Grasp as You Say: Language-guided Dexterous Grasp Generation

Yi-Lin Wei[1], **Jian-Jian Jiang**[1], **Chengyi Xing**[2], **Xian-Tuo Tan**[1], **Xiao-Ming Wu**[1], **Hao Li**[2], **Mark Cutkosky**[2], **Wei-Shi Zheng**[1,3†]

[1] School of Computer Science and Engineering, Sun Yat-sen University, China

[2] Stanford University, USA

[3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China

{weiylin5, jiangjj35, tanxt23, wuxm65}@mail2.sysu.edu.cn

{chengyix, li2053, cutkosky}@stanford.edu wszheng@ieee.org

https://isee-laboratory.github.io/DexGYS/

Abstract

This paper explores a novel task **"Dexterous Grasp as You Say"** (DexGYS), enabling robots to perform dexterous grasping based on human commands expressed in natural language. However, the development of this field is hindered by the lack of datasets with natural human guidance; thus, we propose a language-guided dexterous grasp dataset, named **DexGYSNet**, offering high-quality dexterous grasp annotations along with flexible and fine-grained human language guidance. Our dataset construction is cost-efficient, with the carefully-design hand-object interaction retargeting strategy, and the LLM-assisted language guidance annotation system. Equipped with this dataset, we introduce the **DexGYSGrasp** framework for generating dexterous grasps based on human language instructions, with the capability of producing grasps that are intent-aligned, high quality and diversity. To achieve this capability, our framework decomposes the complex learning process into two manageable progressive objectives and introduce two components to realize them. The first component learns the grasp distribution focusing on intention alignment and generation diversity. And the second component refines the grasp quality while maintaining intention consistency. Extensive experiments are conducted on DexGYSNet and real world environments for validation.

1 Introduction

Enabling robots to perform dexterous grasping based on human language instructions is essential within the robotics and deep learning communities, offering promising applications in industrial production and domestic collaboration scenarios.

With the advancements in data-driven deep learning and the availability of large-scale datasets, robot dexterous grasp methods achieve impressive performance [1, 2, 3, 4, 5, 6, 7]. While previous approaches focus on the grasp stability, they have not fully utilized the potential of dexterous hands for intentional, human-like grasping. Recent studies, known as task-oriented [8] and functional dexterous grasping [9, 10], aim to generate grasps based on specific tasks or functionality of objects. However, these approaches often depend on predefined, fixed and limited tasks or functions, restricting their flexibility and hindering natural human-robot interaction.

In this paper, we explore a novel task, **"Dexterous Grasp as You Say"** (DexGYS), as shown in Figure 1. We can see that natural human guidance is provided in this task, and can be utilized to

†Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-1487

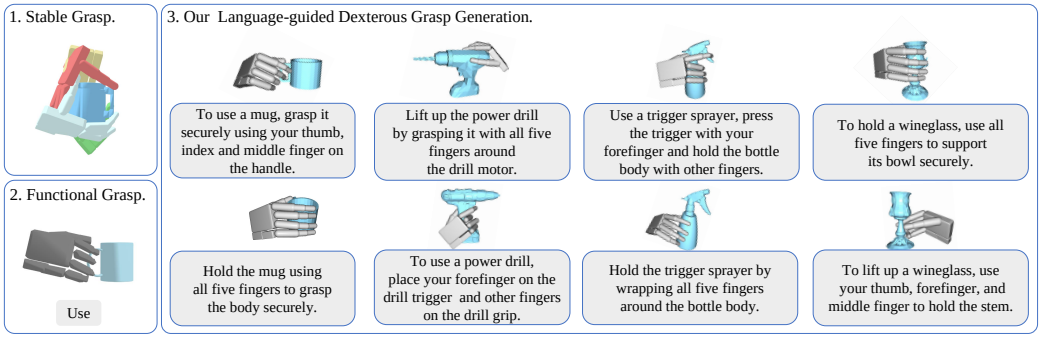


Figure 1: Our Language-guided Task vs. Traditional Dexterous Grasp Tasks. Traditional methods focus either solely on grasp quality or on fixed and limited functionalities. Our approach enables the generation of dexterous grasps based on human language, enhancing natural human-robot interactions.

drive dexterous grasping generation, thereby facilitating more user-friendly human-robot interactions. However, the new task also brings in new challenges. First, the high costs of annotating dexterous pose and the corresponding language guidance, present a barrier for developing and scaling dexterous datasets. Second, the demands of generating dexterous grasps that ensure intention alignment, high quality and diversity, present considerable challenges to the model learning.

To address the first challenge, we propose a large-scale language-guided dexterous grasping dataset **DexGYSNet**. DexGYSNet is constructed in a cost-effective manner by exploiting human grasp behavior and the extensive capabilities of Large Language Models (LLM). Specially, we introduce the Hand-Object Interaction Retargeting (HOIR) strategy to transfer easily-obtained human hand-object interactions to robotic dexterous hand, to maintain contact consistency and high-quality grasp posture. Subsequently, we develop the LLM-assisted Language Guidance Annotation system to produce flexible and fine-grained language guidance for dexterous grasp data with the support of LLM. DexGYSNet dataset comprises 50,000 pairs of high-quality dexterous grasps and their corresponding language guidance, on 1,800 common household objects.

With the support of the dataset, we now turn our way to overcome the second challenge. We propose the **DexGYSGrasp** framework for dexterous grasp generation, which aligns with intentions, ensures high quality, and maintains diversity. At the beginning, we find the difficulty of mastering all objectives simultaneously results from the commonly used penetration loss [7] which used to avoid hand-object penetration. As shown in Figure 2, penetration loss substantially hinders the learning of grasp distribution, causing intention misalignment and reduced diversity. Conversely, despite the high diversity and aligned intention, the removal of penetration loss leads to unacceptable object penetration, making the grasp infeasible. Based on this finding, we design our DexGYSGrasp framework in a progressive strategy, decomposing the complex learning task into two sequential objectives managed by progressive components. Initially, the first component learns a grasp distribution, which focuses on intention consistency and diversity, optimizing effectively without the constraints of penetration loss. Subsequently, the second component refines the initial coarse grasps to high-quality ones with the same intentions and diversity. Our framework allows each component to focus on specific and manageable optimization objective, enhancing the overall performance of the generated grasps.

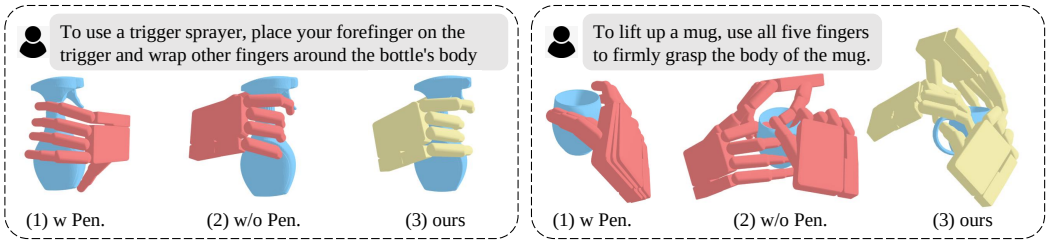
Extensive experiments are conducted on the DexGYSNet dataset and real-world scenarios. The results demonstrate that our methods are capable of generating intention-consistent, high diversity and high quality grasp poses for a wide range of objects.

2 Related work

2.1 Dexterous Grasp Generation

Dexterous hand endows robots with the capability to manipulate objects in a human-like manner. Previous methods have achieved impressive results in ensuring grasp stability by analytical approaches [11, 12, 13, 14, 15] and deep learning methods [5, 3, 7, 16, 17, 18]. However, the full potential of dexterous hands for intentional and human-like grasping has not been completely exploited

https://doi.org/10.52202/079017-1487



(a) Comparison for Intention Consistency and Quality

(b) Comparison for Grasp Diversity and Quality.

Figure 2: Visualization of the impact of penetration loss (Pen. in the figure) on grasp performance: intention alignment, quality, and diversity. (a) illustrates penetration loss **causes intention misalignment** and its absence results in severe object penetration. (b) shows three sampling results under the same conditions, and demonstrates that penetration loss **leads to reduced diversity**.

in these methods. Recently, some works have focused on functional dexterous grasping [8, 9, 10, 19], aiming to achieve human-like capabilities that extend beyond grasp stability alone, but are still lack of flexibility and generalization. In this work, we explore a novel task, Language-guided Dexterous Grasp Generation, which fully leverages the dexterity of robotic hands and enable robot to execute dexterous grasp based on human natural language.

2.2 Grasp Datasets

The development of large-scale datasets has significantly improve the advancement of data-driven grasp methods, including parallel grasp [20, 21, 22, 23, 24], human grasp [25, 26, 27, 28, 29, 30], and dexterous grasp approaches [1, 2, 8, 14, 16]. Despite these advancements, the high cost of data collection remains a significant challenge, particularly in the domain of dexterous hands. Previous datasets for dexterous grasping primarily rely on physical analysis approaches [12, 31] to mitigate this issue. However, these approaches often lack the specific semantic context or corresponding language guidance necessary for constructing our language-guided dexterous task. In this paper, we present DexGYSNet dataset, with a cost-effective construction, providing high-quality dexterous grasp annotation along with flexible and fine-grained human language guidance.

