UGG: Unified Generative Grasping

Jiaxin Lu[1*], Hao Kang[2], Haoxiang Li[3], Bo Liu[2], Yiding Yang[2], Qixing Huang[1], and Gang Hua[4]

[1] University of Texas at Austin, Austin TX 78712
lujiaxin@utexas.edu, huangqx@cs.utexas.edu
[2] ByteDance Inc., Bellevue WA 98004
{haokheseri,richardboliu,yangyd92}@gmail.com
[3] Pixocial Technology, Bellevue WA 98004
lhxustcer@gmail.com
[4] Dolby Laboratories, Bellevue, WA 98004
ganghua@gmail.com

Abstract. Dexterous grasping aims to produce diverse grasping postures with a high grasping success rate. Regression-based methods that directly predict grasping parameters given the object may achieve a high success rate but often lack diversity. Generation-based methods that generate grasping postures conditioned on the object can often produce diverse grasping, but they are insufficient for high grasping success due to lack of discriminative information. To mitigate, we introduce a unified diffusion-based dexterous grasp generation model, dubbed the name UGG, which operates within the object point cloud and hand parameter spaces. Our all-transformer architecture unifies the information from the object, the hand, and the contacts, introducing a novel representation of contact points for improved contact modeling. The flexibility and quality of our model enable the integration of a lightweight discriminator, benefiting from simulated discriminative data, which pushes for a high success rate while preserving high diversity. Beyond grasp generation, our model can also generate objects based on hand information, offering valuable insights into object design and studying how the generative model perceives objects. Our model achieves state-of-the-art dexterous grasping on the large-scale DexGraspNet dataset while facilitating human-centric object design, marking a significant advancement in dexterous grasping research. Our project page is https://jiaxin-lu.github.io/ugg/.

Keywords: Dexterous Grasping · Contact Representation · Generative Model

1 Introduction

The significance of robotic grasping is underscored by its ability to foster a human-like interaction between robotic systems and the environment, playing a fundamental role in supporting robots across a diverse range of tasks—from

[*]This work was initiated when Jiaxin Lu was an intern at Wormpex AI Research.

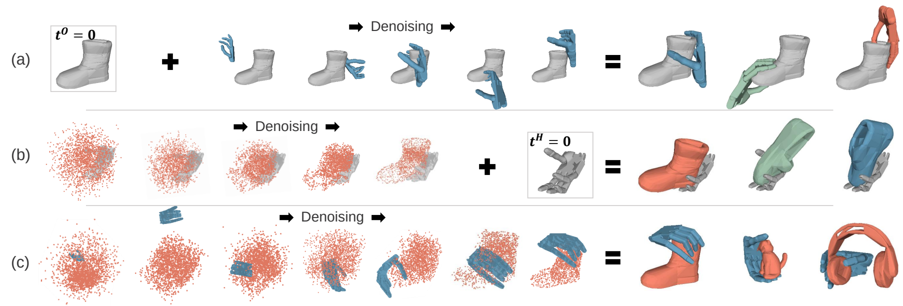

Fig. 1: Overview of three tasks performed by the proposed UGG model: (a) Generating grasps with a fixed object involves denoising hand parameters and transformations for diverse successful postures. (b) Generating objects with a fixed hand posture denoises shape latents for varied object fits. (c) Jointly generating hand posture and object involves simultaneous denoising of both latents, yielding diverse grasps with objects.

straightforward pick-and-place operations [4] to intricate assembly processes [12]. The increased interest in dexterous grasping comes from its ability to anthropomorphically grasp objects of diverse shapes and sizes [1, 64, 76, 81]. This capability not only improves the execution of downstream tasks, but also broadens the scope of robotic applications, including manufacturing, agriculture, healthcare, and extended reality [13, 18, 24, 31, 78].

In order to deal with the dynamically changing environment for planning [69, 76], dexterous grasping aims to produce diverse grasping postures with a high grasping success rate. Before the emergence of large datasets, the field of robotic grasping relied primarily on analytical-based methods, including force closure optimization [11], and simulation-based techniques, such as differentiable contact simulation [64]. With large-scale datasets such as DexGraspNet [70], data-driven approaches have become the focus, generally falling into two main categories, regression-based methods and generation-based methods.

Regression-based methods, as exemplified by DDG [33], directly predict the grasp parameters given the object as input, which is prone to either mode collapse, which hurts diversity, or mode averaging, which degrades regression accuracy and hence success rate. On the contrary, generation-based methods, such as GraspTTA [20], excel at producing a wider variety of grasping strategies. However, because they lack discriminative information on whether grasping could be a success, their success rate may not be sufficient.

To mitigate these insufficiencies, we propose a diffusion-based grasping generation model to guarantee grasp diversity, which is complemented by a simple physics discriminator to push for a high success rate. Our model operates in a latent space of 3D object point clouds and hand-joint parameters, explicitly focusing on diffusing hand positioning and *Contact Anchor*. The anchors are designed to transform the conventional intermediate format of the contact map [14, 20, 34] into a representation that not only aligns with the shape but can also be seamlessly integrated into the generation process. Furthermore, we employ an all-transformer architecture to unify the three elements - object, hand, and contact - into a single unified diffusion model.

Beyond grasp generation, our unified diffusion model exhibits expanded capabilities, generating hand and contact anchors from provided objects and reciprocally creating objects and their associated contact information based on given hand poses, as illustrated in Fig. 1. The generated shapes provide insightful perspectives on how the model perceives objects, thereby enhancing the interpretability of the generation outcomes and facilitating a more effective design of object representations. In addition, qualitative experiments indicate that this model can produce valid solutions for the human-centric design of objects.

In summary, our main contributions are as follows.

- We introduce a unified diffusion model UGG for investigating hand-object interaction tasks through a perspective of generation. Our model seamlessly brings grasping, object generation, and affordance analysis into a cohesive framework.
- We propose a new contact anchors representation instead of using the conventional contact map as navigational cues for affordance information. This representation aligns harmoniously with the point cloud representation, enabling effective participation in the multi-modal generative tasks.
- Leveraging the diversity and validation rate of our model, we introduce a physics discriminator to assess the success of hand-object grasping, achieving state-of-the-art performance in the grasp generation task.
- In particular, UGG exhibits the unique ability to generate objects based on given hand parameters. This not only advances research on object representations in hand-object interaction tasks but also paves the way for human-centric object design studies.

