# 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

# Appendix

## A Augmenting Graspable Pelvis Positions in MOGRAS

For each object and scene receptacle pair in our MOGRAS dataset, we provide not only the generated grasping sequence but also a set of additional pelvis positions from which the object can be feasibly grasped. This augmentation is crucial for training generative models that need to handle a variety of valid starting configurations.

To augment the original pelvis position, we follow a four-step filtering process to identify new, plausible grasping locations:

1. **Initial Sampling:** We randomly sample 5,000 points within a 1-meter radius sphere centered on the object. We discard points that are located above the receptacle or are too far from the ground, as they represent semantically implausible grasping positions.

2. **Height Filtering:** We extract the pelvis height from the generated grasping pose and only keep sampled points within a fixed height range of the original pelvis height along the y-axis. This ensures the new positions are at a realistic height for grasping the object.

3. **Collision Filtering:** We create a standing-room cuboid centered at each remaining point. By checking for intersection between these cuboids and the scene's geometry, we discard any positions that would result in a human-scene collision.

4. **Final Sampling & Orientation:** From the remaining valid positions, we uniformly sample 10 points to ensure diversity. Each of these points serves as a new, collision-free pelvis position. The orientation of the human at each new position is defined by the vector pointing from the sampled pelvis position to the center of the grasped object, ensuring the human faces the object they intend to grasp.

This process provides a robust set of augmented grasping positions, enriching our dataset for training and evaluation of scene-aware grasping models.

# B FLEX Modifications

This section details the technical modifications made to **FLEX** [5] that were critical for adapting it to generate scene-aware, full-body grasps. Our adjustments were designed to improve both the efficiency and physical plausibility of the generated poses, making the method suitable for large-scale data synthesis.

We introduced the following key changes to the FLEX optimization framework:

- **Reduced Optimization Iterations:** We decreased the number of optimization steps, which significantly improved computational efficiency without compromising grasp quality.

- **Scene-Aware Grasp Pose Selection:** Unlike the original FLEX, which is scene-agnostic, our approach explicitly accounts for environmental constraints during grasp pose selection. This ensures that all generated grasps are physically feasible within the surrounding scene.

- **Post-Optimization Refinements:** We added two refinement steps to enhance the realism and stability of the generated grasping poses:

  - **Wrist Alignment Correction:** This post-processing step ensures a natural wrist orientation by aligning the wrist with the target object, which avoids unrealistic rotations and misaligned contacts.

  - **Contact Consistency Check:** We introduced a contact distance threshold to detect and resolve minor misalignments, thereby enforcing stable contact points and preventing compensatory body motions.

These modifications allow our approach to produce high-quality, scene-compliant grasps with reduced inference time compared to FLEX's default settings. This makes our method better suited for large-scale dataset generation and real-time applications.

# C Additional Details of Experiments

This section provides a comprehensive overview of our experimental setup, detailing the implementation specifics of our method and the adjustments made to baseline methods to ensure a fair comparison.

## C.1 Implementation Details

**GNet++ Data Preparation:** GNet++ generates static full-body grasps that avoid scene intersections. For training, we collected all frames from the MOGRAS dataset where the subjects stably grasp objects with their right hand, with no intersection between the human and the scene. This resulted in 11.4k frames for the training set, 1.8k for the testing set, and 932 for the validation set. We did not apply any augmentation during training.

**GNet++ Architecture:** GNet++ features a cVAE architecture that generates static full-body grasps within a scene, conditioned on the specified object, its location, and the surrounding scene. The encoder first processes the human grasp and encodes it into an embedding space, while a pretrained ViT [1]encoder processes the given scene. The decoder then samples from these embeddings and outputs SMPL-X [2] parameters ($\hat{\Theta}$), head direction ($\hat{q}$), and hand offset vectors ($\hat{d}^{h\to o}$). These features are used to refine the predicted SMPL-X parameters, resulting in a more realistic full-body grasp. We trained the model for 30k steps with the Adam optimizer (learning rate: 1e-4, batch size: 128) on a single NVIDIA GeForce RTX 2080 Ti GPU, taking approximately 12 hours.

**Adjustments to Baseline Methods.** To ensure a fair comparison, we conducted a careful evaluation of the baseline methods **GOAL** [4] and **SAGA** [5]. As these models were originally trained on the **GRAB** [3] dataset, they are inherently scene-agnostic. We evaluated them "out-of-the-box" on the MOGRAS test set without any fine-tuning to reflect their original design and to directly assess their performance on our scene-aware task. The results from this initial comparison [mentioned in Table 4 and Figure 5 of the main paper] clearly highlight the domain gap and the limitations of these methods in handling scene constraints.

However, to provide a fairer baseline and isolate the impact of our scene-aware components, we also trained a variant of our model, **GNet++ (w/o scene input)**, on the MOGRAS dataset. This model uses the same architecture as our proposed GNet++ but with the scene-conditioning branch removed. This ablation serves as a strong baseline, as it is trained on the same data as our full model but lacks the ability to reason about the scene. The superior performance of our full GNet++ model over this ablation [Table 4 of the main paper] confirms that our scene-conditioning approach is the primary reason for our improved performance, rather than just a benefit of fine-tuning on MOGRAS.