2.3 Language-guided Robot Grasp

Language-guided robot grasp is important in robotics. Previous works focusing on parallel grippers have made strides in achieving task-oriented grasping [32, 23, 33], language-guided grasping [34, 35] and manipulation [36, 37, 38, 39]. In contrast to parallel grippers, dexterous hand boast a higher number of DOF (e.g., 28 for the Shadow Hand [40]), enabling a broader dexterity. However, this high freedom also presents challenges for model learning. In this paper, we propose the DexGYSGrasp framework, capable of generating intention-aligned dexterous grasps with high-quality and diversity.

3 DexGYSNet Dataset

3.1 Dataset Overview

The DexGYSNet dataset is constructed with a cost-effective strategy, as shown in Figure 3. We first collect object meshes and human grasps data from existing datasets [27]. Subsequently, we develop the Hand-Object Interaction Retargeting (HOIR) strategy to transform human grasps into dexterous grasps with high quality and hand-object interaction consistency. Finally, we implement an LLM-assisted Language Guidance Annotation system, which leverages the knowledge of Large Language Models (LLM) to produce flexible and fine-grained annotations for language guidance.

3.2 Hand-Object Interaction Retargeting

Our Hand-Object Interaction Retargeting (HOIR) aims to transfer human hand-object interaction to dexterous hand-object interaction as shown in Figure 3 . The source MANO [41] hand parameters are denoted as $\mathcal{G}^m \in R^{61}$. And the target dexterous hand parameters are denoted as $\mathcal{G}^{dex} = (r, t, q)$,

https://doi.org/10.52202/079017-1487

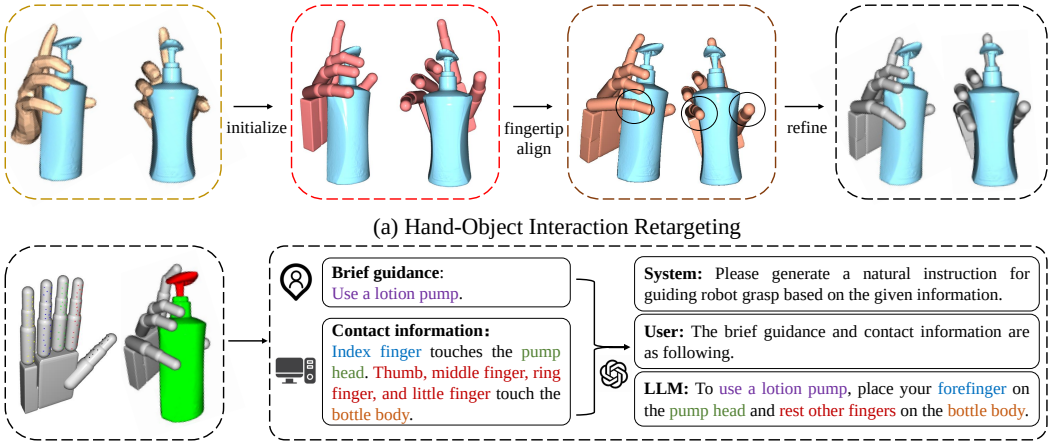


Figure 3: The construction process of the DexGYSNet dataset. (a) The HOIR strategy retargets the human hand to the dexterous hand by three step, maintaining hand-object interaction consistency and avoiding physical infeasibility (shown in black circle). (b) The annotation system automatically annotates language guidance for hand-object pairs with the help of LLM.

where $r \in \mathbf{SO(3)}$ represents the global rotation, $t \in \mathbb{R}^3$ is the translation in world coordinates, and $q \in \mathbb{R}^J$ is the joint angles for a J-DoF dexterous hand, for example $J = 22$ for Shadow Hand[40].

Three steps are within the HOIR: pose initialization, fingertip alignment, and interaction refinement. In the first step, the dexterous poses are initialized by copying parameters from similar structures of human poses to establish better initial values. In the second step, the dexterous poses are optimized in the parameter space to align the fingertip positions $p_k^{dex,ft}$ with those of the human $p_k^{mano,ft}$. This achieves retargeting consistency, and the optimization objective can be formulated as follows:

$$\min_{\mathcal{G}^{dex}=(r,t,q)} \sum_k \|p_k^{dex,ft} - p_k^{mano,ft}\|_2^2. \tag{1}$$

To improve the physical interaction feasibility while maintaining the consistency, the dexterous hand poses are further optimized in the third step by hand-object interaction and physical constraints losses [14]. Two key points are designed to maintain the consistency: preserving the contact area of the optimized pose consistent with the output from the second step, and keeping the translation fixed during this step. The optimization objective can be formulated as follows:

$$\min_{(r,q)} (\lambda_{pen}^1 \mathcal{L}_{pen} + \lambda_{spen}^1 \mathcal{L}_{spen} + \lambda_{joint}^1 \mathcal{L}_{joint} + \lambda_{cmap}^1 \mathcal{L}_{cmap}). \tag{2}$$

Here, the object penetration loss $\mathcal{L}_{pen}$ penalizes the depth of hand-object penetration. The self-penetration loss $\mathcal{L}_{spen}$ penalizes the self-penetration. The joint angle loss $\mathcal{L}_{joint}$ penalizes the out-of-limit joint angles. The contact map loss $\mathcal{L}_{cmap}$ ensures the contact map on the object remains consistent with the output from the second stage. The details of losses can be found in Appendix A.1.5.

3.3 LLM-assisted Language Guidance Annotation

To annotate flexible and fine-grained language guidance for dexterous hand-object pairs with low-cost, we design a coarse-to-fine automated language guidance annotation system with the assistance of the LLM, inspired by [42, 29], as shown in Figure 3. Specially, we initially generate brief guidance based on the object category and the brief human intention (e.g., "using a lotion pump"), which are collected by the human dataset [27]. Subsequently, we compile the contact information for each finger by calculating the distances from the contact anchors on the hand to different parts of the object. We then organize the contact information into language descriptors (e.g. "forefinger touches pump head and other fingers touch the bottle body."). Finally, we input both the brief guidance and the detailed contact information into the GPT3.5 to produce natural annotated guidance (e.g. "To use a lotion pump, press down on the pump head with your forefinger while holding the bottle with your other fingers."). More details about DexGYSNet construction can be found in Appendix A.2.

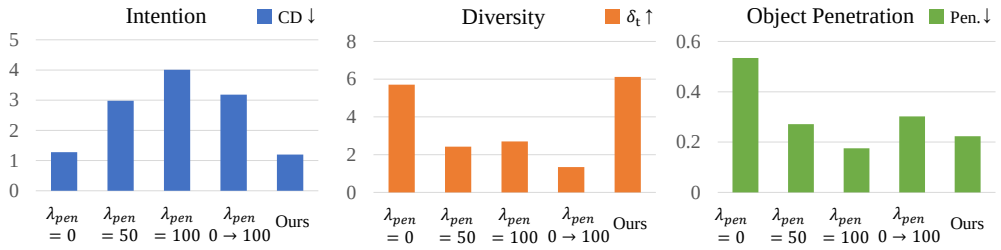


Figure 4: Quantitative experimental results with different object penetration loss weights λ_{pen}. Intention is quantified by the Chamfer distance (CD) between predictions and targets. Diversity is assessed by the standard deviation of hand translation δ_t. Object penetration is evaluated by the penetration depth (Pen.) from the object point cloud to the hand mesh. Our method uniquely achieves high performance in terms of intention consistency, diversity, and penetration avoidance.

4 DexGYSGrasp framework

Given full object point clouds $\mathcal{O}$ and language guidance $\mathcal{L}$ as inputs, our goal is to generate dexterous grasps $\mathcal{G}^{dex}$ with intention alignment, high diversity and high quality.

4.1 Progressive Grasp Objectives.

Learning Challenge in DexGYS. The DexGYS places high demands on intention alignment (e.g., accurately pressing your forefinger on trigger to use the sprayer), high diversity (e.g., holding the bottle using various postures), and high quality (e.g., ensuring stable grasp and avoiding object penetration). However, we find that a single model struggles to meet these requirements simultaneously, due to the optimization challenge caused by the commonly used object penetration loss [43, 14, 16], which is used to prevent hand-object penetration. As shown in Figure 2 and Figure 4, increasing the weight of the penetration loss reduces object penetration but adversely affects intention alignment and generation diversity.

Progressive Grasp Objectives. To address these challenges, we propose to decompose the complex learning objective into two more manageable objectives. The first objective is generative: it focuses on learning the grasp distribution, which does not prioritize quality but focuses on learning the grasp distribution with intention alignment and generation diversity. The second objective is regressive: it aims to refine the coarse grasp to a specific high-quality grasp with same intention. By decomposing the complex objectives, we reduce the learning difficulty of the generative objective as it does not concentrate on quality and avoids using penetration loss which could interfere the learning process. Additionally, the learning of regression is less complex than distributions, as it merely requires adjusting the pose to a specific target within a small space. Hence, we can employ penetration loss to ensure that the refined dexterous hand avoids penetrating the object and with high quality.

