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

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

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

Abstract

In this paper, we study task-oriented human grasp synthesis, a new grasp synthesis task that demands both task and context awareness. At the core of our method is the task-aware contact maps. Unlike traditional contact maps that only reason about the manipulated object and its relation with the hand, our enhanced maps take into account scene and task information. This comprehensive map is critical for hand-object interaction, enabling accurate grasping poses that align with the task. We propose 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 synthesize task-oriented human grasps. We introduce a new dataset and a metric for the proposed task to evaluate our approach. Our experiments validate the importance of modeling both scene and task, demonstrating significant improvements over existing methods in both grasp quality and task performance. See our project page for more details: https://hcis-lab.github.io/TOHGS/

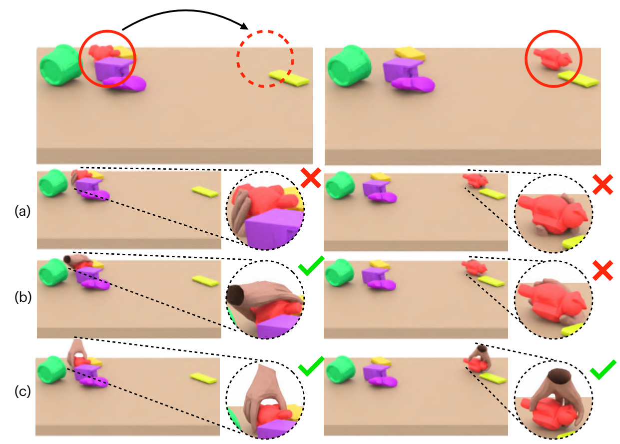

Figure 1. Task-oriented human grasp synthesis aims to synthesize collision-free and task-aware human grasps from initial and goal scene point clouds. Rows (a) and (b) show failures due to a lack of context or task awareness, causing collisions and penetrations. In contrast, (c) demonstrates a successful task-oriented synthesis that considers both scene context and intended task objective.

1. Introduction

Hand-object interaction has been a pivotal topic in computer vision research. We have seen diverse efforts, spanning from hand-object pose estimation [19, 30], 3D hand-object reconstruction [7, 11, 20, 54], human grasp synthesis [15, 24, 25, 28, 32, 42, 45, 49], affordance modeling [18, 28, 32, 39, 55], hand-object interaction dataset [3, 4, 9, 15, 17, 19, 20, 23, 42, 53], to hand motion trajectory [31, 44]. Predicting realistic and viable hand-object interactions opens up new applications in augmented reality [41], robotics [1, 2, 27, 29, 33, 36, 48, 49, 51, 52, 59], and human-robot interactions [10, 14, 50].

Despite significant progress in hand-object interaction, limited attention has been given to grasp motion synthesis that both aligns with task objectives and incorporates scene contextual information. Recent notable work [5, 13, 43, 56–58] focus on generating plausible human grasp motions for a variety of objects, typically with a single objective, such as moving the object to a desired location [5, 13, 58]. GraspXL [56] highlights the importance of incorporating multiple motion objectives, such as the desired heading directions and wrist rotations, while ensuring physical plausibility of the motion. ArtiGrasp [57] extends this by considering desired states of objects, such as opening a cabinet door to a target articulation angle or keeping its base steady in its initial pose. While these approaches show promise in synthesizing grasp motions for specific objectives, they do not consider environmental contexts, such as grasping in a cluttered space for an object relocation task. This limits their applicability to object-singulated manipulation tasks in less complex environments.

To bridge this gap, we study task-oriented human grasp synthesis, a new endeavor that seeks to generate human grasps informed by both scene context and task objectives. Fig. 1 illustrates a *relocation* task, constituting of grasping a target object and relocating it to a designated location. The task presents two unique challenges. First, a model needs to possess scene context awareness to ensure that the syn-

*Corresponding Author

thesized grasp avoids collisions with its surroundings. As shown in the left image of Fig. 1 (a), although the synthesized grasp is valid for the object itself, the lack of contextual awareness causes the hand to penetrate the table during the pickup. Second, the model needs to take the downstream task into account. In Fig. 1 (b), although the synthesized grasp is successful for the pickup (left image), the lack of task awareness causes the hand to collide with the table during the placing pose (right image). Finally, Fig. 1 (c) illustrates a successful synthesis of a task-oriented human grasp that incorporates both the context and task at hand.

In this work, we first construct a new task-oriented grasp dataset to support the development and evaluation of the problem. Despite notable advancements in human grasp synthesis, existing datasets [3, 4, 9, 15, 17, 19, 20, 23, 42, 53] remain largely focused on object-centric grasp synthesis, overlooking the needs of downstream hand manipulation tasks, such as stacking objects onto one another. We focus on three everyday tasks—**Placing**, **Stacking**, and **Shelving**—all of which require scene awareness to avoid collisions with nearby objects, and task awareness to understand the object's affordance. For **Placing** and **Shelving**, we select 104 everyday objects from DexGraspNet [49] including items like bottles, jars, stationery, toys, food, shoes, and 3C electronics. Additionally, we create 23 distinct bricks, derived from fundamental geometric shapes, for the **Stacking** task. For each task, we establish a systematic pipeline for generating ground truth human grasps. Overall, our dataset contains 571,908 task-oriented human grasps for **Placing**, 2,989 for **Stacking**, and 807,028 for **Shelving**.

Existing human grasp synthesis approaches [18, 24, 28, 32] fall short for the proposed task due to two key challenges. First, object-centric representations, such as object affordances represented by contact maps [18, 24, 28, 32], fail to account for object-environment interactions, making them prone to collisions in cluttered scenes. For example, picking and stacking objects in Fig. 1 using object-centric representations would result in collisions with surrounding objects, even if the methods produce stable contacts. Second, current techniques overlook downstream tasks. While the synthesized grasps may be collision-free initially, they can still lead to scene collisions during task execution.