2 Related Work

Dexterous Grasping. Analytical and simulation-based multi-finger grasping methods [10, 11, 27, 35, 45, 49, 51, 54, 64] were extensively investigated prior to the advent of large-scale datasets [65, 70], which exhibit limited generalizability or require heavy computing. Data-driven approaches [47], whether directly regressing grasp based on the object [32, 33, 53] or using generative models for grasp synthesis [20, 38, 39, 72], frequently confront the dilemma of balancing diversity and quality. Meanwhile, contact information is frequently used indirectly as a bridge [6, 14, 22, 28, 29, 34, 43, 58, 67, 82] to improve grasp quality.
Diffusion Models. Diffusion models represent a significant advancement in image generation [17, 23, 52, 55, 56, 60–62, 79]. Expanding into the 3D domain, diffusion models have shown great success in generating point clouds [40, 77, 80], meshes [15, 36], signed distance functions [8, 9], and neural fields [46, 59], facilitating high-quality and controllable 3D synthesis [30, 41, 73]. We adopt the VAEs [26] of LION [77] to produce the object representation, which excels in the precise reconstruction of point clouds. The hierarchical latent representation of

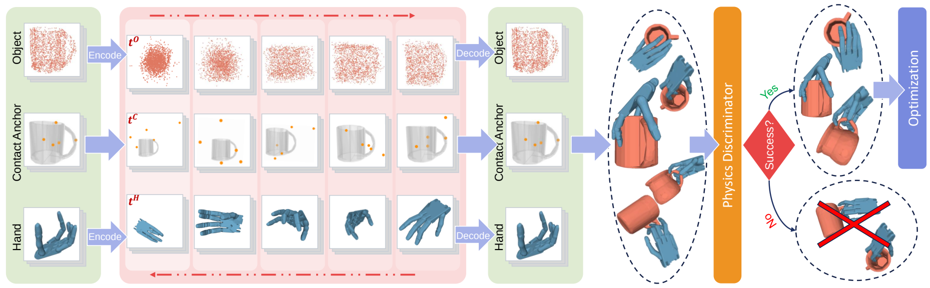

Fig. 2: Overview of the proposed method UGG: Our approach involves encoding and embedding the object, contact anchors, and hand to facilitate the learning of a unified diffusion model. During inference, random seeds are sampled and subjected to a denoising process to generate samples. To discern potentially successful grasps, a physics discriminator is introduced. Subsequently, an optimization stage is undertaken for all selected grasps, utilizing the generated contact anchor and input point cloud.

these VAEs facilitates the subsequent diffusion-based generation of both hand posture and object.

Diffusion models effectively handle spatial connection-based generative challenges, as evidenced by molecular design [74], layout generation [7, 71], gripper pose generation [66] and scene synthesis and planning [19]. Methods in [19, 66] exhibit similarities in diffusion grasping, but our proposed model can simultaneously generate both object and hand. While gripper [66] has fewer parameters, it is not geared for dexterous grasping. The SceneDiffuser [19], though is able to handle dexterous grasping, lacks modeling grasp-specific context, *e.g.*, contact, affecting posture quality. In addition, it conditions posture by object with cross-attention [68], which does not sufficiently consider the alignment of manifold priors [3, 75], is insufficient in dealing with colliding and floating poses.

3 Problem and Our Approach

3.1 Problem Statement

In a broad sense, dexterous grasping is a multistage process that encompasses various aspects such as perception, motion planning, and execution. In a narrow sense, it refers to the process of producing hand postures that can successfully grasp an object. Following the literature, we address it from a narrow sense. More specifically, we assume that the representation of the given object $\boldsymbol{o}$ is a point cloud, *e.g.*, $\boldsymbol{o} \in \mathbb{R}^{N\times 3}$. Let $\boldsymbol{h} = (\boldsymbol{\theta}, \boldsymbol{R}, \boldsymbol{t})$ represent a posture of the hand, where $\boldsymbol{R} \in \mathrm{SO}(3)$ and $\boldsymbol{t} \in \mathbb{R}^3$ denote root rotation and translation, $\boldsymbol{\theta} \in \mathbb{R}^k$ represent the joint angles of the hand model, with k denoting the degree of freedom (*e.g.*, $k = 22$ for the adopted ShadowHand [57] model).

The objective of dexterous grasping is to produce a diverse set $\mathbb{H} = \{\boldsymbol{h}_i\}_{i=1}^m$ of articulated hand postures that can grasp the target object $\boldsymbol{o}$ with a high success rate. An ideal method would produce a highly diverse set of graspings with a high success rate. The definition of what a successful grasp is and the quantification of the diversity of the produced grasping $\mathbb{H}$ could depend on the application. We follow the definition and quantification presented by Wang et al. [70]. In particular, a grasp is considered a success "if it can withstand at least one of the six gravity directions and has a maximum penetration of less than 5mm [70]". The diversity of the produced grasps is measured by the entropy of the joint angles of the generated grasp postures.

3.2 Overview of Our Approach

As shown in Fig. 2, our formulation approaching the dexterous grasping problem is composed of two main parts. The first part is a unified diffusion model which jointly models the generation of grasping hands, objects, and their contact information, which is capable of generating a diverse set of candidate grasping postures given an object of interest. The second part is a lightweight discriminative classifier trained from simulation data that entails discriminative information on whether a grasp is a success.

Unified Grasping Generator. In the natural shape space, dexterous grasping involves three elements, *e.g.*, the *articulated grasping hand*, the *object*, and the *contact* between the two. We regard the problem of dexterous grasping as the interaction among three data manifolds, *i.e.*, the manifold of the articulated grasping hands, and that of the objects, aligned by the manifold of the contact.

Therefore, the question arises as to how to effectively model the interactions among the three manifolds. There are several factors to consider to establish an effective model. The first is how to represent the three manifolds in latent space that effectively model their structures, *e.g.* latent embeddings. The next question is how to model the interactions of the three manifolds to effectively leverage the interaction irregularity in the training data.

As discussed in Sec. 3.1, objects are represented as point clouds, and hands are tuples that include joints and rigid 3D translations. Following LION [77], objects and hands are encoded as shape-latent codes by variational auto-encoders (VAE) [26], details included in Sec. 4.1.

For contact modeling, most previous methods have used the contact map, contact parts, and direction [14, 20, 22, 34]. These representations are typically continuous, posing a challenge in their applicability to the unordered set nature of point clouds during the generative process. The diffusion process does not preserve the mapping between the contact and point cloud. We found a very simple but very effective new representation, *Contact Anchor*, which characterizes the contact information as a set of N_c contact anchors, *e.g.*, $\boldsymbol{c} \in \mathbb{R}^{N_c \times 3}$. Please refer to Sec. 4.2 for details.

To effectively model the interaction among the three manifolds, we cast them into a unified diffusion model by fitting one transformer to model all distributions. Inspired by UniDiffuser [3], we use its transformer-based backbone, known as U-ViT [2], to serve as the core of our joint noise prediction network. Depicted in Fig. 2, the denoising process is trained on asynchronous scheduling of different manifolds, formulating all conditional, marginal, and joint distributions. The UniDiffuser samples combinations of hands and objects not restricted to training data, facilitating a generative model with high generalization capability.

Specifically, the generation behavior can be manipulated by carefully adjusting the sampling scheduler. Fig. 1 shows different generation modes: hand generation by fixing step 0 to the object scheduler (a), object generation by fixing step 0 to the hand scheduler (b), and combined generation without fixing any scheduler (c).