4.2 Progressive Grasp Components

Benefiting from our progressive grasp objectives in Section 4.1, we design the following two simple progressive grasp components, which can achieve intention alignment, high diversity and high quality language-guided dexterous generation.

Intention and Diversity Grasp Component. We introduce intention and diversity grasp component to learn a grasp distribution efficiently, achieve intention aligned and diverse generation. Due to the distribution modeling objective, IDGC is build upon the conditional diffusion model [44, 4] to predict the dexterous pose $\mathcal{G}_0^{dex}$ from noised $\mathcal{G}_T^{dex}$. The input object point clouds $\mathcal{O}$ is encoded by Pointnet++ [45] and language $\mathcal{L}$ is encoded by a pretrained CLIP model [46] as the condition. And we employ DDPM [47] as sampling process, which can be formalized by the following equation:

$$p_\theta\left(\mathcal{G}_0^{dex}|\mathcal{O},\mathcal{L}\right) = p\left(\mathcal{G}_T^{dex}\right)\prod_{t=1}^{T} p\left(\mathcal{G}_{t-1}^{dex}|\mathcal{G}_t^{dex},\mathcal{O},\mathcal{L}\right). \tag{3}$$

https://doi.org/10.52202/079017-1487

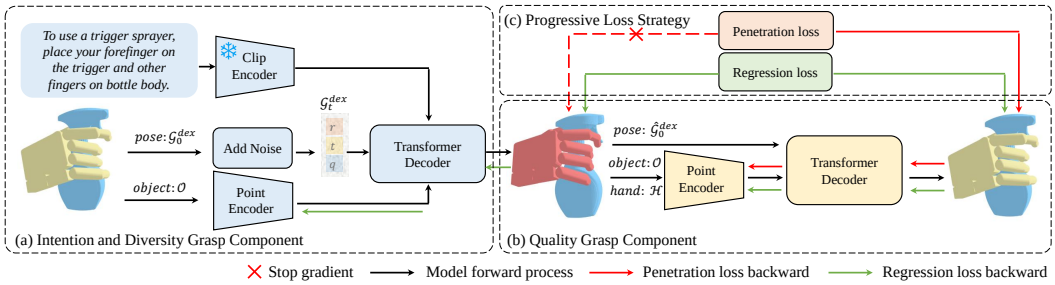


Figure 5: Overview of our framework. (a) With only the regression loss, intention and diversity grasp component is trained to reconstruct the original hand pose from the noise poses, based on language and object condition. (b) With both regression and penetration losses, Quality Grasp Component is trained to refine the coarse pose improve the grasp quality while maintain intension consistency.

Quality Grasp Component. The generated grasps of the first component possess well-aligned intentions and high diversity, but suffer from poor grasp quality due to significant object penetration. Therefore, we introduce Quality Grasp Component to refine the grasp quality while maintaining intention consistency in a regressive manner. Specially, it takes the coarse pose $\hat{\mathcal{G}}^{dex}$, coarse hand point clouds $\mathcal{H}(\hat{\mathcal{G}}^{dex})$ and object point clouds $\mathcal{O}$ as input, and outputs the pose $\Delta\mathcal{G}^{dex}$. The refined grasp is obtained by $\tilde{\mathcal{G}}^{dex} = \hat{\mathcal{G}}^{dex} + \Delta\mathcal{G}^{dex}$. The training pairs of this component are constructed by collecting coarse grasps generated by the first component alongside the most similar ground-truth grasps that share the similar intentions. This ensures the training targets are aligned with the language intention, thereby guaranteeing that the refined grasps maintain consistency with the intended actions.

4.3 Progressive Grasp Loss

Intention and Diversity Grasp Loss. We strategically employ regression losses and exclude object penetration loss to enhance the training efficacy of intention and diversity grasp component. By focusing exclusively on the regression learning, this component facilitates a more effective optimization process, achieving enhancements of intention consistency and grasp diversity. Concretely, we utilize L2 loss for pose parameter regression and incorporate the hand chamfer loss [48] to assist by explicit hand shape. The loss function of intention and diversity grasp component.is defined as:

$$\mathcal{L}_{IDG} = \lambda^2_{para}\mathcal{L}_{para}(\mathcal{G}_0^{dex}, \hat{\mathcal{G}}^{dex}) + \lambda^2_{chamfer}\mathcal{L}_{chamfer}(\mathcal{H}(\mathcal{G}_0^{dex}), \mathcal{H}(\hat{\mathcal{G}^{dex}})), \tag{4}$$

where $\mathcal{H}$ are dexterous hand point clouds of corresponding pose.

Quality Grasp Loss. Benefiting from the simplified training objectives, the quality grasp component focuses solely on refining coarse grasp to a specific target within a relatively constrained space, thereby reducing the negative impact of object penetration. Therefore, we employ the well-designed loss including object penetration. The loss function of quality grasp component can be formulated as:

$$\mathcal{L}_{QG} = \lambda^3_{para}\mathcal{L}_{para} + \lambda^3_{chamfer}\mathcal{L}_{chamfer} + \lambda^3_{pen}\mathcal{L}_{pen} + \lambda^3_{cmap}\mathcal{L}_{cmap} + \lambda^3_{spen}\mathcal{L}_{spen}. \tag{5}$$

More details about loss function and model structure can be found in Appendix A.1.

5 Experiments

5.1 Datasets and Evaluation Metrics

We split the DexDYSNet dataset at the object instance level, using 80% of the objects within each category for training and 20% for evaluation. Notably, none of the objects in the test set appear in the training set, ensuring that all experimental results are evaluated on unseen objects.

Three types of metrics are employed for evaluation from the perspective of intention consistency, grasp quality and grasp diversity. 1) For intention consistency, we employ **Fréchet Inception Distance** (FID), using sampling point cloud features extracted from [49] to calculate $P\text{-}FID$ and rendering image features extracted from [50] to calculate FID. Additionally, **Chamfer distance**

Method	Intention				Quality			Diversity		
	$FID \downarrow$	$P\text{-}FID \downarrow$	$CD \downarrow$	$Con. \downarrow$	$Success \uparrow$	$Q_1 \uparrow$	$Pen. \downarrow$	$\delta_t \uparrow$	$\delta_r \uparrow$	$\delta_q \uparrow$
GraspCVAE[51]	31.26	29.02	3.138	0.096	29.12%	0.054	0.551	0.179	1.762	0.179
GraspTTA[43]	35.41	33.15	12.19	0.111	43.46%	0.071	0.188	2.111	6.150	3.869
SceneDiffuser[4]	20.44	7.932	1.679	0.045	62.24%	0.083	0.253	0.346	3.455	0.387
DGTR[7]	23.31	15.77	2.895	0.078	51.91%	0.078	**0.163**	2.037	14.01	4.299
Ours	**6.538**	**5.595**	**1.198**	**0.036**	**63.31%**	**0.083**	0.223	**6.118**	**55.68**	**6.118**

Table 1: Results on DexGYSNet compared with the SOTA methods.

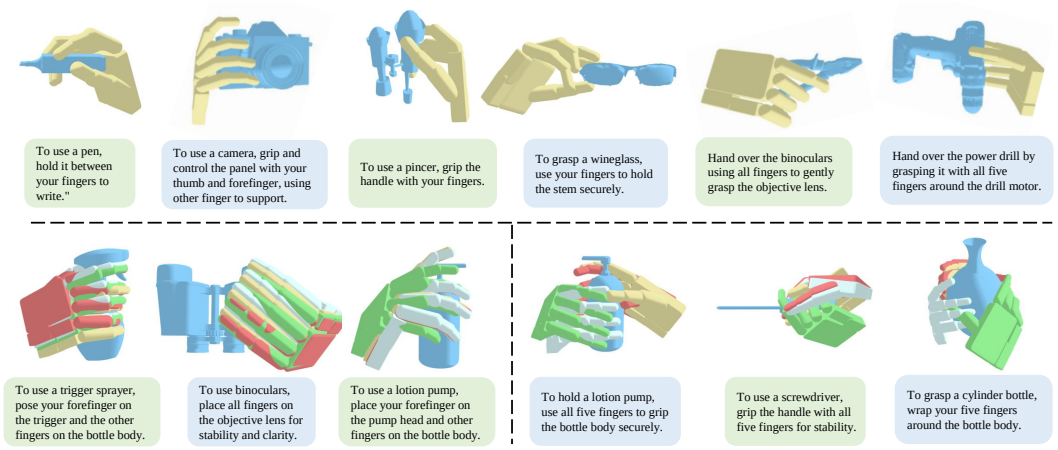


Figure 6: Visualization of generated dexterous grasp. The **top** visualizes one sample for each object and guidance pair. The bottom visualizes four samples, the **bottom left** shows that the generated grasp are consistent with clear and specific guidance, while the **bottom right** shows that the diversity achieved under relatively ambiguous instructions.