To account for both scene and task contexts, we propose a novel two-stage diffusion-based framework. The framework consists of (1) a **ContactDiffuser**, which predicts a new representation called task-aware contact map, given the point clouds of the initial and goal scenes along with their corresponding distance maps, and (2) a **GraspDiffuser**, which synthesizes human grasps based on the predicted task-aware contact map and the object's point cloud. The rationale for using diffusion models is that task-aware contact maps and their corresponding task-oriented human grasps are inherently multimodal, given the diverse combinations of environmental contexts and tasks.

We conduct extensive experiments on the proposed dataset and demonstrate that our framework, with explicit task-aware contextual awareness, outperforms strong baselines in both physical plausibility and collision avoidance. Additionally, we empirically show that task-aware contact maps are more effective at extracting task-relevant information than object-centric contact maps. Our qualitative analysis reveals that the ContactDiffuser generates high-quality contact maps, while the GraspDiffuser produces natural and realistic human grasps. Furthermore, we perform comprehensive ablation studies to validate the significance of the two proposed diffusers. Although our work does not involve motion synthesis directly, it represents a crucial step in incorporating task-aware scene awareness into human grasp synthesis. Moreover, the generated initial and target human grasps can serve as hand pose references to improve motion synthesis, as demonstrated in [13, 57].

Our contributions are summarized as follows. **First**, we introduce a new task, task-oriented human grasp synthesis, along with a new dataset for its development and benchmarking. **Second**, we propose a novel two-stage diffusion-based framework that leverages a new representation, the task-aware contact map, to integrate essential scene and task-specific information. **Third**, we conduct comprehensive experiments to show the effectiveness of our framework over strong baselines.

2. Related Work

Object Affordance. Object offordance has been extensively studied in the community. Zhu et al. [60] introduce a task-oriented object representation that includes affordance basis, functional basis, imagined actions, and physical concepts. The affordance basis and functional basis indicate the area to be grasped by the hands and the part intended to act on a target object, respectively. The authors of ContactDB [3] present a dataset that records contact maps, capturing how humans interact with household objects in grasping and handover scenarios using a thermal camera. AffordPose [23] provides part-level affordance annotations for each object, such as twist, pull, and handle-grasp, to enable fine-grained affordance understanding. The works above aim to ground task-relevant information onto objects to facilitate the generation of corresponding grasps. However, it becomes challenging for tasks like placing or stacking, where goal scene situations (e.g., cluttered scenes) influence how an object should be grasped. Rather than predefining affordances as in previous works, we learn task-aware contact maps that simultaneously account for the scene, task, and goal.

Diffusion Models. Diffusion models [21] have emerged as a powerful generative model and have been leveraged widely across various fields, including generation of im-

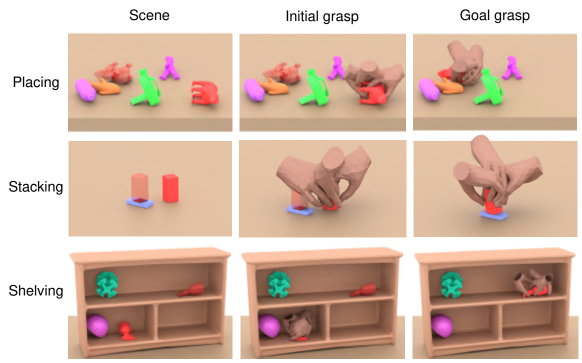

Figure 2. Examples of **Placing**, **Stacking**, and **Shelving**.

ages [38], video [6], human motions [46], 3D objects [34], and robot actions [12]. Recently, Huang et al. [22] propose SceneDiffuser, a diffusion model-based framework for 3D scene-conditioned human motion, robot motion, and dexterous robotics hand motion generation. Inspired by their work, we devise diffusion models for task-oriented human grasp synthesis.

Our work differs from SceneDiffuser in two key aspects. First, while SceneDiffuser focuses on generating dexterous robotic hand grasps for object pick-up tasks, our approach targets a range of everyday tasks, such as stacking and shelving. Second, we explicitly model task-aware scene awareness in human grasp generation, which we demonstrate as essential for producing collision-free grasps that successfully complete tasks. Additionally, our work complements SceneDiffuser, as our framework can serve as an intermediate understanding step for generating a sequence of hand motions across various tasks in future works.

3. Task-Oriented Human Grasp Generation

3.1. Problem Formulation

Given an initial scene point cloud $S_{\texttt{init}}$ and a goal scene point cloud $S_{\texttt{goal}}$, we aim to synthesize task-oriented human grasps G that avoid collisions with surrounding objects and successfully accomplish the desired task. Both scenes contain the target object to be grasped. The proposed task is challenging, as it requires modeling both the environmental context and the task simultaneously.

3.2. Task Specification

We propose three common tasks to instantiate the proposed problem. For each task, we generate diverse configurations by simulating the scenes with a physics engine (PyBullet [16]). The three tasks are illustrated in Fig. 2.

Placing. A target object is placed on a cluttered table and should be picked and moved to a new location on the same table. We randomize the initial and desired positions of the target. First, the object is dropped from a height of 10 cm above the table, and the simulator runs for 5 seconds to allow it to settle into a stable pose. Next, obstacles from our dataset are randomly selected and positioned near the initial and target locations. If no obstacles collide with the grasped object, the configuration is considered valid.

Stacking. Two objects—a base object and an object to be stacked—are placed on a table. The goal is to stack the latter on top of the former. To create diverse configurations, we randomly select bricks from our dataset, assigning one as the base and the other as the object to be stacked. After positioning the stacked object on the base, the simulation runs for 5 seconds. If both objects remain stable at the end, the configuration is considered valid and is saved.