Physics Discriminator. The unified diffusion model is capable of producing a diverse set of grasping postures. However, we can not neglect the hallucination phenomenon, which implies that a significant portion of the generated candidate grasping may not produce a successful grasp.

Given sufficient discriminative guidance, a classifier could be learned to predict whether a candidate's posture would be able to successfully grasp the target object or not. However, discriminative guidance, *e.g.*, the label of whether a grasp would be successful, is not readily available. To obtain such discriminative labels for training, we resort to the physics simulator, Isaac Gym [42], to judge whether a rendered candidate would compose a successful grasping. This discriminator takes advantage of the diversity of the results generated and significantly improves the success rate of grasping, shown in Sec. 6. Furthermore, we demonstrate that with the bless of high initial diversity and quality, the discriminative information can be utilized to adjust success rate and diversity, a point overlooked in related work based on generative models [44].

4 Formulation

The proposed framework comprises three components. First, the embeddings of objects and hands (Sec. 4.1), bridged by the novel contact modeling (Sec. 4.2), formulate a unified representation in latent space. Then, a unified diffusion model (Sec. 4.3) is trained to generate effective and diversified graspings. Finally, a discriminator (Sec. 4.4) is meticulously formulated to ensure a high success rate in hand generation without compromising diversity.

4.1 Unified Shape Encoding

Objects are depicted as point clouds in natural-shape space and hybrid embeddings in latent space. As shown in [77], a hierarchical VAE is capable of capturing global shapes across categories with various scales, while maintaining details for precise reconstruction. Moreover, its latent representation is suitable for diffusion models, facilitating the generation of diverse shapes.

Specifically, given a point cloud $\boldsymbol{o} \in \mathbb{R}^{N\times 3}$, a global embedding $\boldsymbol{g}_o \in \mathbb{R}^{d_g}$ is obtained by a global encoder. A local encoder maps each point in the point cloud to a d_l dimensional local feature regulated by the global embedding, resulting in a local representation $\boldsymbol{l}_o \in \mathbb{R}^{N\times(3+d_l)}$. The hybrid embedding of an object

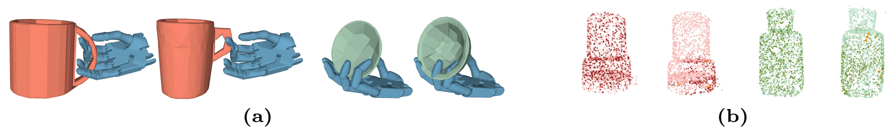

Fig. 3: Motivation of a unified contact modeling. (a) Subtle object changes can lead to grasping failure, as the corresponding adjustment in latent representation may not capture critical details adequately. (b) Contact map fails when embedded into a joint generation model. Left: generated noisy contact map (deeper for closer). Right: generated *Contact Anchor* (yellow) with a grasp and the GT contact map of the generated grasp. Zoom in for better view.

$\boldsymbol{x}_o = \boldsymbol{g}_o \oplus \boldsymbol{l}_o$ is a concatenation of global and local latent codes. Subsequently, a decoder is introduced to reconstruct the latent representation $\mathbf{x}_o$ into its point cloud. The architecture of the encoders and decoders adheres closely to the Point-Voxel CNNs (PVCNNs) [37] framework, which was also utilized in LION [77].
Hands are parameterized as a tuple $(\boldsymbol{\theta}, \boldsymbol{R}, \boldsymbol{t})$, where $\boldsymbol{\theta} \in \mathbb{R}^k$ are joint angles from ShadowHand [57], $\boldsymbol{R} \in \mathrm{SO}(3)$ and $\boldsymbol{t} \in \mathbb{R}^3$ denote 3D rotation and translation. In order to encourage a semantic embedding that facilitates easy reconstruction, a latent code representing the shape of the hand is encoded from $\boldsymbol{\theta}$ to ensure its independence from rigid transformations. Similar to the global embedding of an object, VAE is used with the decoder parameterized as a factorial Gaussian distribution corresponding to an L_2 reconstruction loss, resulting in a latent shape code $\boldsymbol{g}_h$. Adding 3D rotation and translation, the final hand representation is $\boldsymbol{x}_h = \boldsymbol{g}_h \oplus \boldsymbol{r} \oplus \boldsymbol{t}$, where $\boldsymbol{r}$ is the flattened rotation matrix $\boldsymbol{R}$, and $\oplus$ is the concatenation of features.

4.2 Unified Contact Modeling

Object and hand encoders are individually trained to facilitate their own shape reconstruction, without aligning together for good grasping. As shown in Fig. 3a, a subtle change in the object can cause grasping failure. The corresponding change in latent representation may not be significant enough to capture critical detail. This underscores the need to incorporate contact information into the unified representation.

Previous approaches have taken two main routes to capture this information. Pre-processing methods [14, 22, 34] generate contact information directly before employing optimization algorithms to fit a hand model. The post-processing approach, *e.g.*, GraspTTA [20], involves a two-stage training process to first compute a possible contact map based on an initial hand estimate, and then refine the pose on it. Unfortunately, neither of the existing contact modelings successfully accomplishes the joint generation of both hand parameters and a contact model. Meanwhile, our attempts in embedding the contact map modeling into the joint generative model (Fig. 3b) yielded pure noise. We attribute it to the unordered nature of the point cloud, which failed to provide a clear pattern for generative contact modeling.

We propose *Contact Anchor* as a novel modality to convey contact information. It is a set of points located on the meshes of both the object and the hand. In practice, given the distance of a point $\boldsymbol{c}$ to a mesh $\mathbf{S}$ as:

$$d(\boldsymbol{c}, \mathbf{S}) = \min_i ||\boldsymbol{c} - \boldsymbol{s}_i||_2, \quad \forall \boldsymbol{s}_i \in \mathbf{S} \; , \tag{1}$$

Contact Anchor, $\boldsymbol{c} \in \mathbb{R}^{N_c \times 3}$, is composed by randomly selecting N_c points from the object point cloud $\boldsymbol{p}_o$, such that the distance to the hand mesh $\mathbf{S}_h$ is below a threshold η:

$$\boldsymbol{c} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{N_c}], \quad \boldsymbol{c}_i \in_R \{d(\boldsymbol{c}, \mathbf{S}_h) < \eta | \boldsymbol{c} \in \boldsymbol{p}_o\} \; , \tag{2}$$

where $\in_R$ denotes uniformly random sampling of elements from a set.

This design avoids the extensive calculation of mesh intersection, while explicitly revealing grasping patterns. Represented as a set of points, *Contact Anchor* facilitates seamless integration and easy reconstruction through the proposed unified diffusion framework.