(CD), is used to measure the distance between predicted hand point clouds and targets; **Contact distance** ($Con.$) is used to measure the L2 distance of object contact map between the prediction and targets. 2) For grasp quality, **Success rate** in Issac gym and $\mathbf{Q_1}$ [13] measure grasp stability. We set the contact threshold to $1\,\mathrm{cm}$ and set the penetration threshold to $5\,\mathrm{mm}$ following [14]. **Maximal penetration depth** (cm), denoted as $Pen.$, reflects the maximal penetration depth from the object point cloud to hand meshes. 3) For diversity, we employ the **Standard deviation** of translation δ_t, rotation δ_r and joint angle δ_q of eight samples within same condition, following [7]. More details can be found in Appendix A.3.2.

5.2 Implementation Details

For the construction of DexGYSNet, the step 2 and 3 are optimized for 20 and 300 iterations with learning rates of 0.01 and 0.0001 respectively. We set $\lambda^1_{pen} = 100$ and set λ^1_{spen}, λ^1_{joint}, λ^1_{cmap} each to 10. For training our framework, the training epochs are set to 100 for intention and diversity grasp component and 20 for Quality Grasp Component. The loss weights are configured as follows: $\lambda^2_{para} = \lambda^3_{para} = 10$, $\lambda^2_{chamfer} = \lambda^3_{chamfer} = 1$, $\lambda^3_{cmap} = 10$, $\lambda^3_{pen} = 100$, $\lambda^3_{spen} = 10$. Throughout all training processes, the model is optimized with a batch size of 64 using the Adam optimizer, with a weight decay rate of 5.0×10^{-6}. The initial learning rate is 2.0×10^{-4} and decay to 2.0×10^{-5} using a cosine learning rate [52] scheduler. All experiment are implemented with PyTorch on a single RTX 4090 GPU.

5.3 Comparison with SOTA methods

The comparison results are presented in Table 1. We reproduce the SOTA methods to suit our task by concatenating the language condition with the point cloud features, the details can be found in Appendix A.3.3. As seen in the Table, our framework significantly outperforms all previous

https://doi.org/10.52202/079017-1487

	Intention		Quality		Diversity		
	$CD \downarrow$	$Con. \downarrow$	$Q_1 \uparrow$	$Pen. \downarrow$	$\delta_t \uparrow$	$\delta_r \uparrow$	$\delta_q \uparrow$
IDGC ($\lambda^2_{pen} = 0$)	1.276	0.028	0.024	0.534	5.710	54.75	7.741
IDGC ($\lambda^2_{pen} = 50$)	2.980	0.061	0.074	0.271	2.421	33.27	3.391
IDGC ($\lambda^2_{pen} = 100$)	4.009	0.067	0.072	0.175	2.701	38.27	3.785
IDGC ($\lambda^2_{pen} = 500$)	4.185	0.072	0.107	0.037	0.547	8.807	0.481
IDGC ($\lambda^2_{pen} = 0 \rightarrow 100$)	3.181	0.056	0.093	0.302	1.341	16.53	2.211
IDGC ($\lambda^2_{pen} = 0$) + TTA	20.09	0.102	0.057	0.178	4.849	51.91	8.479
IDGC ($\lambda^2_{pen} = 100$) + QGC	2.009	0.042	0.099	0.143	3.414	40.35	2.844
IDGC ($\lambda^2_{pen} = 0$) + QGC	1.198	0.036	0.083	0.223	6.118	55.68	6.118

Table 2: Ablation study for our framework. Intention and diversity grasp component is abbreviated as IDGC, Quality Grasp Component is abbreviated as QGC. λ^2_{pen} is the penetration loss weight tn the training of IDGC. Ours is colored in gray.

step1	step2	step3	Intention $Con. \downarrow$	Quality $Q_1 \uparrow$	Quality $Pen \downarrow$
✓			0.048	0.037	0.572
	✓		0.101	0.833	0.516
✓	✓		0.012	0.029	0.477
✓	✓	✓	0.015	0.063	0.369
all in one stage			0.075	0.090	0.271
w/o fix translation			0.051	0.074	0.332

Table 3: Ablation study for HOIR.

	Intention $CD \downarrow$	Intention $Con. \downarrow$	Quality $Q_1 \uparrow$	Quality $Pen \downarrow$	Diversity $\delta_t \uparrow$
GraspCVAE	3.138	0.096	0.054	0.551	0.179
+ w/o $\mathcal{L}_{pen}$	2.638	0.090	0.005	0.921	0.266
+ QGC	2.425	0.056	0.074	0.261	0.315
SceneDiffuser	1.679	0.045	0.083	0.225	0.346
+ w/o $\mathcal{L}_{pen}$	1.511	0.041	0.019	0.488	6.323
+ QGC	1.495	0.040	0.082	0.241	6.015

Table 4: Plug-and-play Experiments.

methods in terms of intention consistency and grasp diversity, while also achieving comparable performance in grasp quality. Previous methods struggle with learning a robust language conditional grasp distribution due to the optimization challenges outlined in Section 4.1. They often yield misaligned yet high quality grasps, resulting in comparable grasp quality, but less aligned intention and limited diversity compared to our framework. Overall, these results confirm that our framework achieves SOTA performance in generating intention-aligned, high-quality and diverse grasps.

In Figure 6, we visualize the generated grasp to qualitatively demonstrate the grasp generation capabilities of our framework. The bottom figure visualizes the results of four samples, the bottom left highlights our framework's ability to produce precise and consistent grasps under deterministic guidance (e.g., the way to use a trigger sprayer is deterministic). In the other hand, the bottom right illustrates our framework's diversity in generating grasps when provided with ambiguous guidance (e.g., the way to hold a bottle is diverse).

5.4 Necessity of Progressive Components and Losses

The results presented in Table 2 validate the core insight of our framework: decomposing the complex task into progressive objectives, employing progressive components, and learning with progressive losses. The initial four lines of results demonstrate that a single component, without progressive objectives, fails to balance all objectives. Moreover, a single component, even with progressive objectives, that adjusts λ^2_{pen} from 0 to 100 after several training epochs, does not enhance performance. The similar result occurs when using progressive components without corresponding progressive losses, $IDGC(\lambda^2_{pen} = 100) + QGC$. Moreover, the commonly used quality refinement strategy test-time adaptation (TTA) [43], though improves grasp quality but results in extremely poor intention consistency. Overall, only the progressive designs of our DexGYSGrasp framework ensures excellence in intention alignment, high quality and diversity.

5.5 Plug-and-play Experiments

We conducted experiments to evaluate the applicability of our insights to other state-of-the-art (SOTA) methods. Specifically, we trained GraspCAVE and SceneDiffuser without the object penetration

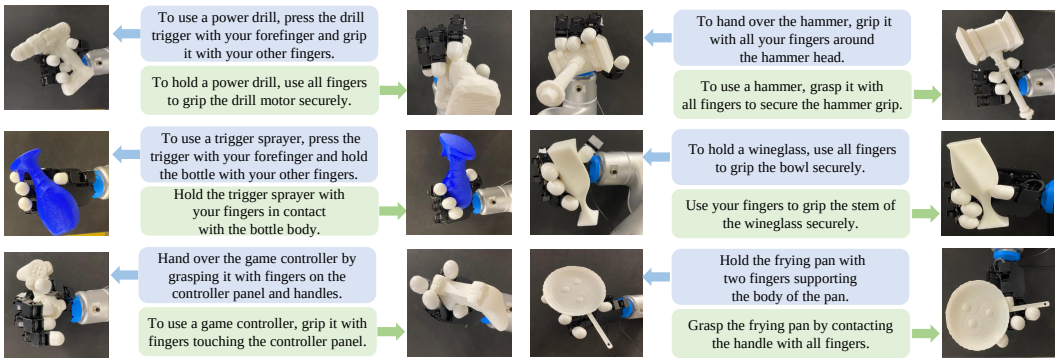


Figure 7: Visualization of real world experiments.

constraint and trained the quality grasp component (QGC) to refine the coarse outcomes. As depicted in Table 4, removing the object penetration loss leads to improved intention consistency, which corroborates our findings discussed in Section 4.1. Moreover, our quality grasp component can significantly enhance grasp quality while maintaining the intention consistency.

5.6 Effectiveness of Hand-Object Interaction Retargeting

We conducted ablation studies to evaluate our Hand-Object Interaction Retargeting (HOIR) strategy in constructing DexGYSNet dataset. As shown in Table 3, our three-step HOIR significantly improves both the quality and the intention consistency progressively. We observed that optimizing all losses in Equations 1 and 2 in one step (*all in one stage*), results in worse contact consistency and better grasp quality. Similar outcomes occur when the root translation is not fixed in step 3 (*w/o fix translation*). We believe this trade-off arises from inherent noise in the hand-object interaction data and the structural differences between human grasps and dexterous hands, making it challenging to excel in all aspects. Overall, we think that three-step HOIR strategy achieves more comprehensive outcomes, especially in the most important aspect of hand object contact consistency.