Shelving. A target object is placed on a shelf and should be moved to a different location on the same shelf. The initial and goal positions are selected randomly. The object is then dropped from 5 cm above its initial position, and the simulation runs for 5 seconds to allow the object to settle. Finally, obstacles are randomly placed on the shelf. If no obstables collide with the object, the configuration is considered valid.

3.3. Dataset

Human Hand Model. We use the MANO hand model [37], which is a 3D mesh consisting of 778 vertices and 1538 faces. The model is controlled by a 10-dimensional parameter β for hand shape and a 51-dimensional parameter θ for joint rotations and root translation. We use a uniform β for all the human grasps.

Object and Scene. We select 104 objects from DexGraspNet [49] and rescale them to be graspable by a hand. These objects are used in the **Placing** and **Shelving** tasks, but many of them are not suitable for **Stacking**. Consequently, we create 23 distinct brick models with simple geometries specifically designed for **Stacking**. We assume all objects are asymmetrical. The shelf models used in **Shelving** are selected from ShapeNet [8].

Dataset Construction. We use DexGraspNet [49] to generate high-quality human grasps. This method leverages a differentiable force closure estimator to efficiently produce diverse and stable grasps at scale. However, it may generate grasps that are not physically plausible or human-like. To mitigate this, we automatically filter out suboptimal grasps based on penetration volume and simulation displacement—common metrics for assessing grasp quality in grasp synthesis [24–26, 32, 42]. We set a threshold of 4×10^{-6} cm^3 for penetration volume and 3 cm for displacement. Subsequently, we manually filter out non-human-like grasps based on joint angles and grasp stability. Finally, we refer to the remaining grasps as prior poses.

For each target, we sample initial and goal positions, along with their corresponding grasp poses from the prior

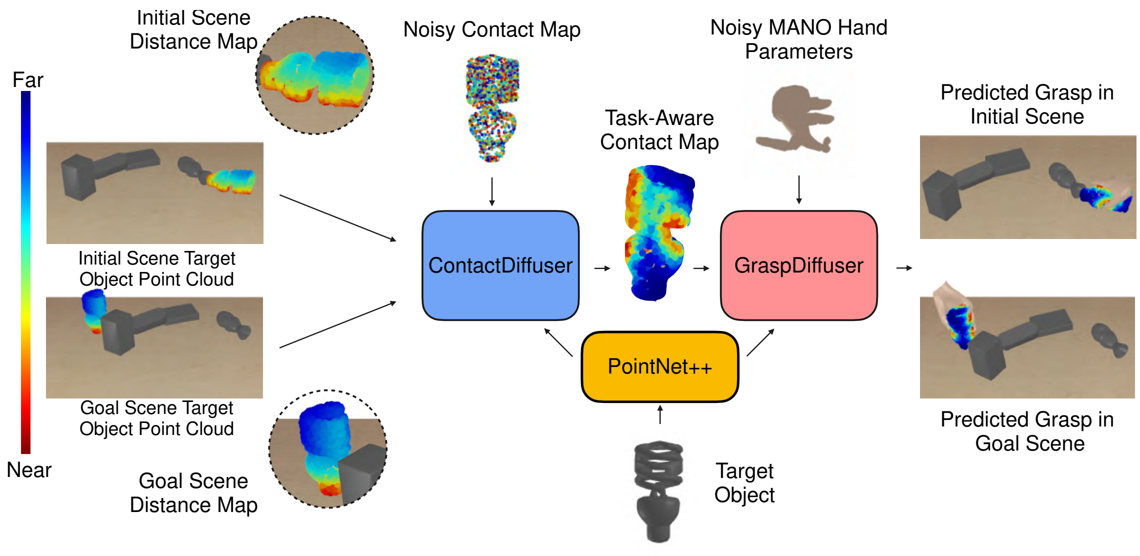

Figure 3. An overview of our proposed task-oriented human grasp synthesis framework. Given 3D initial and goal scenes, our goal is to synthesize task-oriented human grasps that avoid collisions with surrounding objects and achieve the desired task. To capture context- and task-relevant information, our approach integrates the two distance maps, which calculate the shortest distance between the target object and its surroundings in the initial and goal scenes. In addition, we propose to learn a new intermediate representation called task-aware contact maps, integrating environmental context and task-relevant cues to facilitate human grasp synthesis.

pose set. Then, we perform collision checks with all the objects in the scene, excluding the object currently grasped by the hand. If the hand remains collision-free throughout the trial, the configuration is considered valid and is saved. This process expands the prior pose set to over 1.38 million task-oriented grasps.

4. Methodology

The overall framework is shown in Fig. 3. Given the initial scene point cloud $S_{\texttt{init}}$ and the goal scene point cloud $S_{\texttt{goal}}$, our goal is to synthesize a task-oriented human grasp $G = \{\theta, \beta\}$. The full pipeline consists of two stages. In the first stage, we take the input point clouds and predict an intermediate representation, referred to as the task-aware contact map. In the second stage, the task-aware contact map is used to predict the grasp pose. We treat both stages as generative problems and solve them via two separate diffusion models: **ContactDiffuser** and **GraspDiffuser**.