4.3 Unified Generation

UniDiffuser [3] explored a unified perspective on the diffusion model by fitting one transformer to model all distributions. Given our objective of aligning three distributions, specifically object, hand, and contact information, this approach seamlessly aligns with the intuition behind UniDiffuser. Therefore, we leverage its transformer-based backbone, known as U-ViT [2], to serve as the core for our joint noise prediction network.

In our effort to align all distributions using U-ViT, a crucial step in employing transformer-based methods is to tokenize each representation, with each token embedded in $\mathbb{R}^d$. The global embedding $\boldsymbol{g}_o$ of the object point cloud is mapped naturally into a single token. Due to its high dimension, the local embedding, however, is downsampled to N_l tokens by farthest point sampling and k-nearest-neighbor grouping [50]. For hand $(\boldsymbol{g}_h, \boldsymbol{r}, \boldsymbol{t})$, each component is embedded independently into a token, and for contact anchors $\boldsymbol{c}$, each point is represented by one token. In general, an object representation $\boldsymbol{o}$ with $N_l + 1$ tokens, a hand code $\boldsymbol{h}$ with 3 tokens, and a contact map $\boldsymbol{c}$ with N_c tokens are combined as reconstruction data from the unified diffusion pipeline.

We hereby introduce a unified formulation that includes the conditional distributions $q(\boldsymbol{o}_0, \boldsymbol{h}_0|\boldsymbol{c}_0)$, $q(\boldsymbol{o}_0, \boldsymbol{c}_0|\boldsymbol{h}_0)$, $q(\boldsymbol{c}_0, \boldsymbol{o}_0|\boldsymbol{h}_0)$, $q(\boldsymbol{o}_0|\boldsymbol{h}_0, \boldsymbol{c}_0)$, $q(\boldsymbol{h}_0|\boldsymbol{o}_0, \boldsymbol{c}_0)$, $q(\boldsymbol{c}_0|\boldsymbol{o}_0, \boldsymbol{h}_0)$, and the joint distribution $q(\boldsymbol{o}_0, \boldsymbol{h}_0, \boldsymbol{c}_0)$. To achieve this, we introduce the concept of *asynchronized timesteps* for the data, denoted as t^o, t^h and t^c. Expectations modeling takes on a general form expressed as $\mathbb{E}[\boldsymbol{\epsilon}^o, \boldsymbol{\epsilon}^h, \boldsymbol{\epsilon}^c | \boldsymbol{o}_{t^o}, \boldsymbol{h}_{t^h}, \boldsymbol{c}_{t^c}]$. Specifically, when $t^o = 0$, $\mathbb{E}[\boldsymbol{\epsilon}^h, \boldsymbol{\epsilon}^c | \boldsymbol{o}_0, \boldsymbol{h}_{t^h}, \boldsymbol{c}_{t^c}]$ corresponds to the conditional distribution $q(\boldsymbol{h}_0, \boldsymbol{c}_0|\boldsymbol{o}_0)$, *e.g.*, grasping generation, visualized in Fig. 1(a).

By denoting $\boldsymbol{ohc} = (\boldsymbol{o}, \boldsymbol{h}, \boldsymbol{c})$, $\boldsymbol{t} = (t^o, t^h, t^c)$, $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}^o, \boldsymbol{\epsilon}^h, \boldsymbol{\epsilon}^c]$ and $\boldsymbol{ohc_t} = (\boldsymbol{o}_{t^o}, \boldsymbol{h}_{t^h}, \boldsymbol{c}_{t^c})$, within the unified perspective of diffusion models, our objective is formulated as

$$\mathbb{E}_{\boldsymbol{ohc}_0, \boldsymbol{t}, \boldsymbol{\epsilon}} \|\boldsymbol{\epsilon}_\xi(\boldsymbol{ohc_t}, \boldsymbol{t}) - \boldsymbol{\epsilon}\|_2^2 \; , \tag{3}$$

where ξ is the parameters of the noise prediction deep neural network. It is essential to note that this formulation provides a cohesive and integrated framework, allowing us to seamlessly model hand-object interactions from a generative perspective using a single model.

4.4 Physics Discriminator

The proposed unified model generates grasping configurations conditioned on given objects. While providing high diversity in potential postures, generative models contend with a diminished success rate, due to the absence of deterministic objectives during training.

Specifically, the evaluation process involves a physical simulator, Isaac Gym [42], to determine the grasping quality. The simulator is not differentiable and, therefore, cannot be implemented as an objective during training. To bridge this gap, we introduce a lightweight recognition model called the *Physics Discriminator*. This model efficiently identifies successful samples among the generated postures and operates independently from the generative and adaptation modules. We also emphasize its capability of balancing success rate and diversity, an aspect subtlety overlooked by prior methods. This potential becomes evident now when the base model achieves both a commendable success rate and high diversity. Our ablation study in Sec. 6.4 shows that the *Physics Discriminator* enables control over the generation success rate and diversity.

A challenge arises from existing datasets containing only valid grasps, lacking invalid samples for training the discriminator. Fortunately, our generative model can easily create adversarial samples. Ground truths for training are obtained via the physical simulator, labeling successful grasps as $\boldsymbol{y} = [0, 1]^T$ and unsuccessful ones as $\boldsymbol{y} = [1, 0]^T$. Given an object point cloud $\boldsymbol{o}$ and a generated hand $\hat{\boldsymbol{h}}$, the physical discriminator D is trained with a cross-entropy loss:

$$\mathcal{L}_{\text{dis}}(\boldsymbol{o}, \hat{\boldsymbol{h}}, \boldsymbol{y}) = -\sum_{i=0}^{1} y_i \log \frac{\exp[D(\boldsymbol{o}, \hat{\boldsymbol{h}})_i]}{\sum_{j=0}^{1} \exp[D(\boldsymbol{o}, \hat{\boldsymbol{h}})_j]} \ . \tag{4}$$

5 Learning and Inference

Learning. The training of the entire framework can be divided into three stages. In the first stage, VAEs are trained individually to obtain latent representations of objects and hands. Then, a unified diffusion is trained to minimize Eq. 3. Finally, the physical discriminator is optimized on the grasping results generated within the training set.

Inference. In the inference stage, given the target object (a.k.a., setting its diffusion branch to constant time step t_o^o), we run the proposed unified diffusion model to obtain a diverse set of M candidate grasping hands. Then we feed the M candidate grasping hands to the discriminator, and the top N candidates with the highest classification scores are picked up as output. Furthermore, we adopt the ***test-time adaptation*** from [20] to ensure physically plausible results. More specifically, given *Contact Anchor* $\boldsymbol{c} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{N_c}]$, the contact loss is defined as $\mathcal{L}_{\text{cont}} = \sum_{i=1}^{N_c} d(\boldsymbol{c}_i, \mathbf{S}_h)$. The overall loss in test-time optimization is

$$\mathcal{L}_{\text{test}} = \omega_{\text{pen}}\mathcal{L}_{\text{pen}} + \omega_{\text{spen}}\mathcal{L}_{\text{spen}} + \omega_{\text{joint}}\mathcal{L}_{\text{joint}} + \omega_{\text{cont}}\mathcal{L}_{\text{cont}} \;, \tag{5}$$

where $\mathcal{L}_{\text{pen}}$, $\mathcal{L}_{\text{spen}}$, and $\mathcal{L}_{\text{joint}}$ are penetration, self-penetration and joint angle losses from [70] and [20]. We optimize the loss function through gradient descent using ADAM [25] for 100 steps.