5.7 Experiments in Real World

We conducted real-world grasp experiments to verify the practical application of our methods, as shown in Figure 7. The experiments are conducted on an Allegro hand, a Flexiv Rizon 4 arm and an Intel Realsense D415 camera. Although our framework is designed for full object point clouds, we integrate several off-the-shelf methods to enhance its practicality. Specifically, partial object point clouds are obtained through visual grounding [53] and SAM [54], which are then fed into a point cloud completion network [55] to obtain full point clouds. In execution, we first move the arm to the 6-DOF pose of the dexterous hand root node, and then control the dexterous hand joint angles to the predicted poses. Real world experiments further validate the effectiveness of our method. More implementation details can be found in Appendix A.5.

6 Conclusions

We believe that enabling robots to perform high quality dexterous grasps aligned with human language is crucial within the deep learning and robotics communities. In this paper, we explore this novel task, “Dexterous Grasp as You Say” (DexGYS). This task is non-trival, we propose a DexGYSNet dataset and a DexGYSGrasp framework to accomplish it. DexGYSNet dataset is constructed cost-effectively using the object-hand interaction retargeting strategy and the language guidance annotation system assisted by LLMs. Building on DexGYSNet, DexGYSGrasp framework, comprised of two progressive components, which can achieve intention-aligned, high diversity, and high quality dexterous grasp generation. Extensive experiments in DexGYSNet and real-world settings demonstrate that our framework significantly outperforms all SOTA methods, confirming the potential and effectiveness of our approach.

https://doi.org/10.52202/079017-1487

Acknowledgements

This work was supported partially by NSFC(92470202, U21A20471), Guangdong NSF Project (No. 2023B1515040025). Additionally, I sincerely thank the help of Guo-Hao Xu and Dian Zheng for the valuable suggestions for the paper.

References

[1] Min Liu, Zherong Pan, Kai Xu, Kanishka Ganguly, and Dinesh Manocha. Deep differentiable grasp planner for high-dof grippers. *arXiv preprint arXiv:2002.01530*, 2020.

[2] Puhao Li, Tengyu Liu, Yuyang Li, Yiran Geng, Yixin Zhu, Yaodong Yang, and Siyuan Huang. Gendexgrasp: Generalizable dexterous grasping. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[3] Lin Shao, Fabio Ferreira, Mikael Jorda, Varun Nambiar, Jianlan Luo, Eugen Solowjow, Juan Aparicio Ojea, Oussama Khatib, and Jeannette Bohg. Unigrasp: Learning a unified model to grasp with multifingered robotic hands. *IEEE Robotics and Automation Letters*, 2020.

[4] 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 (CVPR)*, 2023.

[5] Jiaxin Lu, Hao Kang, Haoxiang Li, Bo Liu, Yiding Yang, Qixing Huang, and Gang Hua. Ugg: Unified generative grasping. *arXiv preprint arXiv:2311.16917*, 2023.

[6] Zehang Weng, Haofei Lu, Danica Kragic, and Jens Lundell. Dexdiffuser: Generating dexterous grasps with diffusion models. *arXiv preprint arXiv:2402.02989*, 2024.

[7] Guo-Hao Xu, Yi-Lin Wei, Dian Zheng, Xiao-Ming Wu, and Wei-Shi Zheng. Dexterous grasp transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[8] Jiayi Chen, Yuxing Chen, Jialiang Zhang, and He Wang. Task-oriented dexterous grasp synthesis via differentiable grasp wrench boundary estimator. *arXiv preprint arXiv:2309.13586*, 2023.

[9] 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.

[10] Wei Wei, Peng Wang, and Sizhe Wang. Generalized anthropomorphic functional grasping with minimal demonstrations. *arXiv preprint arXiv:2303.17808*, 2023.

[11] Jean Ponce, Steve Sullivan, J-D Boissonnat, and J-P Merlet. On characterizing and computing three- and four-finger force-closure grasps of polyhedral objects. In *[1993] Proceedings IEEE International Conference on Robotics and Automation*, pages 821–827. IEEE, 1993.

[12] Tengyu Liu, Zeyu Liu, Ziyuan Jiao, Yixin Zhu, and Song-Chun Zhu. 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.

[13] Carlo Ferrari, J Canny, et al. Planning optimal grasps. In *Proceedings., 1992 IEEE International Conference on Robotics and Automation, 1992.*, 1992.

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

[15] Dylan Turpin, Liquan Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. In *European Conference on Computer Vision*, 2022.

[16] 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*, 2023.

[17] Kelin Li, Nicholas Baron, Xian Zhang, and Nicolas Rojas. Efficientgrasp: A unified data-efficient learning to grasp method for multi-fingered robot hands. *IEEE Robotics and Automation Letters*, 2022.

[18] Jacob Varley, Jonathan Weisz, Jared Weiss, and Peter Allen. Generating multi-fingered robotic grasps via deep learning. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[19] Tianqiang Zhu, Rina Wu, Xiangbo Lin, and Yi Sun. Toward human-like grasp: Dexterous grasping via semantic representation of object-hand. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15741–15751, 2021.

[20] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11444–11453, 2020.

[21] Clemens Eppner, Arsalan Mousavian, and Dieter Fox. Acronym: A large-scale grasp dataset based on simulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6222–6227. IEEE, 2021.

[22] Xiao-Ming Wu, Jiafeng Cai, Jian-Jian Jiang, Dian Zheng, Yi-Lin Wei, and Wei-Shi Zheng. An economic framework for 6-dof grasp detection. In *European Conference on Computer Vision*, 2024.

[23] Chao Tang, Dehao Huang, Lingxiao Meng, Weiyu Liu, and Hong Zhang. Task-oriented grasp prediction with visual-language inputs. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4881–4888. IEEE, 2023.

[24] Jia-Feng Cai, Zibo Chen, Xiao-Ming Wu, Jian-Jian Jiang, Yi-Lin Wei, and Wei-Shi Zheng. Real-to-sim grasp: Rethinking the gap between simulation and real world in grasp detection. In *Conference on Robot Learning*, 2024.

[25] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11807–11816, 2019.

[26] 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.

[27] 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.

[28] 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.

[29] Kailin Li, Jingbo Wang, Lixin Yang, Cewu Lu, and Bo Dai. Semgrasp: Semantic grasp generation via language aligned discretization. *arXiv preprint arXiv:2404.03590*, 2024.

[30] Yan-Kang Wang, Chengyi Xing, Yi-Lin Wei, Xiao-Ming Wu, and Wei-Shi Zheng. Single-view scene point cloud human grasp generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[31] Andrew T Miller and Peter K Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004.

[32] Adithyavairavan Murali, Weiyu Liu, Kenneth Marino, Sonia Chernova, and Abhinav Gupta. Same object, different grasps: Data and semantic knowledge for task-oriented grasping. In *Conference on Robot Learning*, pages 1540–1557. PMLR, 2021.

[33] Chao Tang, Dehao Huang, Wenqi Ge, Weiyu Liu, and Hong Zhang. Graspgpt: Leveraging semantic knowledge from a large language model for task-oriented grasping. *IEEE Robotics and Automation Letters*, 2023.

[34] Kechun Xu, Shuqi Zhao, Zhongxiang Zhou, Zizhang Li, Huaijin Pi, Yifeng Zhu, Yue Wang, and Rong Xiong. A joint modeling of vision-language-action for target-oriented grasping in clutter. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11597–11604. IEEE, 2023.

[35] Shiyu Jin, Jinxuan Xu, Yutian Lei, and Liangjun Zhang. Reasoning grasping via multimodal large language model. *arXiv preprint arXiv:2402.06798*, 2024.

 https://doi.org/10.52202/079017-1487

[36] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022.

[37] Oier Mees, Lukas Hermann, and Wolfram Burgard. What matters in language conditioned robotic imitation learning over unstructured data. *IEEE Robotics and Automation Letters*, 7(4):11205–11212, 2022.

[38] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, and Aakanksha et al. Chowdhery. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023.

[39] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Proceedings of the 5th Conference on Robot Learning (CoRL)*, 2021.

[40] Shadowrobot. `https://www.shadowrobot.com/dexterous-hand-series/`, 2005.

[41] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands. *ACM Transactions on Graphics*, 36(6):1–17, 2017.

[42] Jieming Cui, Tengyu Liu, Nian Liu, Yaodong Yang, Yixin Zhu, and Siyuan Huang. Anyskill: Learning open-vocabulary physical skill for interactive agents. In *Conference on Computer Vision and Pattern Recognition(CVPR)*, 2024.

[43] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the International Conference on Computer Vision*, 2021.

[44] Julen Urain, Niklas Funk, Jan Peters, and Georgia Chalvatzaki. Se (3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[45] 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*, 2017.

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[47] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[48] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[49] 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.

[50] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

[51] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems*, 2015.

[52] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2016.

[53] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[54] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[55] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *2018 International Conference on 3D Vision (3DV)*, pages 728–737. IEEE, 2018.

[56] Chenxi Wang, Hao-Shu Fang, Minghao Gou, Hongjie Fang, Jin Gao, and Cewu Lu. Graspness discovery in clutters for fast and accurate grasp detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[57] 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*, 2017.

[58] W. Robotics. Allegro robot hand. `https://www.wonikrobotics.com/research-robot-hand`.

[59] Kenneth Shaw, Ananye Agarwal, and Deepak Pathak. Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning. *arXiv preprint arXiv:2309.06440*, 2023.

