MOGRAS: Human Motion with Grasping in 3D Scenes

Kunal Bhosikar[1]
https://kunal-kamalkishor-bhosikar.github.io/
Siddharth Katageri[1]
https://siddharthkatageri.github.io/
Vivek Madhavaram[1]
https://vivekmadhavaram.github.io/vivek_page/
Kai Han[2]
https://www.kaihan.org/
Charu Sharma[1]
https://charusharma.org/

[1] Machine Learning Lab,
International Institute of Information Technology,
Hyderabad, India

[2] Vision AI Lab,
The University of Hong Kong,
Pokfulam Road, Hong Kong

Figure 1: **Overview of the MOGRAS Dataset.** Our MOGRAS dataset provides full-body motion sequences, including both pre-grasp walking and detailed grasping poses, within richly annotated 3D scenes. This unique combination addresses the limitations of existing datasets, bridging the gap between synthetic data and real-world scenarios to enable realistic and complex human-object interactions.

Abstract

Generating realistic full-body motion interacting with objects is critical for applications in robotics, virtual reality, and human-computer interaction. While existing methods can generate full-body motion within 3D scenes, they often lack the fidelity for fine-grained tasks like object grasping. Conversely, methods that generate precise grasping motions typically ignore the surrounding 3D scene. This gap, generating full-body grasping motions that are physically plausible within a 3D scene, remains a significant challenge. To address this, we introduce **MOGRAS** (Human **MO**tion with **GRA**sping in 3D **S**cenes), a large-scale dataset that bridges this gap. MOGRAS provides pre-grasping full-body walking motions and final grasping poses within richly annotated 3D indoor scenes. We leverage MOGRAS to benchmark existing full-body grasping methods and demonstrate their limitations in scene-aware generation. Furthermore, we propose a simple yet effective method to adapt existing approaches to work seamlessly within 3D scenes. Through extensive quantitative and qualitative experiments, we validate the effectiveness of our dataset and highlight the significant improvements our proposed method achieves, paving the way for more realistic human-scene interactions.

1 Introduction

Understanding how humans interact with objects in 3D space is fundamental to numerous domains, including computer vision, robotics, animation, and virtual reality [5, 9, 16, 19, 30]. Accurately modeling these interactions enables behavior analysis, the development of intelligent robotic systems, and the creation of realistic, immersive virtual environments [12, 28]. A key challenge lies in simulating realistic and physically plausible human-object interactions (HOI), which involves modeling contact, force, and object manipulation in complex 3D scenes [15].

Despite recent advances, capturing the nuanced dynamics of HOI remains difficult. While many existing datasets and methods focus on large-scale interactions like sitting on a sofa or using a bed [36], others target fine-grained hand-object manipulation, often ignoring the rest of the body [7, 10, 13, 29]. Some work has extended this to full-body grasping [32], but these approaches typically neglect scene context. This highlights a critical gap: the lack of datasets and models that account for full-body grasping of small objects within real 3D indoor scenes. Addressing this is essential for advancing scene-aware HOI, as performing full-body motion with objects in cluttered environments requires precision in contact and collision handling.

To address this challenge, we introduce MOGRAS, a novel, large-scale synthetic dataset containing full-body motion sequences and grasping interactions in indoor 3D scenes. Each sequence includes both pre-grasp walking motion and the final grasping pose, as shown in Figure 1. Our pipeline ensures minimal collision between the human mesh, the grasped object, and the surrounding scene geometry, enabling realistic motion and contact.

Manually capturing such complex HOI data is prohibitively expensive, requiring specialized hardware and extensive manual labor. To overcome this, we propose an automated data generation framework comprising five key steps: (*i*) placing small objects on surfaces, (*ii*) aligning pre-grasp walking motion, (*iii*) refining the scene to handle collisions, (*iv*) generating a realistic grasping pose, and (*v*) infilling the motion to smoothly transition to the grasp. Given the synthetic nature of our dataset, we conduct an extensive human evaluation to validate its realism and naturalness, ensuring its suitability for training models for multiple downstream tasks.

We benchmark existing grasping models, GOAL [33] and SAGA [37], and observe that they fail to produce plausible grasps without significant scene penetration. To address this, we extend the GOAL's GNet [33] architecture by introducing a novel penetration loss and conditioning the model on the 3D scene mesh. This modification, which we name GNet++, significantly improves grasping quality in cluttered scenes. Our work emphasizes the importance of scene-aware grasping, where the objective is to minimize the intersection between the human body and the scene. We show how our method uses a pretrained model to generate initial grasps and refines them using our proposed loss to reduce collisions and produce physically plausible contact.

