367	**0.9333**
w/o sg	0.5540	12.6511	0.6667
Methods	Scene Interaction		
	Pene. Rate ↓	Pene. Mean ↓	Pene. Max ↓
SCoMoGen	**0.1851**	**0.6562**	**201.8264**
w/o sg	0.3043	336.6398	27326.6445
Methods	Object Interaction		
	Contact. Rate ↑	Pene. Mean ↓	Pene. Max ↓
SCoMoGen	**0.9800**	**0.0007**	0.0113
w/o sg	0.9433	0.0008	0.0083

Table 3: **Ablation studies of SCoMoGen.** The w/o sg means inferencing without scene guidance $\mathcal{L}_f$.

Scene-Guided The Scene-Guided configuration (disabling Eq. 5 in gradient-based guidance) exhibits significant performance degradation across all scene interaction metrics. As shown in Table 3, quantitative results show higher penetration volumes and lower object locomotion stability compared to our full model. Notably, the guidance mechanism improves hand-object alignment precision. Experiments show that scenes are worth using as an additional modality to help generate motions. This is not only to avoid the penetration of human and scenes, but also to improve the quality of humans' activities in the scene, such as manipulating objects.

Conclusions

We present **HOSIG**, a hierarchical framework for synthesizing high-fidelity full-body human-object-scene interactions in complex 3D environments. By decoupling the task into scene-aware grasp pose generation, heuristic navigation planning, and scene-guided controllable motion synthesis, our method addresses critical limitations in existing HOI and HSI approaches. The proposed hierarchical scene perception mechanism ensures collision-free interactions while maintaining precise hand-object contact and natural locomotion. Notably, our framework achieves unlimited motion length through autoregressive generation and requires minimal manual intervention, making it practical for applications in VR, robotics, and animation.

Acknowledgments

This work was supported by the National Key R&D Program of China (2024YFB2809102), the National Natural Science Foundation of China (62332010, 62476131, 62125107, 62377004), the Major Science and Technology Projects in Jiangsu Province under Grant BG2024042 and Fundamental Research Funds for the Central Universities (No. 2233100028).

References

Cen, Z.; Pi, H.; Peng, S.; Shen, Z.; Yang, M.; Zhu, S.; Bao, H.; and Zhou, X. 2024. Generating human motion in 3D scenes from text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1855–1866.

Ghosh, A.; Cheema, N.; Oguz, C.; Theobalt, C.; and Slusallek, P. 2021. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1396–1406.

Ghosh, A.; Dabral, R.; Golyanik, V.; Theobalt, C.; and Slusallek, P. 2023. IMoS: Intent-Driven Full-Body Motion Synthesis for Human-Object Interactions. In *Computer Graphics Forum*, volume 42, 1–12. Wiley Online Library.

Guo, Y.; Yuan, H.; Yang, Y.; Chen, M.; and Wang, M. 2024. Gradient guidance for diffusion models: An optimization perspective. *Advances in Neural Information Processing Systems*, 37: 90736–90770.

Hassan, M.; Guo, Y.; Wang, T.; Black, M.; Fidler, S.; and Peng, X. B. 2023. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–9.

Huang, S.; Wang, Z.; Li, P.; Jia, B.; Liu, T.; Zhu, Y.; Liang, W.; and Zhu, S.-C. 2023. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16750–16761.

Jiang, N.; He, Z.; Wang, Z.; Li, H.; Chen, Y.; Huang, S.; and Zhu, Y. 2024a. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia 2024 Conference Papers*, 1–11.

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

Karunratanakul, K.; Preechakul, K.; Suwajanakorn, S.; and Tang, S. 2023. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2151–2162.

Li, H.; Yu, H.-X.; Li, J.; and Wu, J. 2024a. ZeroHSI: Zero-Shot 4D Human-Scene Interaction by Video Generation. *arXiv preprint arXiv:2412.18600*.

Li, J.; Clegg, A.; Mottaghi, R.; Wu, J.; Puig, X.; and Liu, C. K. 2024b. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, 54–72. Springer.

Li, J.; Wu, J.; and Liu, C. K. 2023. Object Motion Guided Human Motion Synthesis. *ACM Transactions on Graphics (TOG)*, 42: 1 – 11.

Li, L.; and Dai, A. 2024. Genzi: Zero-shot 3d human-scene interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20465–20474.

Liu, L.; and Hodgins, J. 2018. Learning basketball dribbling skills using trajectory optimization and deep reinforcement learning. *Acm transactions on graphics (tog)*, 37(4): 1–14.

Lu, J.; Zhang, H.; Ye, Y.; Shiratori, T.; Starke, S.; and Komura, T. 2024. CHOICE: Coordinated Human-Object Interaction in Cluttered Environments for Pick-and-Place Actions. *ArXiv*, abs/2412.06702.

Mao, W.; Hartley, R. I.; Salzmann, M.; et al. 2022. Contact-aware human motion forecasting. *Advances in Neural Information Processing Systems*, 35: 7356–7367.

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

Peng, X. B.; Chang, M.; Zhang, G.; Abbeel, P.; and Levine, S. 2019. Mcp: Learning composable hierarchical control with multiplicative compositional policies. *Advances in neural information processing systems*, 32.