6 Experiment and Evaluation

6.1 Experiment Setup

Dataset. We benchmark our method using the DexGraspNet [70], a challenging large-scale dexterous grasping benchmark. It encompasses 1.32 million dexterous grasps on 5355 objects, following the structure of ShadowHand [57]. Validation is done in the Isaac Gym simulator [42] and on penetration depth, ensuring grasp effectiveness. The training set includes 4229 objects, and the test set has 1126 objects, with 236 from novel categories. Given the complex nature of this dataset, we employ two subsets, namely "20 objects" and "10 bottles", to generate results and conduct the ablation study, due to the large training time on the entire dataset. We also tested our method on the human-object interaction datasets HO3D [16] and GRAB [5, 63]. Further details on these datasets and results are provided in the appendix.
Metrics. In our evaluation, we adhere to the metrics established in each benchmark, ensuring a fair comparison with baseline methods. These metrics encompass two key aspects: quality and diversity. For all metrics, we follow [20,21,34,70] for implementation. Details of each metric are provided in the supp. material.
Baseline Methods. We compare three methods on the DexGraspNet benchmark dataset: DDG [33], GraspTTA [20], and the generation module in UniDexGrasp [76] (abbr. UDG).

6.2 Results of Dexterous Grasping

Quantitative Results. The quantitative results, as depicted in Tab. 1, underscore the superior performance of UGG compared to baseline methods.

UGG achieves the highest success rate (72.7%), Q_1 (0.063), and the lowest penetration (0.14) among all methods. In terms of diversity, UGG outperforms DDG, GraspTTA, and UniDexGrasp significantly, demonstrating superiority in both mean entropy (7.17) and standard deviation (0.07). This highlights the efficacy of the proposed unified diffusion model.

Table 1: Quantitative results of grasp generation of UGG and benchmark methods on the DexGraspNet.

	Quality			Diversity	
	success ↑	Q_1 ↑	pen ↓	H mean ↑	H std ↓
DDG [33]	67.5	0.058	0.17	5.68	1.99
GraspTTA [20]	24.5	0.027	0.68	6.11	0.56
UDG [76]	23.3	0.056	0.15	6.89	0.08
UGG (w/o disc)	64.1	0.036	0.17	**8.31**	0.28
UGG (ours)	**72.7**	**0.063**	**0.14**	7.17	**0.07**

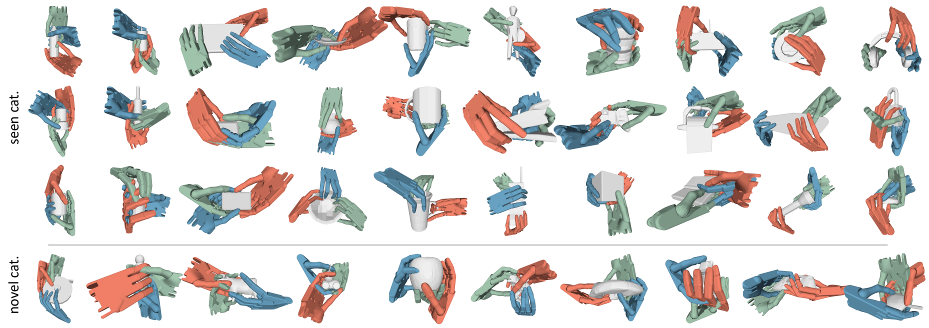

Fig. 4: Visualization of the generated diverse grasps of UGG on the DexGraspNet objects (mesh used only for visualization). Top: grasps of objects from seen categories; Bottom: grasps for objects of novel categories.

In particular, even without the proposed physics discriminator, our method achieves a success rate of 64.1%, only marginally trailing the DDG method. Our approach, drawing insights from both generative and regression-based models, establishes itself as a state-of-the-art performer in terms of both quality and diversity. For seen categories, our method achieves an average success rate of 73.8%, while for novel categories, our model achieves a success rate of 68.1%. These results underscore the strong generalizability of UGG to novel categories.

Qualitative Results. Fig. 4 illustrates the grasps generated by UGG. The top two rows display results for objects within seen categories (though not part of the training set), while the bottom row showcases results for objects falling under novel categories not encountered during training. Each object is accompanied by three distinct generated grasps, highlighting the considerable diversity of the generated grasps. The depiction of mugs in the middle emphasizes the model's capacity to encompass even the handle in its generated grasps. Another noteworthy instance is the headphone in the top right corner, where the generative model exhibits proficiency in devising grasping approaches involving both earpads and the headband.

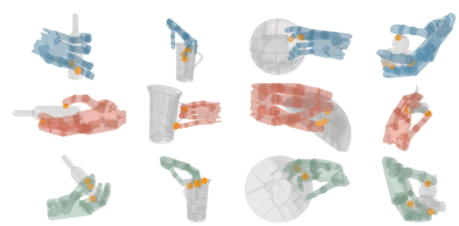

Fig. 5: Visualization of *Contact Anchor*. The yellow points are the generated contact anchors. Zoom in for a better view.

Additionally, Fig. 5 shows the generated contact anchors, which highlight key positions for grasping. Our hand model conforms well to these diverse anchors, avoiding fixed poses. For unseen objects like the bunny car on the right, the anchors effectively identify grasp points. When the contact area between hand and object is small, anchors may appear close due to a strict threshold in training data, which aligns with the precision of the grasping task and does not negatively affect the optimization algorithm.

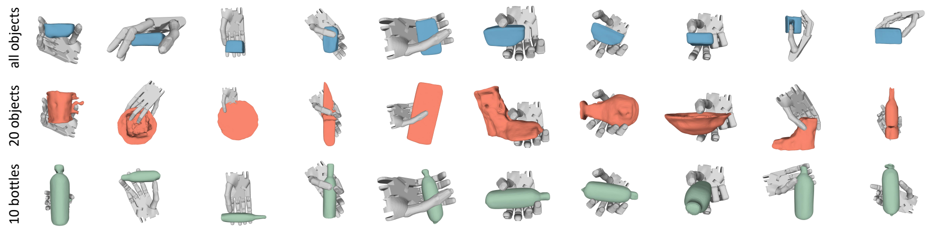

Fig. 6: Visualization of object generation by UGG across three subsets, utilizing grey hand poses subjected to different transformations. The hand poses within each column are the same. The top row presents the results of the model trained on the entire dataset, while the middle and bottom rows exhibit the results of the models trained on subsets comprising 20 objects and 10 bottles, respectively.