Our contributions can be summarized as follows:
(i) We introduce MOGRAS, a large-scale synthetic dataset for scene-aware, full-body grasping. It includes both pre-grasp approaching motion and final grasping poses, generated automatically through our proposed framework.
(ii) We provide a comprehensive benchmark of existing grasping methods on our challenging dataset, highlighting their limitations.
(iii) We propose a novel method, GNet++, that improves scene-aware grasping by leveraging our dataset and introducing a new penetration loss to produce physically plausible results.

2 Related Work

Our work is positioned at the intersection of 3D human-object grasping and human-scene interaction. We review key advancements in both areas to highlight the specific gap our work addresses.

3D Human-Object Grasping. Research in 3D grasp synthesis has historically focused on isolated hand-object interactions. Generative models like Grasping Field [13] and GrabNet [32] used implicit functions and VAEs to produce realistic hand poses. Subsequent work, such as GraspTTA [10] and ContactOpt [7], refined these methods through contact-aware optimization to improve physical plausibility.

More recently, the focus has expanded to full-body grasp synthesis, leveraging datasets like GRAB [32], which provides MoCap data of SMPL-X [23] bodies interacting with objects. Models like GOAL [33] and SAGA [37] have built on this to generate full-body poses for handheld objects. Even more recent works, such as DiffGrasp [41], explore diffusion-based approaches for generating high-quality full-body grasping sequences. However, a critical limitation of these methods is their lack of scene awareness; they generate grasps in isolation, ignoring potential collisions with the surrounding environment. While OmniGrasp [21] explores diverse object grasping with virtual humanoids, it also omits real-world scene context.

A few methods have attempted to incorporate scene constraints. FLEX [34] generates scene-aware full-body grasps by leveraging hand-object and foot-ground contacts, but its optimization-based inference is computationally expensive. VirtualHome [26] simulates household activities but often produces unrealistic interactions with significant penetration due to its reliance on programmatic, pre-defined animations. These approaches show the need for scene context but highlight the difficulty of achieving it efficiently and realistically.

3D Human-Scene Interaction (HSI). Parallel research in human-scene interaction (HSI) aims to generate plausible human motions within rich 3D environments. Early optimization-based methods used physical priors and contact annotations to place humans in scenes [14, 17, 42]. Modern learning-based approaches have improved scalability by conditioning pose generation on scene representations, using either scene-centric [40] or human-centric [8] perspectives.

Large-scale datasets have been pivotal in advancing HSI. Datasets like HUMANISE [36], GTA-IM [2], and more recent works like SMPLOlympics [20] and UniHSI [38] provide diverse human motions in varied environments. TRUMANS [11] further improves data quality with high-fidelity contact supervision. However, the focus of these HSI datasets is on large-scale interactions like sitting, walking, or navigating. They generally lack the fine-grained contact information and specific intent required for precise object grasping. Consequently, models trained on them excel at general scene placement but fail at detailed manipulation tasks.

Positioning and Contribution. The existing literature reveals a clear dichotomy: 3D grasping research is largely scene-agnostic, while 3D HSI research focuses on large-scale interactions and lacks fine-grained grasping detail. This leaves a critical, unaddressed gap for tasks requiring full-body, scene-aware manipulation of small objects. Our work directly targets this gap. MOGRAS is the first large-scale dataset to provide physically plausible, full-body grasping motions within cluttered indoor scenes. By synthesizing realistic approaching motions and collision-free final grasps, we provide the necessary data to train and evaluate the next generation of truly context-aware HOI models.

3 The MOGRAS Dataset

Existing datasets for full-body human-object interaction (HOI) [1, 18, 34, 35] suffer from several limitations. As we outlined in the previous section, datasets capturing human motion in 3D environments tend to focus on interactions with large, static objects like tables or sofas, while overlooking fine-grained interactions with small, handheld objects. Conversely, datasets for small object interaction often omit the 3D scene context entirely, failing to account for critical environmental constraints. Since scene context fundamentally shapes how humans approach, manipulate, and grasp objects, a realistic model of full-body HOI must incorporate both fine-grained interactions and the surrounding 3D scene.