[60] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022.

 https://doi.org/10.52202/079017-1487

A Appendix / supplemental material

A.1 DexGYSGrasp Details

A.1.1 Diffusion Background

The diffusion model is used in our intention and diversity grasp component to generate grasp distribution with aligned intention and high diversity, which represents a class of generative models characterized by a forward process of noise addition and a reverse process of denoising. The forward process entails a Markov Chain that incrementally introduces Gaussian noise into the data across multiple time steps. Originating from the initial data x_0, this process transitions the data to conform with a standard Gaussian distribution x_T after T time steps. This transformation is mathematically formulated as follows:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1}, \tag{6}$$

where α_t denotes a time-dependent noise coefficient, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

Therefore, x_t given x_0 follows a normal distribution,

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}). \tag{7}$$

The first equation delineates the stepwise diffusion, whereas the second equation offers a direct approximation of any intermediate state x_t from x_0.

During the reverse process, the model is trained to closely approximate the reverse conditional distribution $p(x_{t-1}|x_t)$, which is described as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta^2(x_t, t)), \tag{8}$$

where $\mu_\theta(x_t, t)$ and $\sigma_\theta^2(x_t, t)$ are the mean and variance parameters for the distribution of x_{t-1}, respectively, and θ indicates the parameters of the model used to predict ϵ from x_t.

The classical sampling strategy for the reverse process is exemplified by DDPM [47], where the model iteratively learns to reverse the noise addition process to reconstruct the original data from noise. It estimates the distribution $p(x_{t-1}|x_t)$ and predicts the noise ϵ, represented by:

$$\boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right) = \tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t, \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t\right)\right)\right) \tag{9}$$

$$= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right)\right), \tag{10}$$

$$x_{t-1} = \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right) + \sigma_\theta(x_t, t)z, \tag{11}$$

where σ_θ consists of non-trainable, time-dependent constants, and z represents Gaussian noise.

A.1.2 Intention and Diversity Grasp Component

Point Encoder We utilize a three-layer PointNet++ [45] as our point encoder, following recent works [56, 14, 2, 7] in the field of robotic grasping. Specifically, each layer $l_i, i \in 1, 2, 3$, receives point clouds and corresponding features (initially the raw XYZ coordinates for the first layer) from the preceding layer. It then performs down-sampling and feature aggregation using the "set-aggregation" operation [45]. The aggregated features are processed by a three-layer perceptron, which consists of three $Linear - BatchNorm - ReLU$ blocks. The output of point encoder is $\mathcal{F}_{obj} \in \mathbb{R}^{N_{obj} \times C_{obj}}$.

Language Encoder For the language encoder, we employ the CLIP model with the ViT-L/14 architecture [46]. The input text sequence is tokenized and converted into token embeddings with positional embeddings added. This sequence is processed through multiple Transformer encoder layers to obtain the language feature $\mathcal{F}_{lan} \in \mathbb{R}^{N_{lan} \times C_{lan}}$.

Transformer Decoder We employ four layers of MLPs and Transformer as decoder[57, 4]. The time embedding and pose feature are incorporated in MLPs to obtain $\mathcal{F}_{dex\_t}$. Subsequently, the $\mathcal{F}_{dex\_t}$ serves as the query, and the concatenated features of language and object, $\mathcal{F}_{lan\_obj}$, act as key and value in Transformer block. The corss attention process is formalized as:

$$\mathcal{F}_{out} = \text{softmax}\left(\frac{f_q(\mathcal{F}_{dex\_t}) f_k(\mathcal{F}_{lan\_obj})^T}{\sqrt{d_k}}\right) f_v(\mathcal{F}_{lan\_obj}), \tag{12}$$

where f_q, f_k, f_v are MLPs, and d_k is the channel of features. Finally, a MLP is adopted to regress the dexterous grasp parameters $\hat{\mathcal{G}}^{dex}$. $\mathcal{F}_{dex\_t} = f_0(f_1(F_{dex}) + f_2(\mathcal{F}_{time}))$, where f_0, f_1, f_2 are MLPs.

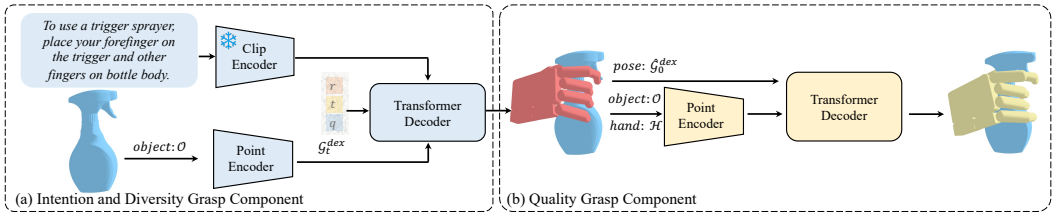


Figure 8: Inference pipeline of our DexGYSGrasp.

A.1.3 Quality Grasp Component

Quality grasp component tasks coarse pose $\hat{\mathcal{G}}^{dex}$, coarse hand point clouds $\mathcal{H}(\hat{\mathcal{G}}^{dex})$ and object point clouds $\mathcal{O}$ as input, and outputs refined grasp $\tilde{\mathcal{G}}^{dex}$. The object and hand are encoded by the PointNet++ same with intention and diversity grasp component. And in the transformer decoder, coarse pose features act as the query, object and hand features serve as key and value.

A.1.4 Inference Pipeline

We also demonstrate the inference pipeline of our DexGYSGrasp, as shown in Fig. 8. We sample a random noise from Gaussian distribution as the input, with the point cloud and language guidance as the conditions. We first generate the coarse grasp by the intention and diversity grasp component, and then refine it with the quality grasp component.

A.1.5 Loss Function

This section provides a detailed exposition of the loss functions utilized during the construction of datasets and the training of models.

Parameter Regression Loss. We utilize the mean squared error (MSE) to quantify the deviation between the generated dexterous hand pose $\hat{\mathcal{G}}^{dex}$ and the ground truth $\mathcal{G}^{dex}$.

$$\mathcal{L}_{para} = \frac{1}{N}\sum_{i=1}^{N} \|\mathcal{G}^{dex,i} - \hat{\mathcal{G}^{dex,i}}\|_2^2. \tag{13}$$

Hand Chamfer Loss. The predicted hand point clouds $\mathcal{H}(\hat{\mathcal{G}}^{dex})$ and the ground truth $\mathcal{H}(\mathcal{G}^{dex})$ are derived by sampling from the hand mesh. We then compute the chamfer distance to assess the discrepancies between the predicted and ground-truth hand shapes.

$$\mathcal{L}_{chamfer} = \sum_{x\in\mathcal{H}(\mathcal{G}^{dex})} \min_{y\in\mathcal{H}(\hat{\mathcal{G}}^{dex})} \|x-y\|_2^2 + \sum_{x\in\mathcal{H}(\hat{\mathcal{G}}^{dex})} \min_{y\in\mathcal{H}(\mathcal{G}^{dex})} \|x-y\|_2^2. \tag{14}$$

Contact Map Loss. The contact map loss $\mathcal{L}_{cmap}$ ensures consistency between the predicted hand contact map $\hat{c}^{obj}$ on object and the target c^{obj}. The contact map is calculate by the distance from object point to the closest dexterous hand point.

$$\mathcal{L}_{cmap} = \sum_i \|c_i^{obj} - \hat{c}_i^{obj}\|_2^2. \tag{15}$$

Object Penetration Loss. The object penetration loss $\mathcal{L}_{pen}$ penalizes the depth of hand-object penetration, where d_i^{sdf} denotes the signed distance from the object point to the hand mesh.

$$\mathcal{L}_{pen} = \sum_i \mathbb{I}(d_i^{sdf} > 0)\cdot d_i^{sdf}. \tag{16}$$

Self-Penetration Loss. The self-penetration loss $\mathcal{L}_{spen}$ punishes the penetration among the different parts of the hand, where $p^{dex,sp}$ denotes predefined anchor spheres on the hand [14].

$$\mathcal{L}_{spen} = \sum_{i,j} \mathbb{I}(i\neq j)\cdot \max(\delta - d(p_i^{dex,sp}, p_j^{dex,sp})). \tag{17}$$

https://doi.org/10.52202/079017-1487

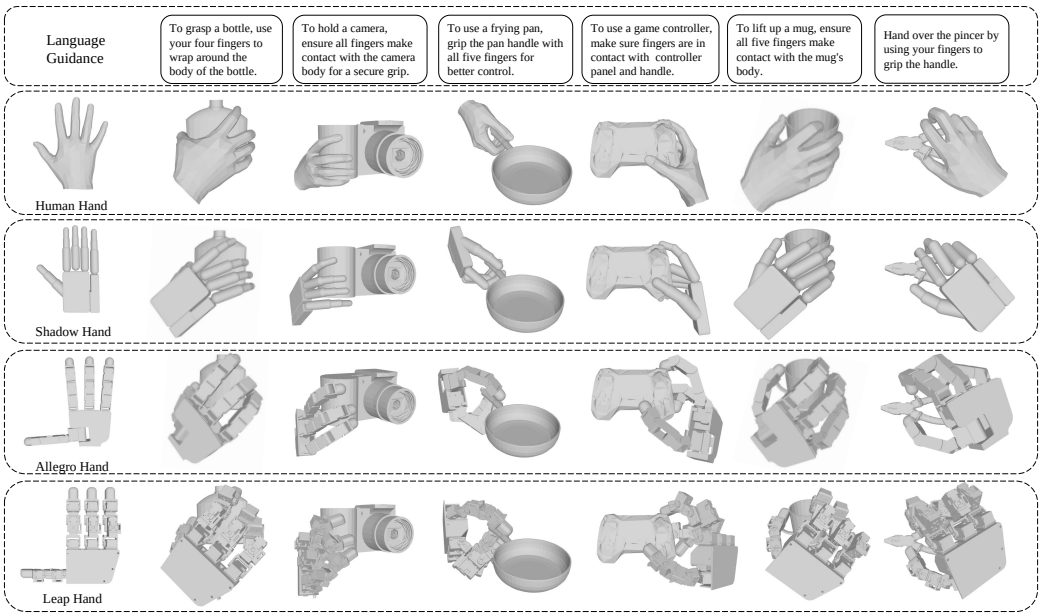


Figure 9: The Extension of DexGYSNet to more dexterous hands.