4.1. Input Preprocessing

Our approach is build upon two types of point cloud based representations: *distance maps* and *task-aware contact maps*. A *distance map* encodes the distance from each point on an object to its closest points in the surrounding scenes. We use two distance maps (i.e., from the initial and goal scene respectively) as the input representation for **ContactDiffuser**. Given input point clouds, we randomly sample N points from a target object mesh, where N is set to 2048 following [22] to obtain the corresponding target object point cloud $O \in \mathbb{R}^{N\times 3}$. To acquire scene point clouds, we merge all points of obstacles (including the table) and perform farthest point sampling to N_s points, where N_s is 6000. We compute the point-to-point shortest distance between a target object point cloud and the scene point clouds. We normalize all distance values d to the range of $[0, 1]$. To stretch the range of all distance values, we rescale the values via the equation $D = 1 - 2 \times \texttt{sigmoid}\big((\alpha \times d) - 0.5\big)$, where the parameter α is set to be 30 in our experiments. Through the above process, we obtain the two distance maps: the initial scene distance map $D_{\texttt{init}} \in \mathbb{R}^{N\times 1}$ and the goal scene distance map $D_{\texttt{goal}} \in \mathbb{R}^{N\times 1}$. On the other hand, a *task-aware contact map* $\mathcal{C}_{\texttt{task}} \in \mathbb{R}^{N\times 1}$ encodes the distance from each point on the object to its closest point on the hand mesh model in a grasp pose. We apply the same process to hand meshes and object point cloud, with an α of 100, to generate ground-truth task-aware contact maps for training and evaluation.

4.2. Contact- and Task-aware Diffusers

We use diffusion models [21] for predicting task-aware contact maps and task-oriented human grasps. While previous work has applied diffusion models to learn multi-modal distributions for other domains, we adapt them for grasp synthesis and specifically investigate their effectiveness in predicting our proposed representations.

ContactDiffuser. To capture context- and task-relevant information, our approach leverages the two distance maps, i.e., $D_{\texttt{init}}$ and $D_{\texttt{goal}}$, as valuable guidance for generating task-aware contact maps. Task-aware contact maps integrate environmental context (e.g., multiple objects) and task-relevant details (e.g., relocating one object onto another) to facilitate human grasp synthesis. The representation contrasts with existing object-centric maps [18, 24, 28, 32], which primarily focus on predicting suitable contact points on an object's surface for grasping.

We adapt the architecture, a Transformer-based model, proposed in [22] for task-aware contact map prediction. This architecture processes the current noisy contact map x_t, timestep t, and a set of conditioning variables C. The conditioning C includes the target object's point cloud and the initial and goal distance maps. We concatenate each point's initial and goal distances and 3-dimensional coordinates to form a 5-dimensional point cloud. We then utilize PointNet++ [35] to encode the point cloud and extract the corresponding local features. However, the point-order invariance of PointNet++ presents a challenge in mapping features between the point cloud and the contact map. In natural language processing, sinusoidal positional encoding is often added to retain the sequence order of elements. Instead, we incorporate extracted features as positional embeddings, ensuring that the current noisy contact map x_t can be guided to generate the corresponding task-aware contact maps in the denoising process.

GraspDiffuser. GraspDiffuser takes as inputs the noisy MANO parameters [33], target object features extracted via PointNet++, and predicted task-oriented contact maps $\hat{\mathcal{C}}_{\texttt{task}}$. First, we concatenate $\hat{\mathcal{C}}_{\texttt{task}}$ with the target object features to form task-aware object features $\mathcal{O}_{\texttt{task}}$. We then apply self-attention [47] separately to the noisy MANO parameters [33] and task-aware object features $\mathcal{O}_{task}$, allowing us to capture the spatial relationships among the various joints and object features, respectively. Subsequently, we apply cross-attention to establish the correspondence between the object features and the MANO parameters, enabling a more nuanced understanding of the interplay between the object's geometry and the hand model. The design leads to a more accurate prediction of human grasp. The detailed architectures of ContactDiffuser and GraspDiffuser can be found in the Supplementary Material.

4.3. Loss Function

We train the two networks independently. The ContactDiffuser network leverages the simplified training objective, denoted as $\mathcal{L}(\theta)$, proposed by Ho et al., [21]. For the GraspDiffuser network, we follow previous works [24, 26, 28, 32] that incorporate three additional auxiliary losses to synthesize realistic and physically plausible hand grasp synthesis. (1) **Reconstruction Loss**: the first objective is the mesh reconstruction error. Specifically, we calculate the error between the ground truth hand mesh and the predicted hand mesh generated using the MANO hand model. The loss is denoted as $L_{\texttt{recon}} = \|\hat{M} - M\|^2$. (2) **Penetration Loss**: We denote object points that are inside a hand mesh as $O_{\texttt{in}}$. The penetration loss is defined as $L_{\texttt{penetr}} = \frac{1}{|O_{\texttt{in}}|}\sum_{p \in O_{\texttt{in}}} \min_i \|p - \hat{V}_i\|_2$, which calculate the distance between an object's point and the closest point on hand vertices. (3) **Consistency Loss**: We enforce that the generated grasps align with the regions the task-aware contact map indicates as graspable. The consistency loss is defined as $L_{\texttt{cons}} = \|C - C'\|^2$, where C is the ground truth contact map. To obtain C', we first calculate the shortest distance between each vertex of the target object O and the vertices of the predicted hand mesh $\hat{M}$, then normalize these distance maps to form C', with ranges $[0, 1]$. The overall training loss of the framework is defined as follows: $L = w_1 \cdot \mathcal{L}(\theta) + w_2 \cdot L_{\texttt{recon}} + w_3 \cdot L_{\texttt{penetr}} + w_4 \cdot L_{\texttt{cons}}$, where the parameters w_1, w_2, w_3, and w_4 are set empirically as 15, 1, 5, and 0.002, respectively, for **Placing** and **Shelving**. For **Stacking**, they are set as 10, 1, 3, and 0.005, respectively.

5. Experiments

Our experiments aim to answer the following questions. **Can the proposed method synthesize high-quality task-oriented human grasps?** Despite significant progress in human grasp synthesis [22, 24, 32, 45], we aim to highlight the remaining solution gap. **Is the task-aware contact map crucial for task-oriented human grasp synthesis?** Modeling contextual and task information is challenging, and we aim to determine whether the proposed task-aware contact map can effectively capture both. **Do the proposed ContactDiffuser and GraspDiffuser play crucial roles in synthesizing high-quality task-oriented human grasps?**

5.1. Baselines

Human Grasp Synthesis. We compare the following human grasp synthesis baselines with GraspDiffuser in our experiments.

