Task-Oriented Human Grasp Synthesis via Context- and Task-Aware Diffusers

An-Lun Liu[1] Yu-Wei Chao[2] Yi-Ting Chen[1]

National Yang Ming Chiao Tung University[1] NVIDIA[2]

{liuallen871219.cs10,ychen}@nycu.edu.tw ychao@nvidia.com

Abstract

In this paper, we study task-oriented human grasp synthesis, a new task aiming at synthesizing human grasps that require the awareness of its task and context. At the core of our method is the task-aware contact maps. Unlike traditional contact maps that only reason about the object itself and its relation with the hand, our enhanced maps take into account scene and task information. This comprehensive map is critical in hand-object interaction, leading to accurate grasping poses that align with the task. We proposed a two-stage pipeline that first constructs a task-aware contact map informed by the scene and task. In the subsequent stage, we use this contact map to predict task-oriented grasping poses. To validate our approach, we introduced a new dataset for task-oriented grasp synthesis. Our experiments demonstrate the superior performance of our approach, surpassing existing methods on both grasp quality and task performance.

1. Introduction

In this paper, we explore task-oriented human grasp synthesis, which aims to generate grasps that consider environmental context and manipulation goals.

We construct a new task-oriented grasp dataset to support the development and evaluation of this problem. We select three everyday tasks—placing, stacking and shelving—because they require spatial awareness to avoid collisions with nearby objects, as well as to discern the object's affordance relevant to the task. We have selected 104 daily objects from DexGraspNet [15], which include items like bottles, jars, stationery, toys, food, shoes, and 3C electronics for the task of **Placing and Shelving**. Additionally, we have created 23 distinct bricks, each derived from fundamental geometric shapes, for the task of **Stacking**. For each task, we establish a systematic pipeline to generate ground truth human grasps. Overall, our dataset contains 571,908 task-oriented human grasps for placing, 2,989 human grasps for stacking, and 807,028 for shelving.

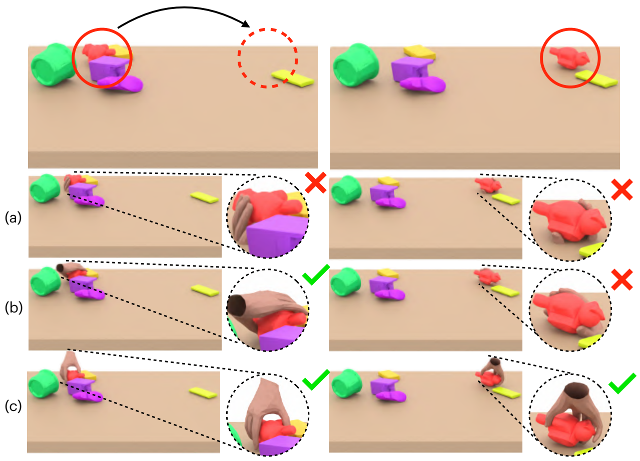

Figure 1. The Task-oriented Human Grasp Synthesis aims to anticipate human-like, collision-free, and task-aware grasping poses given the initial and intended goal scene. In this work, we introduce **Task-aware contact map**, an enhanced contact representation that specifies **where** and **how** to grasp an object depending on various tasks and environments. We designed 3 daily scenarios: placing; handover; and stacking to evaluate predicted grasping poses. Our experiments show significant improvement in past work.

We highlight two challenges that current human grasp synthesis algorithms [4, 8, 10, 11] for task-oriented human grasp synthesis. First, current techniques prioritize stable contact with an object's surface through object-centric representations, such as object affordance represented by contact map[4, 8, 10, 11]. However, object affordance does not account for object-environment interaction, which is vital for lifting and stacking bricks. Therefore, object-centric methods may struggle or fail in cluttered scenes. Second, current techniques do not consider downstream tasks. Hence, the synthesized grasps, while they maybe collision-free at the initial grasping stage, can result in scene collision at task completion.