Joint Angle Loss. Given the physical structure limitations of the robotic hand, each joint has designated upper and lower limits. The joint angle loss penalizes deviations from these limits.

$$\mathcal{L}_{joint} = \sum_{i} (\max(q_i - q_i^{max}) + \max(q_i^{min} - q_i)). \tag{18}$$

A.2 DexGYSNet Datasets Details

A.2.1 Prompt of LLM

We introduce the prompt for using GPT-3.5 in this section.

System Prompt: "You are an assistant in creating language instruction, aimed at guiding robot on how to grasp objects. Given a brief instruction and a fine-gained interaction information. Your task is generate a natural and more informative instruction. The instruction should start with the given brief instruction, which is limited in a sentence and about 10-15 words."

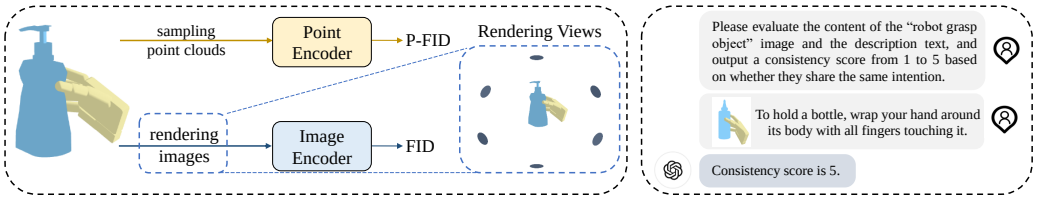
User Prompt: "Brief instruction: To <*brief intention*> a <*object category*>. Hand-object interaction information: <*contact information*>. "

The *brief intention* and *object category* are sourced from the hand-object dataset OakInk [27]. The *contact information* is derived by calculating the distances from predefined contact anchors on each finger to the segmentation parts of the object. Details on predefined contact anchors are available in DexGraspNet [14], and segmentations are annotated in OakInk [27].

An example of user prompt is: "Brief instruction: To use a trigger sprayer. Hand-object interaction information: forefinger touches the trigger. thumb, middle finger, ring finger and little finger touches the finger." An example of LLM output is: "To use a trigger sprayer, press the trigger with your forefinger and hold the bottle with your other fingers."

A.2.2 DexGYSNet Extension

Our cost-effective dataset construction strategy can be easily extended to various types of dexterous hands. As shown in Figure 9, besides the Shadow Hand[40], which features a highly biomimetic design replicating most degrees of freedom of human hands at a high cost of $100,000, we also expand to the Allegro Hand [58] and Leap Hand [59]. These latter models, while offering fewer



(a) The evaluation pipeline of P-FID and FID.

(b) LLM-assisted evaluation.

Figure 10: (a) Evaluation of intention consistency using the Fréchet Inception Distance between the <generation hand and object> and the ground truth. (b) When the ground truth is not available (e.g., evaluation on a 3D object dataset), we employ GPT4-o for evaluation.

degrees of freedom, are significantly more affordable, costing $16,000 and $2,000, respectively, making them practical for promoting the use of robotic arms in laboratory environments. We have trained our method on the DexGYSNet dataset using the Allegro Hand and implemented it in real robot experiments.

A.3 Implementation Details

A.3.1 Dataset Split

We split the DexDYS dataset at the level of object instances. Specially, for all objects within each category, 80% of the objects instances are used for training and 20% for evaluation. Concretely, the training set includes approximately 1,200 objects with 40k grasps, while the evaluation set comprises about 300 objects with 10k grasps. Therefore, all objects in the test set of DexGYSNet don't exist in the training set.

A.3.2 Metrics Detials

Target Assignment. For target assignment in the testing phase, the grasp targets of an object-guidance pair consist of all poses that share the same contact part and brief guidance. And the matrices of intention consistency are calculated by comparing the prediction to the most similar grasp target.

Fréchet Inception Distance, which is commonly used in generative task[60] by measuring the distance between the generated distribution and the ground truth distribution. We use sampling point cloud features extracted from[49] to calculate P-FID and rendering image features extracted from[50] to calculate FID. The details are shown in Figure 10 (a).

Chamfer distance, denoted as CD, is used to measure the distance between predicted hand point clouds and targets to measure the consistency from the aspect of hand consistency. Please look at Equation 14 for details.

Contact distance, denoted as $Con.$ to measure the L2 distance of object contact map between the prediction and targets to measure the consistency from the aspect of object contact consistency. Please look at Equation 15 for details.

Success rate. We evaluate the grasp success rate in Issac Gym simulation environment. To simulate the force exerted by dexterous hands grasping objects in real environments, we contract each finger in the direction of the object. If the grasp can withstand at least one of the six directions of gravity, it is considered successful.

Mean Q_1. Intuitively, the Q_1 metric reflects the norm of the smallest wrench which can disrupt the stability of a grasp. We follow [14] to set the contact threshold to 1cm and set the penetration threshold to 5mm. Any grasp with its maximal penetration depth greater than 5mm is considered invalid and we set the Q_1 of it to 0 before taking the average.

Maximal penetration depth, which is the maximal penetration depth from the object point cloud to hand meshes.

Diversity. We use the standard deviation of translation δ_t, rotation δ_r, and joint angle δ_q to measure the diversity of generated grasps. We perform eight samples in the intention and diversity component under the same input conditions, and each sample is individually sent to the quality component

https://doi.org/10.52202/079017-1487

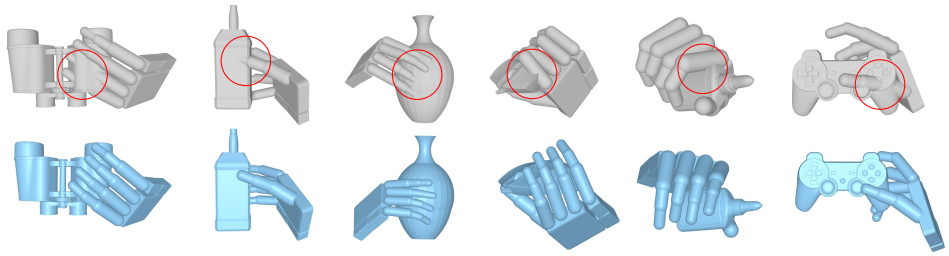


Figure 11: Visualization of grasps before and after quality grasp component. Our quality grasp component improves grasp quality and maintains intention consistency.

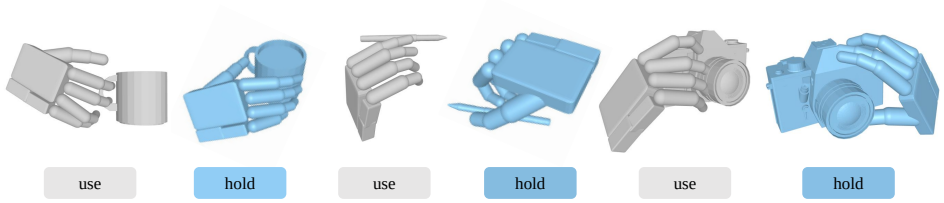


Figure 12: Visualization of our DexGYSGrasp framework with task-oriented simple input.

for refinement. Before calculation, δ_r and δ_q are converted to Euler angles in degrees, while δ_t is measured in centimeters.

A.3.3 Implementation Details of SOTA Methods

We replicate SOTA methods on our DexGYSNet dataset using the same encoder structure and the loss functions defined in Equation 5 to ensure fair comparison. Specifically, we reimplement GraspCAVE based on [51], GraspTTA [43], SceneDiffuser [4], and DGTR [7]. To introduce language information, we use an identical CLIP language encoder. For GraspCAVE, we concatenate the language feature, object feature, and latent feature to send to the decoder. Based on GraspCAVE, GraspTTA employs a test-time adaptation strategy for quality refinement. For SceneDiffuser, we concatenate the language and object features as the model condition. For DGTR, the language and object features are concatenated to send to its transformer decoder.