GraspTTA [24]: GraspTTA utilizes CVAE [40] to generate an initial coarse human grasp and obtains the final grasp through test-time adaptation.

Modified GraspTTA [24]: We modify GraspTTA by augmenting input point clouds with predicted task-oriented contact maps.

F-GraspTTA [24]: We estimate the transformation between the initial and goal scene point clouds via iterative closest point. Next, we use GraspTTA to predict human grasps and compute initial and goal Object Penetration Points (OPPs) for each predicted grasp. If a predicted grasp intersects with either scene—indicated by an OPP exceeding a set threshold—we discard that grasp. The thresholds for Init and Goal OPP are set at 0.05. However, the method requires accurate pose transformation of the object between initial and goal states.

FLEX [45]: FLEX is initially designed to generate 3D full-body human grasps. We re-purpose the method to generate human grasps. Note that the penetration losses in both the initial and goal scenes drive the optimization.

ContactGen [32]: We train ContactGen on our dataset to predict object-centric contact representation, including the contact map, hand-part map, and direction map. Then, we apply their grasp synthesis solver to obtain predicted grasps.

SceneDiffuser [22]: We reimplement SceneDiffuser [22] to predict MANO [37] hand parameters. We do not perform optimization during inference, as we apply three auxiliary losses to generate physically plausible human grasps.

Contact Map Generation.

ContactGen [32]: We choose ContactGen as a baseline and compare it with the proposed ContactDiffser. Specifically, we adapt the sequential Conditional Variational Autoencoder (CVAE) from ContactGen to generate only a contact map, referring to this variant as **ContactCVAE**.

5.2. Evaluation Metrics

For **Placing** and **Shelving**, we test on 21 unseen objects. As for **Stacking**, we test on 6 unseen bricks. For each object, we test on 10 different task configurations. For every object in each task configuration, we predict 16 grasps for the evaluation. The quality of predicted grasps is evaluated based on their physical plausibility, stability, and collision avoidance, following prior works [24–26, 32, 42, 45]. The following introduces the metrics.

Penetration Volume (PV): We calculate the penetration volume by converting the meshes into 1mm cubes and measuring the overlap between these voxels.

Simulation Displacement (SD): We simulate the object and predicted grasps in PyBullet [16] 1 second and then calculate the displacement of the object's center of mass.

Contact Ratio (CR): Contact percentage of predicted grasps with objects.

Qualified Ratio (QR): The metric jointly considers both penetration volume and simulation displacement. Notably, a higher penetration volume generally results in lower simulation displacement, which is undesirable. 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.

Diversity Score (DS): We follow FLEX [45] to compute the average L_2 pairwise distance to assess the diversity.

Obstacle Penetration Percentage (OPP): We compute the penetration percentage of human grasp vertices in obstacles for initial and goal scenes [45].

Task Score (TS): We propose a metric called **TS** to evaluate the quality of task-oriented human grasp synthesis. An effective metric should consider physical plausibility, stability, and collision avoidance in both the initial and goal scenes. Thus, we define **TS** as $\mathrm{TS} = \mathrm{QR} \times (1 - \mathrm{Init\ OPP}) \times (1 - \mathrm{Goal\ OPP})$.

5.3. Results and Discussions

Can the proposed method synthesize high-quality task-oriented human grasps? We report our empirical studies in Table 1. While GraspTTA, ContactGen, and SceneDiffuser have low penetration volumes (PV↓), they struggle with stable grasps (SD↓) in **Placing** and **Shelving** tasks. In **Placing** and **Shelving**, F-GraspTTA [24] demonstrates strong performance in the Init **OPP**. However, due to inaccuracies in the transformation obtained from ICP, F-GraspTTA still struggles with the Goal **OPP**. In **Stacking**, fails to generate suitable human grasps for evaluation, as none of the predicted grasps meet the filtering criteria. FLEX seeks a balance between penetration volume and simulation displacement. ContactGen [32] exhibits a lower Init **OPP** compared to Goal **OPP**. This is attributed to its optimization strategy, which initially orients the hand toward the target object, resulting in a favorable initial **OPP**. However, due to the lack of scene and task modeling, ContactGen produces a higher goal **OPP**.

The **Stacking** task is more challenging due to the smaller size of the objects compared to those in **Placing** and **Shelving**. GraspTTA [24] cannot synthesize proper grasps (**QR**↑) and experience severe mode collapse (**DS**↑). Similarly, FLEX [45] struggles to achieve stable grasps in the **Stacking** task due to small bricks. The proposed method demonstrates favorable grasp synthesis (**QR**↑) in the challenging task. In the Supplementary Material, we provide a thorough analysis of different **QR** thresholds and their impacts on various baselines.

Our method demonstrates a favorable balance between penetration volume and simulation displacement, as reflected in **QR**. Most importantly, it achieves the best performance in **TS** among all tasks, demonstrating the proposed framework can synthesize high-quality task-oriented human grasps in terms of physical plausibility, stability, and collision avoidance in initial and goal scenes.

Is the task-aware contact map crucial for task-oriented human grasp synthesis? In Table 2, we compare the influence of different contact maps, i.e., object-centric (OC) and task-aware (TA) contact maps, to the quality of human grasp synthesis. Incorporating TA into both Modified-GraspTTA [24] and our framework results in significant performance improvements in the **TS** metric compared to using OC. Additionally, we report results using ground-truth contact maps, which serve as an upper bound for the two-stage framework. In particular, GT produces low OPP in the initial and goal scenes. The observed performance gap highlights a promising research direction for the community to address collectively. A noteworthy observation is that M-GraspTTA outperforms our method in **Stacking** when TA is incorporated. This is attributed to the smaller object size

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. Note FLEX [45] struggles to synthesize stable grasps in **Stacking** due to its inability to handle small bricks.