To this end, we propose a new two-stage diffusion-based framework, driven by the proposed task-aware contact map, to address the synthesis of task-oriented human grasps. A task-aware contact map incorporates crucial information about the context and the task. The two-stage diffusion-based framework consists of (1) **ContactDiffuser**, which predicts a task-aware contact map for an object, given the point clouds of the initial and goal scenes along with the initial and goal distance maps, and (2) **GraspDiffuser**, which synthesizes human grasps from the predicted task-aware contact map and the object's point cloud. We conduct extensive experiments on the proposed dataset and demonstrate that our framework surpasses strong baselines on grasp synthesis in terms of physical plausibility, and collision avoidance.

2. Related Work

Human Grasp Synthesis. The study of human grasp synthesis is segmented into three distinct streams: object model-based grasps [4, 8–11, 13] emphasize optimal hand-object contact and allow approaching the object from any direction, functional grasps [2, 7, 16] consider an object's affordance, requiring a grasp on the object's part that enables its intended use, and scene-aware grasps [3, 14] account for the object's surroundings to prevent collisions and enable effective grasp synthesis in cluttered environments. In sum, object model-based grasps establish a foundation by concentrating on contact points. Functional grasps broaden this approach by identifying how to hold an object for its intended use. Scene-aware grasps extend model-based grasps by factoring in contextual information by including nearby objects, ensuring the grasp is aware of the environment. To the best of our knowledge, we are the first work to study task-oriented human grasp synthesis, which encompasses obstacle awareness, environmental context, and the intended task.

Object Affordance. ContactDB [1] introduce a dataset containing contact maps for household objects, which effectively capture the intricate hand-object contact in grasping and handover scenarios via the use of a thermal camera. AffordPose [7] explores affordance-aware hand-object interactions, offering part-level affordance annotations for each object, such as twist, pull, and handle-grasp, expanding on the intentions explored in ContactDB. The work mentioned above attempts to map specific grasping poses with tasks. We observe that the mapping becomes challenging for tasks like placing or stacking, where the same task can have varying goals, influencing how an object should be grasped. Besides, grasping in unconstrained settings needs to handle cluttered scenes with various objects. Depending on the task and goal, humans determine where to interact with or how to avoid colliding with their surroundings. In this work, we propose a task-aware contact map that considers the scene, the task, and the goal jointly.

3. Task-oriented Human Grasp Dataset

3.1. Task Description

We design three daily tasks for evaluating task-oriented human grasp synthesis. We employ PyBullet as our physics simulator to generate diverse task configurations. Task configuration is composed of initial and goal position of target object and position of obstacles.

3.2. Data Generation

3D Human Hand Model. We adopt MANO [12], a differentiable 3D human hand model. With a mesh has 778 vertices and 1538 faces, MANO provides a comprehensive representation and can be integrated into training pipelines.

Object and Scene. We select 104 objects from DexGraspNet [15], rescaling them to be graspable with one hand. These objects are used in placing. For stacking tasks, many of these objects aren't appropriate. Thus, we create 24 distinct bricks based on simple geometry that are conducive to stacking.

Task-oriented Human Grasps Generation. To produce high-quality human grasps, we utilize DexGraspNet [15] to generate human grasps. However, it may produce grasps that are not physically plausible or human-l ike. To mitigate this, we filter out inferior grasps using penetration volume and simulation displacement. Both metrics are frequently used in grasp synthesis studies [8, 9, 9, 11, 13] to assess the quality of grasp poses. We set the thresholds at $4 \times 10^{-6}\,\mathrm{cm}^3$ and $3\,\mathrm{cm}$, respectively, to initially filter out lower quality grasps. Subsequently, we manually remove any grasping poses that do not appear human-like. Based on specific task configurations, we adjust the hand's position, transitioning it from its initial state to its intended goal state. During this transition, we conduct a thorough collision detection analysis involving all objects present in the scene, except for the object currently being grasped by the hand. If, throughout this process, the hand remains free from any collisions, we consider the trial as successful and save the configuration.

4. Methodology

Fig. 2 shows an overview of our pipeline with two key components: ContactDiffusers for task-aware contact map prediction, and GraspDiffusers for human grasp synthesis.