## C.2 Ground-Truth Contact Annotation

To define ground-truth contact in the MOGRAS dataset, we use a distance-based threshold. For each frame, we compute the closest distance from every human vertex to the object surface. A human vertex is considered in contact with the object if its distance to the object is less than a predefined threshold of **2 cm**. This threshold accounts for minor inaccuracies in mesh alignment while ensuring that contact is physically plausible. For the grab data we use the semantic hand annotations as defined in the original GRAB [3] dataset. Similarly, foot-floor contact is defined as any vertex on the feet or lower leg with a distance to the ground plane ($z = 0$) less than **2 cm**.

## C.3 Analysis of Contact Precision and Recall

Our evaluation reports both **contact precision** and **recall** to provide a nuanced understanding of model performance.

- **Precision:** A high precision indicates that when a model predicts a contact point, it is likely to be a true positive. Our models, particularly **GNet++**, achieve high precision, suggesting they are good at avoiding false-positive contacts (e.g., predicting a hand is touching a part of the object it's not actually holding).

- **Recall:** Recall measures the model's ability to find all true contact points. A lower recall indicates that the model is missing some real contact points. This often happens because the model might be too conservative, or it fails to predict contacts that are present in the ground truth but are subtle (e.g., a pinky finger lightly touching the side of an object).

In our experiments, the ablation study (Table 4) shows that removing the penetration loss and scene conditioning significantly impacts recall. This suggests that without proper context and collision avoidance, the models struggle to maintain the full, intricate contact points necessary for a stable, realistic grasp.

## C.4 Human Study

We conducted a human study with 75 participants (38 males, 37 females) to compare the quality of the generated full-body grasps. For each method (GOAL [4], SAGA [6], and GNet++), we rendered 10 human grasp poses in the same scene from the MOGRAS test set. To ensure a fair horizontal comparison, we carefully minimized visual obstructions in the renderings.

Participants were shown 30 images of human grasps synthesized by our method and the baselines. They were asked to rate the full-body grasps on a 1-to-5 scale based on two criteria:

- **Full-body grasp quality:** 1 = poor grasp, 5 = good grasp.

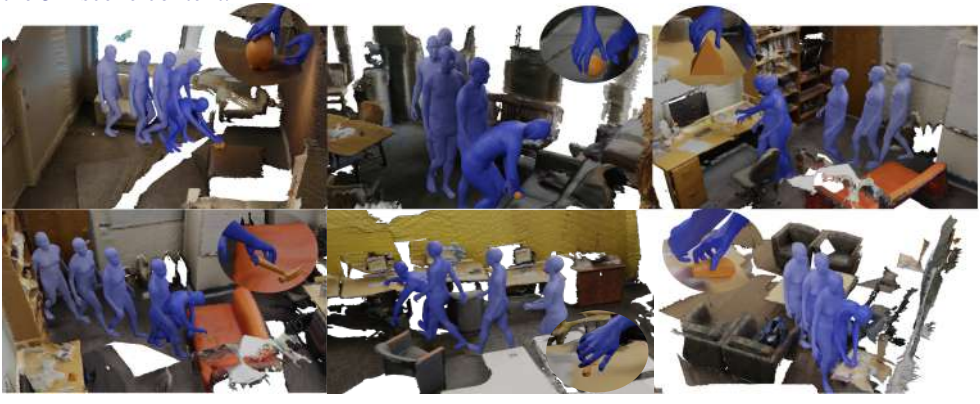
- **Body-scene intersection:** 1 = significant intersection, 5 = no intersection.

Table 1: **Comparison of Grasp Quality and Scene Intersection Across Methods:** Average participant ratings from a human study comparing the full-body grasp quality (Average Grasp Rating) and the extent of scene intersection (Average Scene Intersection Rating) for GOAL, SAGA, and GNet++ methods on the MOGRAS dataset. Higher scores indicate better performance, with GNet++ achieving the highest ratings in both categories.

| | Average Grasp Rating | Average Scene Intersection Rating |
|---|---|---|
| **GOAL [4]** | 3.39 | 3.00 |
| **SAGA [6]** | 2.79 | 2.63 |
| **GNet++** | **3.42** | **4.00** |

The average ratings for full-body grasp quality and scene intersection are presented in Table 1 as the Average Grasp Rating and the Average Scene Intersection Rating, respectively. The results show that users consistently rated GNet++ higher in both criteria, suggesting superior full-body grasp quality and better scene compliance.

Figure 1: **Examples from our proposed dataset showcasing full-body object interaction within 3D indoor scenes.** The samples demonstrate the pre-grasping walking motion, the grasping pose, and the hand-object contact. The dataset captures the intricate interactions between humans and small objects while adhering to the constraints of the surrounding scene, addressing the limitations of existing datasets that either neglect small object interactions or the 3D scene context.



## References


[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.