	Method	PV(Avg/Std)↓	SD(Avg/Std)↓	CR(%)↑	QR(%)↑	DS ↑	Init OPP(%)↓	Goal OPP(%)↓	TS↑
Placing	GraspTTA [24]	1.83/2.51	2.55/2.90	98.92	59.10	75.01	21.36	17.67	0.382
	F-GraspTTA [24]	1.85/2.29	2.47/2.85	99.78	58.46	75.55	0.54	15.41	0.491
	ContactGen [32]	1.44/1.99	3.85/6.75	93.57	47.29	**87.01**	**5.92**	17.67	0.366
	SceneDiffuser[22]	**1.39**/2.04	3.10/3.49	96.51	54.58	67.79	20.88	16.17	0.362
	FLEX [45]	2.61/2.86	1.62/2.31	**99.88**	58.09	32.92	6.84	**5.55**	0.511
	Ours	2.36/2.55	**1.44**/1.97	99.34	**64.61**	41.91	6.82	6.35	**0.563**
Stacking	GraspTTA [24]	4.31/2.77	**0.28**/0.31	**100.00**	34.87	0.30	26.05	8.32	0.236
	F-GraspTTA [24]	0/0	0/0	0	0	0	0	0	0
	ContactGen [32]	0.66/0.66	1.85/2.94	94.89	78.35	62.53	7.80	9.44	0.638
	SceneDiffuser[22]	0.53/0.81	1.62/2.53	94.85	78.35	63.13	25.45	9.66	0.527
	FLEX [45]	**0**/0	10.65 / 0	0/0	0.00	**107.87**	**0.5**	**0**	0
	Ours	1.13/1.22	0.97/1.59	95.79	**84.54**	48.45	14.76	4.55	**0.687**
Shelving	GraspTTA [24]	1.84/2.34	2.54/2.91	99.04	58.86	74.63	15.70	13.61	0.428
	F-GraspTTA [24]	1.88/2.08	2.68/3.12	99.50	58.08	74.84	1.00	11.09	0.511
	ContactGen [32]	**1.43**/2.12	3.90/4.03	93.13	46.32	**89.79**	6.42	13.17	0.376
	SceneDiffuser[22]	1.37/2.20	3.35/3.68	95.91	50.86	68.18	14.60	13.77	0.376
	FLEX [45]	2.80/3.00	**1.57**/1.96	**99.90**	57.45	29.02	**4.41**	**4.46**	0.524
	Ours	2.15/2.84	1.63/2.17	99.48	**66.94**	52.74	8.55	10.06	**0.550**

Table 2. A comparison between Object-Centric (OC), Task-Aware (TA), and Ground-Truth (GT) contact maps.

	Method	Type	QR(%)↑	Init OPP(%)↓	Goal OPP(%)↓	TS ↑
Placing	M-GraspTTA [24]	OC	51.39	19.38	16.54	0.345
		TA	42.08	9.61	5.64	0.358
		GT	40.00	3.91	3.99	0.369
	Ours	OC	64.46	17.95	16.92	0.439
		TA	65.29	7.27	5.79	**0.570**
		GT	68.30	1.69	1.66	0.660
Stacking	M-GraspTTA [24]	OC	90.16	26.76	9.48	0.507
		TA	87.52	16.05	5.54	**0.694**
		GT	93.92	6.67	2.70	0.852
	Ours	OC	82.77	25.05	9.51	0.561
		TA	84.31	14.94	4.51	0.684
		GT	83.33	3.12	1.11	0.798
Shelving	M-GraspTTA [24]	OC	50.66	13.80	12.44	0.382
		TA	52.51	11.05	11.65	**0.412**
		GT	41.91	4.62	4.49	0.381
	Ours	OC	65.10	13.66	13.68	0.485
		TA	67.49	8.72	10.31	**0.552**
		GT	67.36	4.71	4.96	0.610

in **Stacking**. Additional analysis, including qualitative insights, is provided in the following sections.

Do the proposed ContactDiffuser and GraspDiffuser play crucial roles in synthesizing high-quality task-oriented human grasps? Ablation studies of our method are reported in Table 3 and Table 4. We use ContactDiffuser as the task-aware contact map predictor and study the difference between GraspDiffuser and M-GraspTTA. In Table 3, we show that GraspDiffuser significantly improves the quality of task-oriented human grasp in **Placing** and **Shelving**. In Table 4, we use GraspDiffuser as the grasp predictor and study the influence of different contact map predictors. ContactDiffuser generates favorable task-oriented contact maps, resulting in superior task-oriented human grasps in terms of **TS** for the **Placing** and **Shelving** tasks. Note that the proposed method performs on par with other baselines in **Stacking**, highlighted in magenta. This is due to the proposed method cannot produce distinguishable task-oriented contact maps for grasp synthesis, as shown in Fig. 4.

Table 3. Ablation study of GraspDiffuser. We denote GraspDiffuser as **GD**, ContactDiffuser as **CD**, and Modified-GraspTTA as **M-GraspTTA**.

	Method	QR(%)↑	Init OPP(%)↓	Goal OPP(%)↓	TS ↑
Placing	CD + M-GraspTTA [24]	49.13	9.70	6.23	0.416
	CD + GD	64.61	6.82	6.35	**0.563**
Stacking	CD + M-GraspTTA [24]	86.68	15.74	5.56	**0.689**
	CD + GD	84.54	14.76	4.55	0.687
Shelving	CD + M-GraspTTA [24]	53.11	9.56	9.45	0.434
	CD + GD	66.94	8.55	10.06	**0.550**

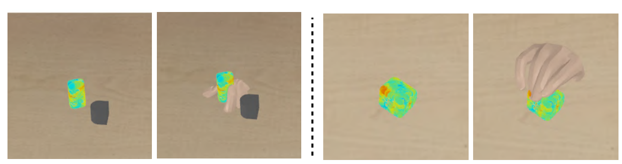

Figure 4. Failure examples. The proposed method struggles with predicting reliable task-aware contact maps for small objects.