4.1. Problem Formulation

Given a 3D initial scene S_i and goal scene S_g, we aim to predict task-oriented human grasp G for accomplishing a desired task. Both scenes contain the target object O to be grasped. Our pipeline consists of two stages. In the first stage, ContactDiffusers generates task-aware contact map C_{task} given S_i and S_g. $C_{\text{task}} = D_{\text{contact}}(S_i, S_g)$. In the second stage, GraspDiffusers synthesizes task-oriented human

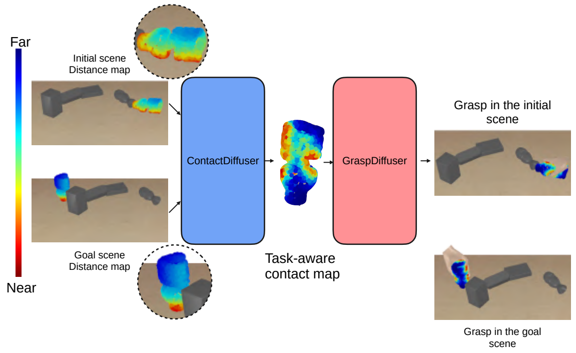

Figure 2. An overview of our proposed pipeline.

grasp with the target object O and the task-aware contact map, $G = D_{\text{grasp}}(O, C_{\text{task}})$.

We present an improved representation of contact maps, termed task-aware contact maps. Unlike traditional object-centric contact maps [4, 8, 10, 11] that primarily focus on predicting suitable contact with the object's surface for grasping. This object-centric modeling is insufficient in real-world scenarios. There are two key factors for a task: the affordance of the object, which dictates how humans grasp different parts of the object to achieve various goals or tasks, and the context of the scene, which often contains multiple objects. Depending on the specific tasks and objectives, humans decide whether to interact with these objects or avoid collisions. To capture the information relevant to the context and task, we use **distance map**, represented by D. This map can be obtained by computing the shortest distance between the target object and the scene, thereby implicitly incorporating the information of surroundings into objects. When engaging in grasping, we prioritize contact with regions on the object that are further from any obstacles, thereby minimizing the possibility of collision with nearby obstacles. Beyond addressing concerns of collision, the distance map also provides insights into the ways objects interact with their environment, reflecting their state or goal. We can obtain two maps, D_{init} and D_{goal} using the initial scene and the goal scene. By incorporating this additional information, our task-aware contact map becomes a more sophisticated and effective instrument for modeling hand-object interaction.

4.2. Context- and Task-aware Diffusers

To generate diverse and realistic human-object interaction, we leverage the power of the diffusion model [5]. **ContactDiffuser.** We follow recent work on human motion generation [6] and use Transformer encoder-based model. This model processes the current noisy contact map x_t, the time step t, and a set of conditioning variables C. The conditioning C includes the object's point cloud, along with initial and goal distance maps, denoted as D_{init} and D_{goal}, respectively. Given that the prediction of contact maps is a per-point task, we utilize PointNet++ to derive local features. **GraspDiffuser.** We first employ self-attention on noisy MANO parameters and object feature to discern the relationships among the different joints and object feature. Subsequently, we apply cross-attention to establish the correspondence between the object and the MANO parameters. This allows for a more nuanced understanding of the interplay between the object's geometry and the hand model, leading to more accurate and realistic grasps.

5. Experiments

We perform extensive experiments on our dataset. For placing, we test 21 unseen objects. As for stacking, we test on 6 unseen bricks. For every object in each task configuration, we predict 16 grasps for the evaluation. The quality of predicted grasps is evaluated based on their physically plausibility, stability, and collision avoidance. Our experiments aim to answer the following research questions. **Can the proposed method synthesize high-quality task-oriented human grasps?** While there is significant progress in human grasp synthesis, we are interested in identifying the solution gap for the synthesis of task-oriented human grasps.