6.3 Other Generative Results

In addition to the well-defined task of dexterous grasping by hand generation, the proposed framework can also generate objects based on a given hand, or a pair of an object and a hand from scratch. Both exhibit a successful grasping. In this part, we show the qualitative results of these generations. The mesh is constructed based on the generated point cloud using the off-the-shelf method *Shape As Points* (SAP) [48].

Object Generation. In Fig. 6, we present the results of object generation based on the hand poses provided using three models trained on *the entire dataset*, the *20 objects subset*, and the *10 bottles subset*, respectively. The tests maintain consistent hand joint parameters for each model (in each column), with variations in rotation and translation.

Overall, the models exhibit the capability to generate valid objects corresponding to the given hand poses. We observe that objects generated from the entire dataset predominantly take on the form of small cubes that fit within the hand without extending significantly beyond. Several explanations emerge: 1) the entire dataset comprises a substantial number of objects with an abstract cube-like shape. 2) Learning to grasp on the entire dataset enables the model to grasp the abstract concept of a graspable shape for object generation. 3) The "graspability" of an object is primarily determined by the portion within the hand. These generation results provide insight into how the model perceives the object during hand pose generation, offering a valuable perspective for designing more efficient 3D representations in grasping tasks.

When presented with a smaller subset of objects, our model demonstrates the ability to generate a diverse range of objects across different categories, as evidenced in the middle and bottom rows of Fig. 6. This aligns with the common practice in point cloud generation methods to operate on a per-category basis. Objects such as bowl, plate, and mug in the second row demonstrate a design adapted to the hand pose. The leftmost bottle in the third row, with a slight

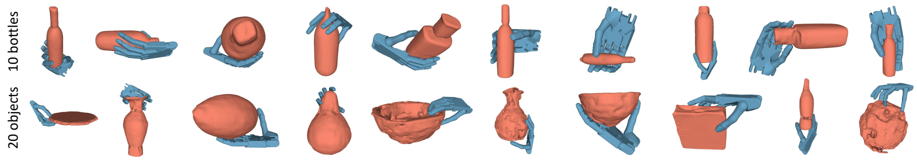

Fig. 7: Joint generation visualization of UGG, illustrating simultaneous generation of hand and objects. Results are presented for models trained on two subsets.

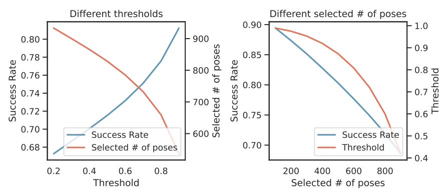

Fig. 8: Success rate after the discriminator. Left: over the changes of threshold; Right: over the number of selected poses.

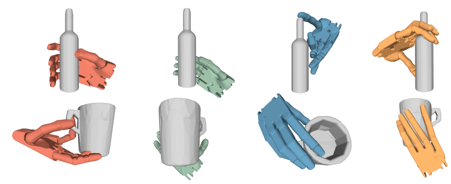

Fig. 9: Four types of failure cases in the generation process of UGG.

bump on the right for holding, further showcases the potential of our models in human-centric object design.

Joint Generation. Fig. 7 showcases the results of joint generation of objects and hands. We exclude the model trained on the entire dataset, as explained earlier, due to its consistent generation of cube-like objects. The visualization highlights the generation of a diverse array of objects and hand poses concurrently. This capacity contributes valuable additional data pairs to the grasping task, even when provided with a minimal initial dataset.

6.4 Ablation Study

Physics Discriminator. We conduct an ablation study on the physics discriminator to assess its performance and how the hyper-parameter helps to balance quality and diversity. We consider two settings: 1) Setting different thresholds for predicted probabilities and 2) Selecting a different number of poses by ranking the predicted probabilities. A good discriminator is expected to show an increasing success rate with an increase in the threshold and a decreasing success rate with a larger number of selected poses. We generated a total of 1000 poses for this study.

Fig. 8 presents the results in these settings. The performance of our physics discriminator aligns with the anticipated behavior of a competent discriminator. Noteworthy observations include that more than half of the generated poses have a confidence score exceeding 0.9, with more than 81% of them being successful. This is attributed to the robust nature of our generation module. Furthermore, for a specific grasping task, simultaneously generating multiple poses and selecting the highest confidence of the physics discriminator can significantly improve the success rate. We believe there is substantial potential in developing physics discriminators to assist in real-world pose generation tasks.

Ablation on Hand Encoder and Contact Anchor. We conducted an ablation study on our UGG model to assess each component's effectiveness. The baseline model is without hand joint encoding and the contact anchor. Another baseline includes hand joint encoding but not the contact anchor. Due to dataset complexity, we trained and tested three models on a smaller subset of 20 objects (details in the appendix). Each model was tested without the physics discriminator for a fair comparison of generative performance.

Tab. 2 shows that the model with both components outperforms two baselines in success rate and diversity. This superiority is attributed to the effectiveness of the proposed contact anchors, which provide valuable guidance for grasp generation during the generation and optimization phases. The model with the hand encoder improves diversity without significantly affecting the success rate. Encoding hand joint parameters acts like data augmentation, introducing increased diversity. The baseline without the hand encoder and contact anchor is similar to a variant of SceneDiffuser [19]. This ablation study highlights the substantial contributions of our proposed encoding schemes and contact modeling to the success of UGG.

Table 2: The effectiveness of different generative model components. Results are shown on a 20-object subset.

hand encoder	contact anchor	Quality success ↑	Q_1 ↑	pen ↓	Diversity H mean ↑	H std ↓
		54.3	0.061	0.15	7.92	0.24
✓		53.5	**0.072**	0.18	7.98	0.21
✓	✓	**59.9**	0.037	**0.16**	**7.99**	**0.19**

Limitation and Failure Case Analysis. Fig. 9 explores four key failure types in our generation results. One challenge is significant penetration between the hand and object (red), contrasting with cases where they are completely detached (green). We also observe grasps that lean more towards touch than a full grasp (blue), sometimes even fooling our physics discriminator. Additionally, we visualize instances where the discriminator erroneously categorizes successful grasps as failures (yellow). In general, most of the presented cases only require subtle adjustments to be valid. Although our methodology currently stands as the state-of-the-art in dexterous grasping, we acknowledge the need for future endeavors to refine the physics discriminator and strike a more optimal balance between penetration, simulated success, and grasp diversity.

7 Conclusions

This work tackles the task of dexterous grasping. By introducing a novel representation of contact information, *Contact Anchor*, for the first time, it models objects, hands, and contacts in a unified diffusion process. A physics discriminator is carefully designed to utilize the diversity of a generative model and further push the success rate of grasping. Combining all these novelties results in an effective framework that outperforms all state-of-the-art results on both success rate and diversity. Additionally, grasping generation from hand to object, or even both from scratch, becomes possible with the proposed unified design.