Table 4. Ablation study of ContactDiffuser. We denote GraspDiffuser as **GD**, ContactDiffuser as **CD**, and ContactCVAE as **CC**.

	Method	QR(%)↑	Init OPP(%)↓	Goal OPP(%)↓	TS ↑
Placing	CC [32] + GD	62.90	7.91	**5.14**	0.549
	CD + GD	**64.61**	6.82	6.35	**0.563**
Stacking	CC [32] + GD	88.68	**15.62**	2.89	**0.690**
	CD + GD	84.54	14.76	4.55	0.687
Shelving	CC [32] + GD	59.06	11.49	**10.19**	0.469
	CD + GD	66.94	8.55	10.06	**0.550**

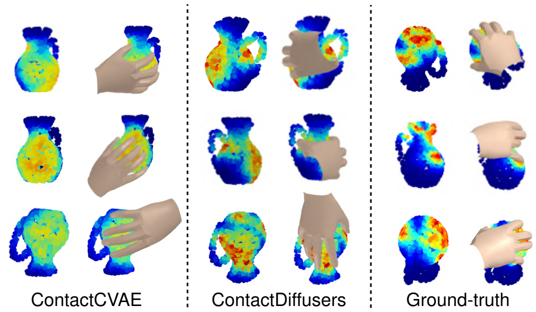

Figure 5. Visualization of predicted contact maps via ContactCVAE [32], ContactDiffuser, the ground-truth contact maps, and the corresponding synthesized human grasps.

5.4. Qualitative Results

Contact Map. Fig. 5 shows the results of human grasp synthesis using ContactCVAE [32], ContactDiffuser, and ground truth. ContactCVAE [32] tends to predict over-smooth contact maps because it fails to capture multimodal distributions of human grasps under diverse combinations of environmental contexts and tasks. In contrast, ContactDiffuser shows a strong capability to generate contact maps that closely resemble ground-truth contact maps.

Synthesized Human Grasps. ContactGen [32], GraspTTA[24], and SceneDiffuser's predictions collide with the scene severely, as shown in Fig. 6. FLEX [45] synthesize grasps with unrealistic contact and fail to generate a grasp for **Stacking**. Our method produces high-quality human grasps and avoids collision with obstacles.

Human Study. We conduct a human study to evaluate the perceptual quality of different human grasp synthesis methods. The study involves 10 subjects without prior research experience in human grasp synthesis. For **Placing** and **Shelving**, we select 3 unseen objects, and for **Stacking**, we choose 2 unseen bricks. For each object, we randomly sampled 8 grasps from the GraspTTA [24], FLEX [45], and Ours. Subjects are asked to rate each human grasp on a scale from 1 to 6, where 1 indicates grasps that are unnatural, unstable, and collide with the scene, and 6 indicates grasps that are natural, stable, and free of collisions with the scene. Table 5 presents each method's mean and standard deviation. Our method demonstrates promising results on **Placing** and **Stacking**, compared to GraspTTA and FLEX. FLEX performs favorably in **Shelving**. This is because human subjects can easily identify penetrations with the scene, resulting in better subjective scores. This observation aligns with the lower OPPs for FLEX in Table 1.

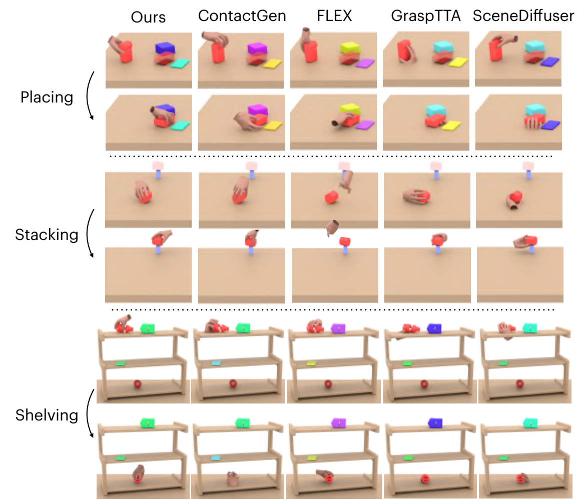

Figure 6. Visualization of predicted human grasps by Ours, ContactGen [32], FLEX [45], GraspTTA [24], and SceneDiffuser [22].

Table 5. Human Study of Synthesized Human Grasps.

Method	Placing Mean↑	Placing Std	Stacking Mean↑	Stacking Std	Shelving Mean↑	Shelving Std
GraspTTA [24]	2.06	0.69	2.65	0.84	2.54	0.56
FLEX [45]	2.65	1.03	1	0	**3.73**	0.82
Ours	**3.41**	0.94	**4.43**	0.93	3.38	0.55

6. Conculsion

In this work, we present a task-oriented human grasp synthesis task and a new dataset for development and benchmarking. We show that existing algorithms fail to generate high-quality grasps due to limited context and task modeling. We propose a novel two-stage diffusion-based framework that explicitly integrates essential context and task information to address this. Our thorough quantitative and qualitative experiments validate our proposed framework's effectiveness compared to strong baselines.

Limitations and Future Work. Our method cannot reliably predict contact maps for small objects, which may lead to collisions. We assume the initial and goal grasps are coherent, though we plan to further evaluate this assumption. At present, our experiments are conducted on a synthetic dataset that provides the goal scene point cloud as input. However, obtaining the target scene point cloud in real-world scenarios may be non-trivial. We plan to extend our work to the real world and aim to integrate task-oriented human grasp synthesis with hand-object motion synthesis.