To address these shortcomings, we introduce MOGRAS, a large-scale, synthetic dataset featuring full-body grasping in diverse 3D indoor scenes. Manually acquiring such a dataset would be prohibitively expensive and time-consuming. To overcome this, we propose an automated synthesis pipeline that efficiently generates high-quality interaction sequences. The core idea is to first align an existing human walk motion to a 3D scene, then synthesize a physically plausible grasping pose, and finally infill the motion to smoothly transition from the walk to the grasp. Figure 2 provides an overview of our data generation process.

Our pipeline consists of five key steps, which we detail below.

3.1 Walk Motion Alignment and Object Placement

We follow the methodology of HUMANISE [36] to align walking motions with 3D scenes. We use walking sequences from AMASS [22] and 3D environments from ScanNetv2 [3]. Motion labels from BABEL [27] are used to segment walk motions, while graspable objects are sourced from GRAB [32]. Our dataset includes 2,988 walking motion clips, each 1 to 4 seconds in duration.

For each sequence, we first identify a scene element, or *receptacle* (e.g., a table, counter, or desk), from ScanNet's [3] semantic annotations. We then optimize for a global translation t and rotation R to ensure the human body is collision-free and the walking trajectory terminates near the chosen receptacle [36].

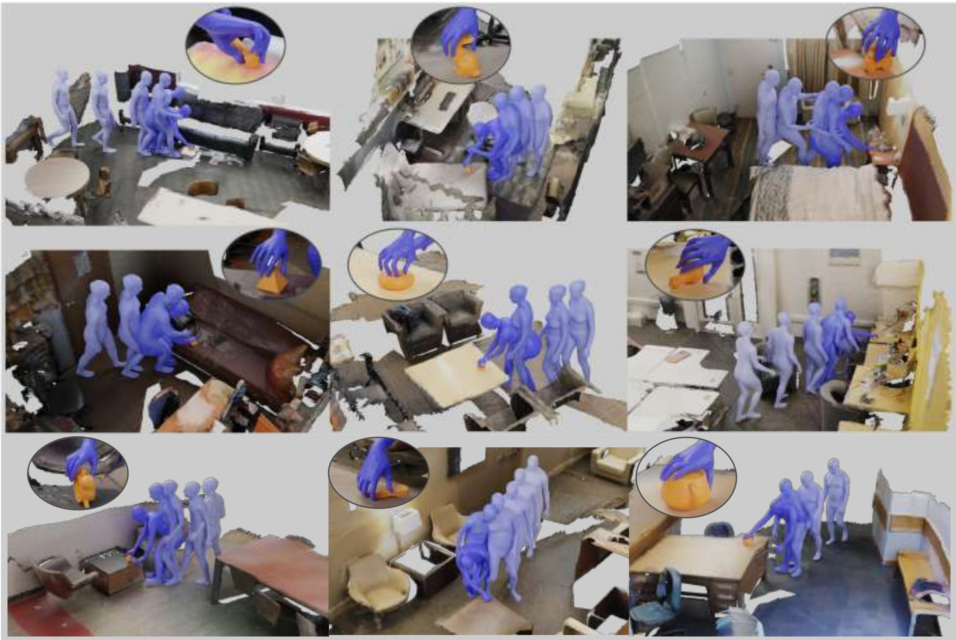

Figure 2: **Examples from the MOGRAS dataset.** MOGRAS provides full-body motion sequences, including a pre-grasp walking phase and a final grasping pose, all within richly annotated 3D scenes. The dataset captures subtle, physically plausible interactions with small objects and their environment, addressing a key gap in existing research.

Once the motion is aligned, we place a graspable object on the receptacle's surface to ensure it is reachable from the final human position. We achieve this by extracting the receptacle mesh, identifying a suitable placement point using a KD-tree, and placing the object at the highest point on the surface to prevent floating artifacts. This pipeline generalizes well across diverse scenes with high-quality semantic segmentation (see Figure 2).

3.2 Refining Scenes from ScanNet

ScanNet [3] scenes often suffer from floor misalignment with the ideal $z = 0$ ground plane, leading to foot-floating or penetration artifacts in the generated motion. A global rigid transform is inadequate due to spatially varying deviations, while non-rigid warping risks distorting object layouts. To address this, we propose a piecewise rigid alignment method that adjusts floor geometry while preserving the overall scene structure (see Figure 3).