Acknowledgements

We would like to acknowledge NSF IIS-2047677, HDR-1934932, and CCF-2019844.

References

1. Bao, C., Xu, H., Qin, Y., Wang, X.: Dexart: Benchmarking generalizable dexterous manipulation with articulated objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21190–21200 (2023)
2. Bao, F., Li, C., Cao, Y., Zhu, J.: All are worth words: a vit backbone for score-based diffusion models. In: NeurIPS 2022 Workshop on Score-Based Methods (2022)
3. Bao, F., Nie, S., Xue, K., Li, C., Pu, S., Wang, Y., Yue, G., Cao, Y., Su, H., Zhu, J.: One transformer fits all distributions in multi-modal diffusion at scale. In: International Conference on Machine Learning (2023)
4. Berscheid, L., Meißner, P., Kröger, T.: Self-supervised learning for precise pick-and-place without object model. IEEE Robotics and Automation Letters **5**(3), 4828–4835 (2020)
5. Brahmbhatt, S., Ham, C., Kemp, C.C., Hays, J.: ContactDB: Analyzing and predicting grasp contact via thermal imaging. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019), `https://contactdb.cc.gatech.edu`
6. Brahmbhatt, S., Handa, A., Hays, J., Fox, D.: Contactgrasp: Functional multi-finger grasp synthesis from contact. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2386–2393. IEEE (2019)
7. Chai, S., Zhuang, L., Yan, F.: Layoutdm: Transformer-based diffusion model for layout generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18349–18358 (June 2023)
8. Cheng, Y.C., Lee, H.Y., Tulyakov, S., Schwing, A.G., Gui, L.Y.: Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4456–4465 (June 2023)
9. Chou, G., Bahat, Y., Heide, F.: Diffusion-sdf: Conditional generative modeling of signed distance functions. In: The IEEE International Conference on Computer Vision (ICCV) (2023)
10. Ciocarlie, M., Goldfeder, C., Allen, P.: Dexterous grasping via eigengrasps: A low-dimensional approach to a high-complexity problem. In: Robotics: Science and systems manipulation workshop-sensing and adapting to the real world (2007)
11. Dai, H., Majumdar, A., Tedrake, R.: Synthesis and optimization of force closure grasps via sequential semidefinite programming. In: ISRR (1). Springer Proceedings in Advanced Robotics, vol. 2, pp. 285–305. Springer (2015)
12. Dogar, M., Spielberg, A., Baker, S., Rus, D.: Multi-robot grasp planning for sequential assembly operations. Autonomous Robots **43**, 649–664 (2019)
13. Fan, P., Yan, B., Wang, M., Lei, X., Liu, Z., Yang, F.: Three-finger grasp planning and experimental analysis of picking patterns for robotic apple harvesting. Computers and Electronics in Agriculture **188**, 106353 (2021)
14. Grady, P., Tang, C., Twigg, C.D., Vo, M., Brahmbhatt, S., Kemp, C.C.: ContactOpt: Optimizing contact to improve grasps. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

15. Gupta, A., Xiong, W., Nie, Y., Jones, I., Oğuz, B.: 3dgen: Triplane latent diffusion for textured mesh generation. arXiv preprint arXiv:2303.05371 (2023)
16. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: CVPR (2020)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
18. Hu, J., Watkins, D., Allen, P.: Teleoperated robot grasping in virtual reality spaces (2023)
19. Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
20. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. In: Proceedings of the International Conference on Computer Vision (2021)
21. Karunratanakul, K., Spurr, A., Fan, Z., Hilliges, O., Tang, S.: A skeleton-driven neural occupancy representation for articulated hands. In: International Conference on 3D Vision (3DV) (2021)
22. Karunratanakul, K., Yang, J., Zhang, Y., Black, M.J., Muandet, K., Tang, S.: Grasping field: Learning implicit representations for human grasps. 2020 International Conference on 3D Vision (3DV) pp. 333–344 (2020)
23. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6007–6017 (June 2023)
24. Kim, U., Jung, D., Jeong, H., Park, J., Jung, H.M., Cheong, J., Choi, H.R., Do, H., Park, C.: Integrated linkage-driven dexterous anthropomorphic robotic hand. Nature communications **12**(1), 7177 (2021)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
26. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
27. Krug, R., Dimitrov, D., Charusta, K., Iliev, B.: On the efficient computation of independent contact regions for force closure grasps. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 586–591 (2010)
28. Li, K., Baron, N., Zhang, X., Rojas, N.: Efficientgrasp: A unified data-efficient learning to grasp method for multi-fingered robot hands. IEEE Robotics and Automation Letters **7**(4), 8619–8626 (2022)
29. Li, P., Liu, T., Li, Y., Geng, Y., Zhu, Y., Yang, Y., Huang, S.: Gendexgrasp: Generalizable dexterous grasping. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 8068–8074 (2023)
30. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 300–309 (June 2023)
31. Liu, H., Selvaggio, M., Ferrentino, P., Moccia, R., Pirozzi, S., Bracale, U., Ficuciello, F.: The musha hand ii:a multifunctional hand for robot-assisted laparoscopic surgery. IEEE/ASME Transactions on Mechatronics **26**(1), 393–404 (2021)