Peng, X. B.; Ma, Z.; Abbeel, P.; Levine, S.; and Kanazawa, A. 2021. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4): 1–20.

Petrov, I. A.; Marin, R.; Chibane, J.; and Pons-Moll, G. 2024. TriDi: Trilateral Diffusion of 3D Humans, Objects, and Interactions. *arXiv preprint arXiv:2412.06334*.

Rempe, D.; Luo, Z.; Bin Peng, X.; Yuan, Y.; Kitani, K.; Kreis, K.; Fidler, S.; and Litany, O. 2023. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13756–13766.

Shi, Y.; Wang, J.; Jiang, X.; and Dai, B. 2023. Controllable motion diffusion model. *arXiv preprint arXiv:2306.00416*.

Song, W.; Zhang, X.; Li, S.; Gao, Y.; Hao, A.; Hou, X.; Chen, C.; Li, N.; and Qin, H. 2024. Hoianimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 811–820.

Taheri, O.; Choutas, V.; Black, M. J.; and Tzionas, D. 2022. Goal: Generating 4d whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13263–13273.

Taheri, O.; Zhou, Y.; Tzionas, D.; Zhou, Y.; Ceylan, D.; Pirk, S.; and Black, M. J. 2024. Grip: Generating interaction poses using spatial cues and latent consistency. In *2024 International Conference on 3D Vision (3DV)*, 933–943. IEEE.

Wang, J.; Luo, Z.; Yuan, Y.; Li, Y.; and Dai, B. 2024a. Pacer+: On-demand pedestrian animation controller in driving scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 718–728.

Wang, J.; Rong, Y.; Liu, J.; Yan, S.; Lin, D.; and Dai, B. 2022a. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20460–20469.

Wang, Y.; Lin, J.; Zeng, A.; Luo, Z.; Zhang, J.; and Zhang, L. 2023. Physhoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*.

Wang, Z.; Chen, Y.; Jia, B.; Li, P.; Zhang, J.; Zhang, J.; Liu, T.; Zhu, Y.; Liang, W.; and Huang, S. 2024b. Move as you say interact as you can: Language-guided human motion generation with scene affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 433–444.

Wang, Z.; Chen, Y.; Liu, T.; Zhu, Y.; Liang, W.; and Huang, S. 2022b. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35: 14959–14971.

Wu, Y.; Wang, J.; Zhang, Y.; Zhang, S.; Hilliges, O.; Yu, F.; and Tang, S. 2022. Saga: Stochastic whole-body grasping with contact. In *European Conference on Computer Vision*, 257–274. Springer.

Wu, Z.; Li, J.; Xu, P.; and Liu, C. K. 2024. Human-object interaction from human-level instructions. *arXiv preprint arXiv:2406.17840*.

Xiao, Z.; Wang, T.; Wang, J.; Cao, J.; Zhang, W.; Dai, B.; Lin, D.; and Pang, J. 2023. Unified human-scene interaction via prompted chain-of-contacts. *arXiv preprint arXiv:2309.07918*.

Xu, S.; Li, Z.; Wang, Y.-X.; and Gui, L.-Y. 2023. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14928–14940.

Xuan, H.; Li, X.; Zhang, J.; Zhang, H.; Liu, Y.; and Li, K. 2023. Narrator: Towards natural control of human-scene interaction generation via relationship reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22268–22278.

Yang, J.; Niu, X.; Jiang, N.; Zhang, R.; and Huang, S. 2024. F-HOI: Toward Fine-grained Semantic-Aligned 3D Human-Object Interactions. In *European Conference on Computer Vision*, 91–110. Springer.

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

YUAN, Y.; and Makoviychuk, V. 2023. Learning physically simulated tennis skills from broadcast videos.

Zhang, J.; Zhang, Y.; An, L.; Li, M.; Zhang, H.; Hu, Z.; and Liu, Y. 2024a. ManiDext: Hand-Object Manipulation Synthesis via Continuous Correspondence Embeddings and Residual-Guided Diffusion. *ArXiv*, abs/2409.09300.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, Y.; Hassan, M.; Neumann, H.; Black, M. J.; and Tang, S. 2020. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6194–6204.

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

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

Zhang, Y.; Yang, H.; Luo, C.; Peng, J.; Wang, Y.; and Zhang, Z. 2024b. OOD-HOI: Text-Driven 3D Whole-Body Human-Object Interactions Generation Beyond Training Domains. *arXiv preprint arXiv:2411.18660*.

Zhao, K.; Wang, S.; Zhang, Y.; Beeler, T.; and Tang, S. 2022. Compositional human-scene interaction synthesis with semantic control. In *European Conference on Computer Vision*, 311–327. Springer.

Zhao, K.; Zhang, Y.; Wang, S.; Beeler, T.; and Tang, S. 2023. Synthesizing diverse human motions in 3d indoor scenes. In *Proceedings of the IEEE/CVF international conference on computer vision*, 14738–14749.

Zheng, Y.; Shi, Y.; Cui, Y.; Zhao, Z.; Luo, Z.; and Zhou, W. 2023. Coop: Decoupling and coupling of whole-body grasping pose generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2163–2173.