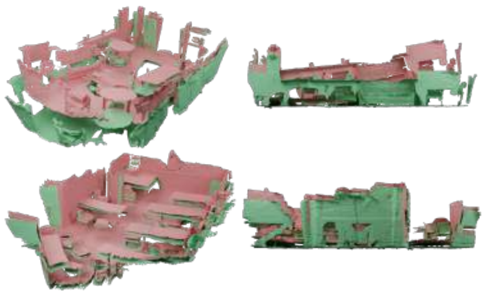

Figure 3: **Comparison of Original and Refined ScanNet Scenes.** Original scenes (pink) often have floor misalignment, causing artifacts like floating feet. Our refined scenes (green) correct this, improving physical plausibility and interaction quality.

Our solution applies local rigid transforms computed via ICP over fixed-size sliding windows on the $x-y$ plane. For each window, we extract floor vertices and align them to the $z = 0$ plane by optimizing for vertical translation t_z and in-plane rotations $\mathsf{R}_x, \mathsf{R}_y$. This strategy yields significantly better floor alignment, as quantified in Table 1, and enables more physically plausible human-scene interactions.

3.3 Generating Grasping Poses

We generate realistic full-body grasping poses conditioned on object placement and scene context using FLEX [34], an optimization-based method that searches the latent spaces of VPoser [24] (body pose) and GrabNet [32] (hand pose). Although FLEX [34] yields high-quality poses, its computational cost is high. To enable large-scale generation, we adapt FLEX [34] by reducing optimization iterations and applying two post-optimization refinements, achieving a significant speedup without compromising pose quality. We provide implementation details of these modifications in the Supplementary Appendix.

Table 1: **Quantitative Floor Alignment.** This table shows that our method significantly reduces the deviation of floor vertices from the ground plane ($z = 0$) in refined ScanNet scenes, demonstrating improved alignment.

	Average Mean	**Average Standard Deviation**	**Global Mean Mean**	**Global Standard Deviation**
Original	0.1175	0.1303	0.1396	0.2591
Refined	**0.0045**	**0.0208**	**0.0039**	**0.0394**

3.4 Generating Final Approaching Motion

To create a coherent sequence, we seamlessly integrate the generated grasping pose (Section 3.3) with the preceding walking motion (Section 3.1) using PriorMDM [31], a motion infilling model. We adopt its root-trajectory variant, which generates intermediate poses by interpolating between the pelvis of the final walking frame and the grasping pose. To avoid unnatural backtracking, we adjust the grasping location to ensure a forward motion, preserving trajectory realism. The infill duration scales with the distance between the start and end poses, allowing for smooth, adaptive transitions. Additionally, for each object-receptacle pair, we provide both the generated motion and multiple pelvis locations near the grasp. These serve as valid, collision-free grasp targets to support downstream tasks like evaluation or learning.

3.5 Dataset Scale and Quality

MOGRAS is a large-scale, high-quality dataset for full-body human-object interaction. As detailed in Table 2, MOGRAS contains over 14k motion sequences across 500 scenes and 47 objects. Unlike prior work that focuses on either large-scale interactions or scene-agnostic grasping, MOGRAS uniquely combines full-body motion, fine-grained grasping, and 3D scene context (Table 3).

Table 2: **Statistics of the MOGRAS Dataset Splits.** This table presents a breakdown of the training, validation, and test sets. The splits are structured to ensure diversity and broad coverage of scenes and object interactions.

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

To ensure realism, we use a multi-stage quality assurance process that includes automated penetration filtering and a human study. In this study, 75 participants rated our

Table 3: **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.

Dataset	#Samples	#Frames	#Scenes	#Objects	3D Scene Inclusion	Fine-Grained Grasping	Full-Body Motion
HUMANISE [36]	19.6k	1.2M	643	✗	✓	✗	✓
GRAB [32]	1.3k	1.6M	✗	51	✗	✓	✓
OakInk [39]	50k	0.23M	✗	1800	✗	✓	✗
SceneFun3D [4]	14.8k	✗	710	✗	✓	✗	✗
ScenePlan [38]	1.1k	✗	10	40	✓	✗	✓
MOGRAS	14.2k	0.8M	507	47	✓	✓	✓

motions for naturalness and interaction realism, yielding high scores of 4.2/5 and 4.5/5, respectively. Motions below a rating of 3.5 were removed, ensuring our synthetic data is physically plausible and suitable for training models.