32. Liu, M., Pan, Z., Xu, K., Ganguly, K., Manocha, D.: Generating grasp poses for a high-dof gripper using neural networks. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1518–1525. IEEE (2019)
33. Liu, M., Pan, Z., Xu, K., Ganguly, K., Manocha, D.: Deep differentiable grasp planner for high-dof grippers. ArXiv **abs/2002.01530** (2020)
34. Liu, S., Zhou, Y., Yang, J., Gupta, S., Wang, S.: Contactgen: Generative contact modeling for grasp generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
35. Liu, T., Liu, Z., Jiao, Z., Zhu, Y., Zhu, S.C.: Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator. IEEE Robotics and Automation Letters **7**(1), 470–477 (2021)
36. Liu, Z., Feng, Y., Black, M.J., Nowrouzezahrai, D., Paull, L., Liu, W.: Meshdiffusion: Score-based generative 3d mesh modeling. In: International Conference on Learning Representations (2023)
37. Liu, Z., Tang, H., Lin, Y., Han, S.: Point-voxel cnn for efficient 3d deep learning. In: Advances in Neural Information Processing Systems (2019)
38. Lundell, J., Corona, E., Le, T.N., Verdoja, F., Weinzaepfel, P., Rogez, G., Moreno-Noguer, F., Kyrki, V.: Multi-fingan: Generative coarse-to-fine sampling of multi-finger grasps. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 4495–4501. IEEE (2021)
39. Lundell, J., Verdoja, F., Kyrki, V.: Ddgc: Generative deep dexterous grasping in clutter. IEEE Robotics and Automation Letters **6**(4), 6899–6906 (2021)
40. Luo, S., Hu, W.: Diffusion probabilistic models for 3d point cloud generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2837–2845 (2021)
41. Lyu, Z., Wang, J., An, Y., Zhang, Y., Lin, D., Dai, B.: Controllable mesh generation through sparse latent point diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 271–280 (2023)
42. Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., et al.: Isaac gym: High performance gpu-based physics simulation for robot learning. arXiv preprint arXiv:2108.10470 (2021)
43. Mandikal, P., Grauman, K.: Learning dexterous grasping with object-centric visual affordances. In: 2021 IEEE international conference on robotics and automation (ICRA). pp. 6169–6176. IEEE (2021)
44. Mayer, V., Feng, Q., Deng, J., Shi, Y., Chen, Z., Knoll, A.: Ffhnet: Generating multi-fingered robotic grasps for unknown objects in real-time. 2022 International Conference on Robotics and Automation (ICRA) pp. 762–769 (2022)
45. Miller, A., Allen, P.: Graspit! a versatile simulator for robotic grasping. IEEE Robotics and Automation Magazine **11**(4), 110–122 (2004)
46. Müller, N., Siddiqui, Y., Porzi, L., Bulo, S.R., Kontschieder, P., Nießner, M.: Diffrf: Rendering-guided 3d radiance field diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4328–4338 (2023)
47. Newbury, R., Gu, M., Chumbley, L., Mousavian, A., Eppner, C., Leitner, J., Bohg, J., Morales, A., Asfour, T., Kragic, D., et al.: Deep learning approaches to grasp synthesis: A review. IEEE Transactions on Robotics (2023)
48. Peng, S., Jiang, C.M., Liao, Y., Niemeyer, M., Pollefeys, M., Geiger, A.: Shape as points: A differentiable poisson solver. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)

49. Prattichizzo, D., Malvezzi, M., Gabiccini, M., Bicchi, A.: On the manipulability ellipsoids of underactuated robotic hands with compliance. Robotics and Autonomous Systems **60**(3), 337–346 (2012)
50. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
51. Rodriguez, A., Mason, M.T., Ferry, S.: From caging to grasping. The International Journal of Robotics Research **31**(7), 886–900 (2012)
52. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
53. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Trans. Graph. **36**(6) (nov 2017). `https://doi.org/10.1145/3130800.3130883`
54. Rosales, C., Suárez, R., Gabiccini, M., Bicchi, A.: On the synthesis of feasible and prehensile robotic grasps. In: 2012 IEEE international conference on robotics and automation. pp. 550–556. IEEE (2012)
55. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22500–22510 (June 2023)
56. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
57. Shadowhand. `https://www.shadowrobot.com/dexterous-hand-series/` (2005)
58. Shao, L., Ferreira, F., Jorda, M., Nambiar, V., Luo, J., Solowjow, E., Ojea, J.A., Khatib, O., Bohg, J.: Unigrasp: Learning a unified model to grasp with multi-fingered robotic hands. IEEE Robotics and Automation Letters **5**(2), 2286–2293 (2020)
59. Shue, J.R., Chan, E.R., Po, R., Ankner, Z., Wu, J., Wetzstein, G.: 3d neural field generation using triplane diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20875–20886 (June 2023)
60. Singh, J., Gould, S., Zheng, L.: High-fidelity guided image synthesis with latent diffusion models. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5997–6006. IEEE (2023)
61. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
62. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
63. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: European Conference on Computer Vision (ECCV) (2020), `https://grab.is.tue.mpg.de`
64. Turpin, D., Wang, L., Heiden, E., Chen, Y.C., Macklin, M., Tsogkas, S., Dickinson, S., Garg, A.: Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. In: European Conference on Computer Vision. pp. 201–221. Springer (2022)
65. Turpin, D., Zhong, T., Zhang, S., Zhu, G., Heiden, E., Macklin, M., Tsogkas, S., Dickinson, S., Garg, A.: Fast-grasp'd: Dexterous multi-finger grasp generation through differentiable simulation. In: ICRA (2023)

66. Urain, J., Funk, N., Chalvatzaki, G., Peters, J.: Se (3)-diffusionfields: Learning cost functions for joint grasp and motion optimization through diffusion. arXiv preprint arXiv:2209.03855 (2022)
67. Varley, J., Weisz, J., Weiss, J., Allen, P.: Generating multi-fingered robotic grasps via deep learning. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4415–4420 (2015)
68. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
69. Wan, W., Geng, H., Liu, Y., Shan, Z., Yang, Y., Yi, L., Wang, H.: Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. arXiv preprint arXiv:2304.00464 (2023)
70. Wang, R., Zhang, J., Chen, J., Xu, Y., Li, P., Liu, T., Wang, H.: Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 11359–11366. IEEE (2023)
71. Wei, Q.A., Ding, S., Park, J.J., Sajnani, R., Poulenard, A., Sridhar, S., Guibas, L.: Lego-net: Learning regular rearrangements of objects in rooms. arXiv preprint arXiv:2301.09629 (2023)
72. Wei, W., Li, D., Wang, P., Li, Y., Li, W., Luo, Y., Zhong, J.: Dvgg: Deep variational grasp generation for dextrous manipulation. IEEE Robotics and Automation Letters **7**(2), 1659–1666 (2022)
73. Xu, J., Wang, X., Cheng, W., Cao, Y.P., Shan, Y., Qie, X., Gao, S.: Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20908–20918 (June 2023)
74. Xu, M., Powers, A., Dror, R., Ermon, S., Leskovec, J.: Geometric latent diffusion models for 3d molecule generation. In: International Conference on Machine Learning. PMLR (2023)
75. Xu, X., Wang, Z., Zhang, G., Wang, K., Shi, H.: Versatile diffusion: Text, images and variations all in one diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7754–7765 (2023)
76. Xu, Y., Wan, W., Zhang, J., Liu, H., Shan, Z., Shen, H., Wang, R., Geng, H., Weng, Y., Chen, J., 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. pp. 4737–4746 (2023)
77. Zeng, X., Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K.: Lion: Latent point diffusion models for 3d shape generation. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
78. Zhang, H., Ye, Y., Shiratori, T., Komura, T.: Manipnet: Neural manipulation synthesis with a hand-object spatial representation. ACM Trans. Graph. **40**(4) (jul 2021). `https://doi.org/10.1145/3450626.3459830`
79. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
80. Zhou, L., Du, Y., Wu, J.: 3d shape generation and completion through point-voxel diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5826–5835 (2021)
81. Zhu, T., Wu, R., Hang, J., Lin, X., Sun, Y.: Toward human-like grasp: Functional grasp by dexterous robotic hand via object-hand semantic representation. IEEE

Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12521–12534 (2023)
82. Zhu, T., Wu, R., Lin, X., Sun, Y.: Toward human-like grasp: Dexterous grasping via semantic representation of object-hand. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15741–15751 (2021)