5.1. Human Grasp Synthesis Baselines

We compare the following baselines in our experiments.
GraspTTA [8]: GraspTTA utilizes CVAE to generate an initial coarse human grasp and obtains the final grasp synthesis through test-time adaptation, guided by a predicted contact map. **Modified GraspTTA [8]**: We modify GraspTTA by augmenting input point clouds with the contact maps. **FLEX [14]**: FLEX is designed to generate 3D full-body human grasps. We re-purpose the method to the generation of human grasps. Note that the optimization is driven by the penetration losses in both initial and goal scenes. **ContactGen [11]**: We train ContactGen to predict the proposed object-centric representation. We then use their method to generate the contact map, hand-part map, direction map, and human grasp accordingly. **SceneDiffuser [6]**: We reimplement SceneDiffuser [6] to predict MANO [12] hand parameters. We do not perform optimization during inference because we apply three auxiliary losses for generating physically plausible human grasps.

5.2. Metrics

We evaluate the synthesized human grasps based on their physical plausibility, stability, and diversity, following prior works [8, 11, 13, 14]. We propose a new metric called **Task Score (TS)** to evaluate the quality of task-oriented human grasp synthesis. **Penetration Volume (PV):** We calculate the penetration volume by converting the meshes into 1mm cubes and calculating the overlap of these voxels. **Simulation Displacement (SD):** We simulate the object and predicted grasps in PyBullet for 1 sec. and then compute

Table 1. **Task-oriented Human Grasp Synthesis Evaluation. PV:** Penetration Volume, **SD:** Simulation Displacement, **CR:** Contact Ratio, **QR:** Qualified Ratio, **OPP:** Obstacle Penetration Percentage, and **TS**: Task Score. The table reports the proposed method can synthesize favorable task-oriented human grasp synthesis, compared to strong baselines. FLEX [14] struggles to synthesize stable grasps in **Stacking** due to its inability to handle small bricks. Please see Fig. 3 for qualitative results.

	Method	PV↓	SD↓	QR(%)↑	Init OPP(%)↓	Goal OPP(%)↓	TS↑
Placing	GraspTTA [8]	1.85	2.60	58.57	21.67	17.91	0.376
	ContactGen [11]	1.40	3.85	46.84	**5.56**	17.26	0.366
	SceneDiffuser[6]	**1.37**	3.19	53.12	20.00	16.91	0.353
	FLEX [14]	2.50	1.62	59.10	6.74	**5.61**	0.520
	Ours	2.40	**1.42**	**65.29**	7.27	5.79	**0.570**
Stacking	GraspTTA [8]	4.30	**0.28**	35.00	26.04	8.32	0.237
	ContactGen [11]	0.66	1.87	76.43	8.42	9.75	0.631
	SceneDiffuser[6]	0.53	1.64	77.72	25.30	9.81	0.523
	FLEX [14]	**0**	10.65	0.00	**0**	**0**	0
	Ours	1.09	1.03	**84.31**	14.94	4.51	**0.684**
Shelving	GraspTTA [8]	1.78	2.56	58.94	15.46	13.43	0.431
	ContactGen [11]	**1.43**	3.90	46.32	6.42	13.17	0.376
	SceneDiffuser[6]	1.38	3.31	51.48	14.52	13.59	0.380
	FLEX [14]	2.81	**1.54**	57.26	**4.39**	**4.47**	0.522
	Ours	2.12	1.62	**67.49**	8.72	10.31	**0.552**

the object's center of mass displacement. **Qualified Ratio (QR):** The metric jointly considers both penetration volume and simulation displacement. Note that, a higher penetration volume generally leads to a lower simulation displacement, which is not satisfactory. We set thresholds at $3 \times 10^{-6}\,\mathrm{cm}^3$ and $2\,\mathrm{cm}$ for penetration volume and simulation displacement, respectively. We calculate the percentage of predicted grasps that satisfy both criteria. **Obstacle Penetration Percentage (OPP):** We compute the penetration percentage of human grasp vertices in obstacles for initial and goal scenes [14]. **Task Score (TS):** A proper metric for task-oriented human grasp synthesis should take grasp physically-plausibility, stability, and collision avoidance in both initial and goal scenes. Thus, we propose a new metric and define it as $\mathrm{TS} = \mathrm{QR} \times (1 - \mathrm{Init\ OPP}) \times (1 - \mathrm{Goal\ OPP})$.