4 Scene-Aware Grasping with GNet++

We benchmark our MOGRAS dataset on the task of scene-aware grasping, focusing specifically on the static grasp pose rather than full motion sequences. This decision reflects the fine-grained challenge of generating physically plausible grasps in complex 3D scenes, where avoiding human-scene and human-object intersections is critical. Our proposed method, GNet++, is an extension of the pretrained GNet [33] model, which we use as our baseline to refine scene-agnostic grasps into context-aware and physically valid ones.

4.1 Preliminaries: GOAL and GNet

SMPL-X [23] is a parametric 3D human model with 10,475 vertices, controlled by shape, pose, and expression parameters. It is widely used to represent human bodies, faces, and hands in a unified mesh.

GOAL [33] is a two-stage framework for human-object interaction. Given an object and an initial human pose, it first uses GNet to predict a realistic whole-body grasping pose. The second stage, MNet, then generates a full motion trajectory. Our work focuses on improving the GNet stage.

GNet is a conditional variational autoencoder (cVAE) that generates full-body grasp poses conditioned on object geometry and position. Its encoder maps input parameters (SMPL-X [23], BPS [25] object encoding, and object translation) to a latent space. The decoder uses a sampled latent grasp code and object condition to predict SMPL-X [23] parameters, a head orientation, and offset vectors for refining hand-object contact.

4.2 Our Method: GNet++

We introduce GNet++, a scene-aware extension of GNet, as illustrated in Figure 4. We make two key modifications to enable it to reason about environmental constraints:

1. Scene-Conditioned Architecture: We augment GNet's input to include scene information. The 3D scene is encoded into a feature vector using a pretrained Vision Transformer (ViT) [6] on a bird's-eye-view (BEV) projection. This scene embedding is fused with the object encoding and injected into both the encoder and decoder of the cVAE, allowing the model to learn the relationship between the grasp pose and the surrounding geometry. While the BEV projection effectively captures floor-level and planar obstacles, we acknowledge its limitation in fully representing vertical or overhanging obstacles (e.g., shelves) that are not directly projected onto the ground plane. Our proposed penetration loss helps mitigate this by explicitly penalizing human-scene intersections in 3D space.

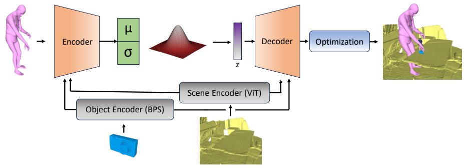

Figure 4: **GNet++ Architecture Overview.** We extend the GNet architecture to create GNet++ by injecting a scene embedding into both the encoder and decoder, enabling it to generate context-aware grasps that avoid collisions.

2. Penetration-Aware Loss: To explicitly discourage collisions, we introduce a new loss term, $(\mathcal{L}_{pen})$. We first voxelize the 3D scene around the object. Any voxel below an occupied one is also marked as occupied to account for unreachable spaces. During training, we penalize any predicted human mesh vertex that falls within these occupied voxels. This loss, combined with fine-tuning on our MOGRAS dataset, teaches the model to generate physically plausible poses that respect scene boundaries.

4.3 Experiments and Evaluation

We evaluate GNet++ on the test set of MOGRAS, comparing it against the original GOAL [33], SAGA [37], and ablations of our own model. For a comprehensive overview of our baseline evaluation and experimental setup, please refer to the Appendix.

Metrics. To assess physical plausibility and interaction realism, we report performance on four key metrics:

Human-Scene Penetration$(\downarrow)$**:** The proportion of human body vertices located inside occupied voxels of the 3D scene, indicating undesired interpenetrations.

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

Human-Object Penetration$(\downarrow)$**:** Mean signed distance function (SDF) values computed at hand vertices.

Contact Precision/Recall$(\uparrow)$**:** F1 score measuring how accurately predicted contact points align with ground-truth contact areas on objects and the floor. For a detailed definition of our ground-truth contact annotations, please see the Appendix.

Quantitative Results. Table 4 shows that GNet++ significantly outperforms all baselines. Compared to GOAL, GNet++ reduces scene penetration by 28% and object penetration by 58%, while dramatically improving floor contact (0.92 vs. 0.76 F1). Figure 5 visually confirms these findings, showing that GOAL and SAGA frequently generate poses that intersect with the table, whereas GNet++ produces a realistic, collision-free crouch.