Acknowledgements. The work is sponsored in part by the National Science and Technology Council under grants 113-2634-F-002-007-, 113-2628-E-A49-022- and 114-2628-E-A49-007-, the Higher Education Sprout Project of National Yang Ming Chiao Tung University, and the Ministry of Education, the Yushan Fellow Program Administrative Support Grant.

References

[1] Ananye Agarwal, Shagun Uppal, Kenneth Shaw, and Deepak Pathak. Dexterous functional grasping. In *7th Annual Conference on Robot Learning*, 2023. 1

[2] Chen Bao, Helin Xu, Yuzhe Qin, and Xiaolong Wang. Dexart: Benchmarking generalizable dexterous manipulation with articulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21190–21200, 2023. 1

[3] 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. 1, 2

[4] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 361–378. Springer, 2020. 1, 2

[5] J. Braun, S. Christen, M. Kocabas, E. Aksan, and O. Hilliges. Physically Plausible Full-body Hand-Object Interaction Synthesis. In *International Conference on 3D Vision*, 2024. 1

[6] Tim Brooks, Bill Peebles, Connor Homes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Wing Yin Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 3

[7] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021. 1

[8] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

[9] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 1, 2

[10] Yu-Wei Chao, Chris Paxton, Yu Xiang, Wei Yang, Balakumar Sundaralingam, Tao Chen, Adithyavairavan Murali, Maya Cakmak, and Dieter Fox. Handoversim: A simulation framework and benchmark for human-to-robot object handovers. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6941–6947. IEEE, 2022. 1

[11] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *European Conference on Computer Vision*, pages 231–248. Springer, 2022. 1

[12] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023. 3

[13] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20577–20586, 2022. 1, 2

[14] Sammy Christen, Lan Feng, Wei Yang, Yu-Wei Chao, Otmar Hilliges, and Jie Song. Synh2r: Synthesizing hand-object motions for learning human-to-robot handovers. *arXiv preprint arXiv:2311.05599*, 2023. 1

[15] 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. 1, 2

[16] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. `http://pybullet.org`, 2016–2021. 3, 6

[17] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[18] 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, 5

[19] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 1, 2

[20] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 1, 2

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4, 5

[22] 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, 5, 6, 7, 8

[23] 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. 1, 2

[24] 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, 5, 6, 7, 8

[25] 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. 1

[26] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *2021 International Conference on 3D Vision (3DV)*, pages 11–21. IEEE, 2021. 3, 5, 6

[27] Ninad Khargonkar, Neil Song, Zesheng Xu, Balakrishnan Prabhakaran, and Yu Xiang. Neuralgrasps: Learning implicit representations for grasps of multiple robotic hands. In *Conference on Robot Learning*, pages 516–526. PMLR, 2023. 1

[28] 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, 2, 5

[29] Qingtao Liu, Yu Cui, Zhengnan Sun, Haoming Li, Gaofeng Li, Lin Shao, Jiming Chen, and Qi Ye. Dexrepnet: Learning dexterous robotic grasping network with geometric and spatial hand-object representations. *arXiv preprint arXiv:2303.09806*, 2023. 1

[30] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14687–14697, 2021. 1

[31] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3292, 2022. 1

[32] 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, 5, 6, 7, 8

[33] Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*, pages 651–661. PMLR, 2022. 1, 5

[34] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 3

[35] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 5

[36] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, pages 570–587. Springer, 2022. 1

[37] 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. 3, 6

[38] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3

[39] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 1

[40] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 5

[41] Ryo Suzuki, Adnan Karim, Tian Xia, Hooman Hedayati, and Nicolai Marquardt. Augmented reality and robotics: A survey and taxonomy for ar-enhanced human-robot interaction and robotic interfaces. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2022. Association for Computing Machinery. 1

[42] 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. 1, 2, 3, 6

[43] O. Taheri, V. Choutas, M.J. Black, and D. Tzionas. GOAL: Generating 4D Wholebody Motion for Hand-Object Grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[44] Shuhan Tan, Tushar Nagarajan, and Kristen Grauman. Egodistill: Egocentric head motion distillation for efficient video understanding. *Advances in Neural Information Processing Systems*, 36, 2023. 1

[45] 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. 1, 5, 6, 7, 8

[46] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[48] Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. Unidexgrasp++: Im-

proving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. *arXiv preprint arXiv:2304.00464*, 2023. 1

[49] 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*, pages 11359–11366. IEEE, 2023. 1, 2, 3

[50] Tianhao Wu, Mingdong Wu, Jiyao Zhang, Yunchong Gan, and Hao Dong. Learning score-based grasping primitive for human-assisting dexterous grasping. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[51] Yueh-Hua Wu, Jiashun Wang, and Xiaolong Wang. Learning generalizable dexterous manipulation from human grasp affordance. In *Conference on Robot Learning*, pages 618–629. PMLR, 2023. 1

[52] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4737–4746, 2023. 1

[53] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20953–20962, 2022. 1, 2

[54] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19717–19728, 2023. 1

[55] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22479–22489, 2023. 1

[56] Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. Graspxl: Generating grasping motions for diverse objects at scale. In *European Conference on Computer Vision*, 2024. 1

[57] Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. ArtiGrasp: Physically Plausible Synthesis of Bi-Manual Dexterous Grasping and Articulation. In *International Conference on 3D Vision*, 2024. 1, 2

[58] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. CAMS: CAnonicalized Manipulation Spaces for Category-Level Functional Hand-Object Manipulation Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[59] 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. 1

[60] Yixin Zhu, Yibiao Zhao, and Song Chun Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2