5.3. Results

Can the proposed method synthesize high-quality task-oriented human grasps? We report our empirical studies in Table 1. GraspTTA, ContactGen, and SceneDiffuser often result in lower penetration volumes (PV↓) but struggle with stable grasps (SD↓) in **Placing**. FLEX seeks a balance between penetration volume and simulation displacement. Our method provides the optimal balance, as measured by QR. The **Stacking** task is much harder as the objects are smaller than **Placing**. GraspTTA [8] suffers from unsatisfactory grasp synthesis (**QR**↑) and severe mode collapse (**DS**↑). FLEX [14] struggles with stable grasps in **Stacking** due to its inability to handle small bricks The proposed method demonstrates favorable grasp synthesis (**QR**↑) in the challenging task. For **OPP**, ContactGen [11] shows a lower Init **OPP** than Goal **OPP**, due to its grasp synthesis strategy, which begins with the hand oriented face down towards the ground. Our method achieves the best performance in terms of **TS**, which evaluates the synthesis of human grasps for physical plausibility, stability, and collision avoidance in initial and goal scenes.

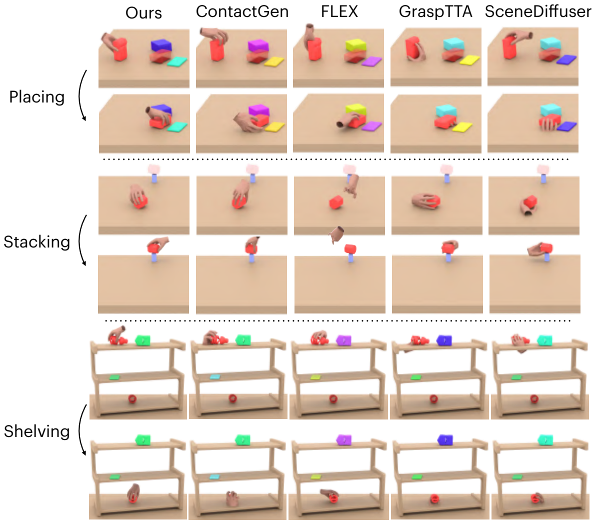

Figure 3. Visualization of predicted grasping poses from our method, ContactGen [11], FLEX [14], GraspTTA [8], and SceneDiffuser [6].

5.4. Qualitative Results

Synthesized Human Grasps. The prediction results of ContactGen [11], GraspTTA[8], and SceneDiffuser collide with the scene severely, as shown in Fig. 3. FLEX [14] synthesize grasps with unrealistic contact and fail to generate a grasp for **Stacking**. In contrast, our method produces high-quality human grasps and can avoid collision with obstacle.

6. Conculsion

In this work, we present a new task called task-oriented human grasp synthesis along with a new dataset for development and benchmarking. We show existing human grasp synthesis algorithm struggle with generating high-quality human grasps for the designed tasks. To this end, we propose a novel two-stage diffusion-based framework powered by the task-aware contact map to incorporate crucial information about the context and the task. We perform comprehensive quantitative and qualitative experiments to validate the effectiveness of our proposed framework in comparison to strong baselines.

References

[1] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8709–8719, 2019. 2

[2] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393. IEEE, 2019. 2

[3] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5031–5041, 2020. 2

[4] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1471–1481, 2021. 1, 2, 3

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[6] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023. 3, 4

[7] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14713–14724, 2023. 2

[8] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11107–11116, 2021. 1, 2, 3, 4

[9] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 2

[10] Haoming Li, Xinzhuo Lin, Yang Zhou, Xiang Li, Yuchi Huo, Jiming Chen, and Qi Ye. Contact2grasp: 3d grasp synthesis via hand-object contact constraint. *arXiv preprint arXiv:2210.09245*, 2022. 1, 3

[11] Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. Contactgen: Generative contact modeling for grasp generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20609–20620, 2023. 1, 2, 3, 4

[12] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2, 3

[13] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020. 2, 3

[14] Purva Tendulkar, Dídac Surís, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21179–21189, 2023. 2, 3, 4

[15] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11359–11366. IEEE, 2023. 1, 2

[16] Tianqiang Zhu, Rina Wu, Jinglue Hang, Xiangbo Lin, and Yi Sun. Toward human-like grasp: Functional grasp by dexterous robotic hand via object-hand semantic representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2