Ablation Studies. Our ablations in Table 4 isolate the impact of our key contributions.

Table 4: **Quantitative Comparison on Scene-Aware Grasping.** We evaluate penetration (↓) and contact quality (↑). Our full GNet++ model, fine-tuned on MOGRAS with the proposed penetration loss, significantly outperforms both state-of-the-art baselines and its ablated variants across all metrics. This demonstrates the importance of both scene-aware conditioning and explicit collision penalization.

Method	Penetration (↓)			Object Contact (↑)			Floor Contact (↑)		
	Scene	Object	Floor	Precision	Recall	F1	Precision	Recall	F1
GOAL [33]	4.35%	3.49%	3.62%	0.85	0.76	0.80	0.70	0.83	0.76
SAGA [37]	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**

Effect of MOGRAS: Fine-tuning on our dataset (comparing row 3 vs. row 5) is crucial for reducing floor and scene penetration, as it provides the necessary examples of full-body interaction in context.

Effect of Penetration Loss: Adding our explicit penetration loss (comparing row 4 vs. row 5) provides the largest single improvement in scene penetration, proving its effectiveness in enforcing physical constraints.

These results confirm that both our MOGRAS dataset and the proposed GNet++ architecture are essential for achieving high-quality, scene-aware grasping.

5 Conclusion

In this work, we introduced MOGRAS, a large-scale dataset that addresses a critical gap in human-scene interaction research by providing full-body grasping motions within complex 3D scenes. We also proposed GNet++, a novel generative model that leverages this dataset to produce physically plausible and scene-aware grasps. Our extensive quantitative and qualitative evaluations demonstrate that GNet++ significantly outperforms existing baselines by generating human meshes with minimal scene and object penetration, thereby improving the realism and physical accuracy of human-object interactions.

Beyond its immediate use for scene-aware grasping, the MOGRAS dataset serves as a versatile resource for future research in motion generation, simulation, and scene understanding. By providing a rich foundation of physically plausible human-scene interactions, we hope our work encourages the community to explore more complex, dynamic, and context-aware human behaviors in virtual environments.

References

[1] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions, 2022.

[2] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. *CoRR*, abs/2007.03672, 2020. URL https://arxiv.org/abs/2007.03672.

Figure 5: **Qualitative Comparison of Generated Grasps.** We visualize grasp poses from GNet++ (blue), GOAL (green), and SAGA (pink). In these cluttered scenes, both GOAL and SAGA produce significant penetration with the table and surrounding objects. In contrast, GNet++ successfully generates a physically plausible and collision-free pose by adapting the body to the environment.

[3] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2443, 2017. doi: 10.1109/CVPR.2017.261.

[4] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. SceneFun3D: Fine-Grained Functionality and Affordance Understanding in 3D Scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[5] Renaud Detry, Dirk Kraft, Anders Glent Buch, Norbert Krueger, and Justus Piater. Refining grasp affordance models by experience, 2010-01-01. ISSN 2577-087X.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[7] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. Contactopt: Optimizing contact to improve grasps. *CoRR*, abs/2104.07267, 2021. URL https://arxiv.org/abs/2104.07267.

[8] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14708–14718, June 2021.

[9] Kaijen Hsiao and Tomas Lozano-Perez. Imitation learning of whole-body grasps. *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5657–5662, 2006. URL https://api.semanticscholar.org/CorpusID:2468294.

[10] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. *CoRR*, abs/2104.03304, 2021. URL https://arxiv.org/abs/2104.03304.

[11] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan 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.

[12] Maciej Kalisiak and Michiel Panne. A grasp-based motion planning algorithm for character animation. *The Journal of Visualization and Computer Animation*, 12, 07 2001. doi: 10.1002/vis.250.

[13] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J. Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. *CoRR*, abs/2008.04451, 2020. URL https://arxiv.org/abs/2008.04451.

[14] Vladimir G. Kim, Siddhartha Chaudhuri, Leonidas Guibas, and Thomas Funkhouser. Shape2pose: human-centric shape analysis. *ACM Trans. Graph.*, 33(4), jul 2014. ISSN 0730-0301. doi: 10.1145/2601097.2601117. URL https://doi.org/10.1145/2601097.2601117.

[15] Robert Krug, Dimitar Dimitrov, Krzysztof Charusta, and Boyko Iliev. On the efficient computation of independent contact regions for force closure grasps. In *IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems (IROS 2010) :*, pages 586–591, 2010. ISBN 978-1-4244-6675-7. doi: 10.1109/IROS.2010.5654380.

[16] Paul G. Kry and Dinesh K. Pai. Interaction capture and synthesis. *ACM Trans. Graph.*, 25(3):872–880, jul 2006. ISSN 0730-0301. doi: 10.1145/1141911.1141969. URL https://doi.org/10.1145/1141911.1141969.

[17] Kurt Leimer, Andreas Winkler, Stefan Ohrhallinger, and Przemyslaw Musialski. Pose to seat: Automated design of body-supporting surfaces. *CoRR*, abs/2003.10435, 2020. URL https://arxiv.org/abs/2003.10435.

[18] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Trans. Graph.*, 42(6), 2023.

[19] Ying Li, Jiaxin L. Fu, and Nancy S. Pollard. Data-driven grasp synthesis using shape matching and task-based pruning. *IEEE Transactions on Visualization and Computer Graphics*, 13(4):732–747, jul 2007. ISSN 1077-2626. doi: 10.1109/TVCG.2007.1033. URL https://doi.org/10.1109/TVCG.2007.1033.

[20] Zhengyi Luo, Jiashun Wang, Kangni Liu, Haotian Zhang, Chen Tessler, Jingbo Wang, Ye Yuan, Jinkun Cao, Zihui Lin, Fengyi Wang, Jessica Hodgins, and Kris Kitani. Smplolympics: Sports environments for physically simulated humanoids, 2024. URL https://arxiv.org/abs/2407.00187.

[21] Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Omnigrasp: Grasping diverse objects with simulated humanoids, 2025. URL https://arxiv.org/abs/2407.11385.

[22] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: archive of motion capture as surface shapes. *CoRR*, abs/1904.03278, 2019. URL http://arxiv.org/abs/1904.03278.

[23] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. *CoRR*, abs/1904.05866, 2019. URL http://arxiv.org/abs/1904.05866.

[24] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael 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)*, 2019.

[25] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. *CoRR*, abs/1908.09186, 2019. URL http://arxiv.org/abs/1908.09186.

[26] Xavier Puig, Kevin Kyunghwan Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2018. URL https://api.semanticscholar.org/CorpusID:49317780.

[27] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael 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.

[28] Hans Rijpkema and Michael Girard. Computer animation of knowledge-based human grasping. In *Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '91, page 339–348, New York, NY, USA, 1991. Association for Computing Machinery. ISBN 0897914368. doi: 10.1145/122718.122754. URL https://doi.org/10.1145/122718.122754.

[29] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *CoRR*, abs/2201.02610, 2022. URL https://arxiv.org/abs/2201.02610.

[30] Jungwon Seo, Soonkyum Kim, and Vijay Kumar. Planar, bimanual, whole-arm grasping. In *2012 IEEE International Conference on Robotics and Automation*, pages 3271–3277, 2012. doi: 10.1109/ICRA.2012.6225086.

[31] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023.

[32] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. URL https://grab.is.tue.mpg.de.

[33] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://goal.is.tue.mpg.de.

[34] Purva Tendulkar, Dídac Surís, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[35] Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physhoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023.

[36] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[37] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[38] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts, 2024. URL https://arxiv.org/abs/2309.07918.

[39] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction, 2022. URL https://arxiv.org/abs/2203.15709.

[40] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. Generating person-scene interactions in 3d scenes. *CoRR*, abs/2008.05570, 2020. URL https://arxiv.org/abs/2008.05570.

[41] Yonghao Zhang, Qiang He, Yanguang Wan, Yinda Zhang, Xiaoming Deng, Cuixia Ma, and Hongan Wang. Diffgrasp: Whole-body grasping synthesis guided by object motion using a diffusion model, 2024. URL https://arxiv.org/abs/2412.20657.

[42] Youyi Zheng, Han Liu, Julie Dorsey, and Niloy J. Mitra. Ergonomics-inspired reshaping and exploration of collections of models. *IEEE Transactions on Visualization and Computer Graphics*, 22(6):1732–1744, 2016. doi: 10.1109/TVCG.2015.2448084.