A.4 Additional Experiments

A.4.1 Qualitative Experiments of Quality Grasp Component.

We provide additional qualitative results to verify the effectiveness of Quality Grasp Component. Figure 11 shows the grasps before and after the application of the Quality Grasp Component, demonstrating that QGC can prevent object penetration and maintain consistency with the original intention.

A.4.2 Qualitative Experiments of Task-oriented Guidance.

We conduct qualitative experiments to demonstrate the generalization of our DexGYSGrasp framework to task-oriented or functional grasp task. Specifically, we input task-oriented guidance (e.g., "use" or "hold") into our framework, which has been trained on DexGYSNet. As shown in Figure 12, our DexGYSGrasp framework exhibits good compatibility with these inputs. This further confirms that our approach enables more flexible and natural human-robot interactions.

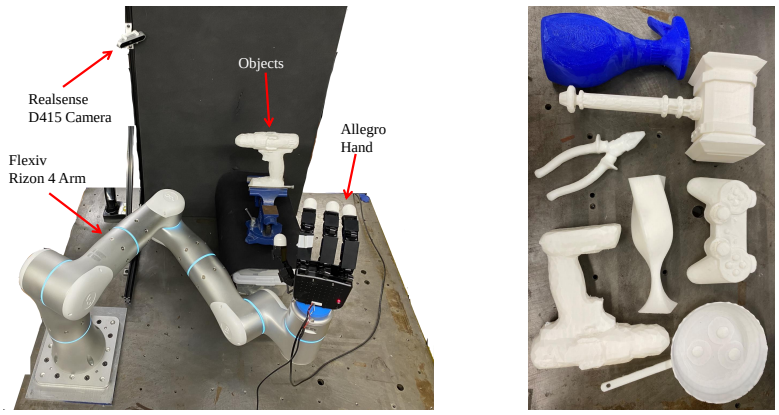


Figure 13: The illustration of our real world experiments settings.

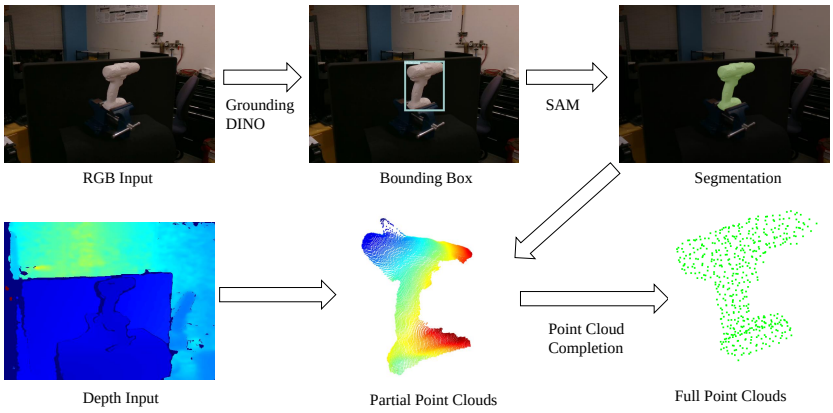


Figure 14: Real world experiments pipeline.

A.5 Real World Experiments Details

Experimental Environment Figure 13 shows the settings of our real world experiments. The experiments are conducted on Allegro hand, a Flexiv Rizon4 arm and an Intel Realsense D415 camera. The experimental object is a 3D printed object from test set of DexGYSNet.

Experiment Pipeline Our DexGYSGrasp takes full point clouds as input following recent works in dexterous grasping [14, 4, 7]. To make our methods more practical, we employ three off-the-shelf models in a cascade to obtain a full point cloud from scene point cloud. As shown in Figure 14, we input the object category and the RGB image into an open-set detection model [53] to detect the bounding box of the object. This bounding box is then used as a prompt for SAM [54] to obtain the segmentation of the object. Next, we crop the target depth image using the segmentation map and the depth input. Finally, we convert the partial depth image into point clouds and feed it into a point completion network [55] to obtain the final full point clouds. Then, the full point clouds are fed into our framework to obtain the dexterous grasp pose, which is then transformed into the real coordinate system. In execution, we first move the arm to the 6-DOF pose of the dexterous hand root node, and then control the joint angles to achieve the target pose.

Experiment Results The experiment results are presented in Table 5. For each object, we command robot with different language instruction, and each instruction is tested five times, resulting in a total of ten grasping trials per object. A grasp is deemed successful if it aligns with the intended instruction and maintains stability, preventing the object from falling. Our method demonstrates a moderate success rate, indicating its effectiveness. Further research on real-world scenarios is recommended to enhance the robustness of our approach.

https://doi.org/10.52202/079017-1487

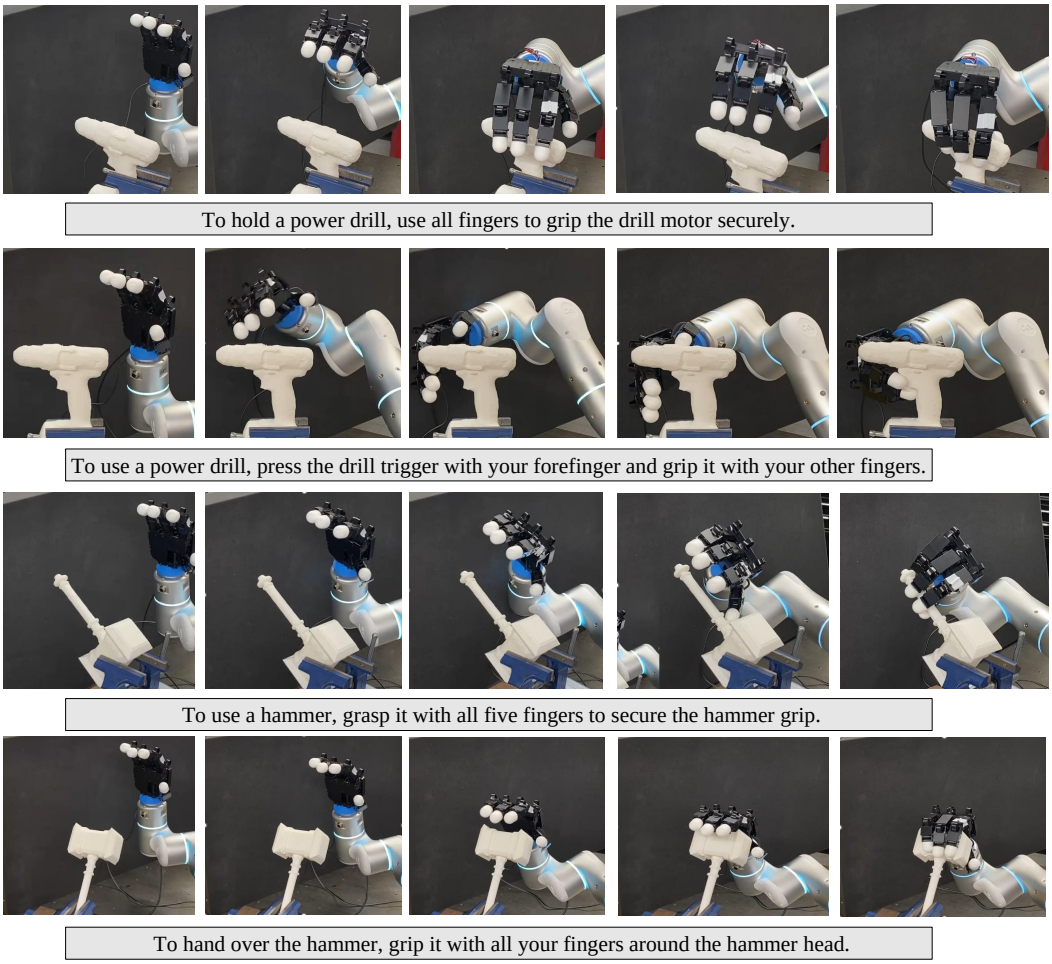


Figure 15: Real world experiments.

	Power drill	Hammer	Trigger sprayer	Game controller	Pincer	Frying pan	Wineglass
Success	9/10	4/10	3/10	8/10	5/10	3/10	6/10

Table 5: Real word experiment results

A.6 Societal Impacts and Limitations

The core innovation of this paper has a significant positive impact on society. We propose a novel task: language-guided dexterous grasp generation, which can promote human-robot interaction and expedite the deployment of robots in real-world scenarios. Additionally, we introduce an innovative framework to accomplish this task. Our approach can generate high-quality grasps while ensuring consistency of intent and diversity of grasps.

However, our method still faces some challenges in real-world deployment. Due to limitations in the current development of robotic arm control and physical structures, we cannot guarantee success in every grasp execution in the real world. In future work, we will further enhance the quality of grasp generation to improve the success rate in real-world scenarios.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work in the conclusion and supplementary material.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

https://doi.org/10.52202/079017-1487

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully disclose all the information needed to reproduce the main experimental results of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer:[No]

   Justification: We provide examples of our dataset in supplementary materials. We promise to release all code and the complete dataset after the publication of this paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all the training and test details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported because performing multiple runs for each experiment would be too computationally expensive.

   Guidelines:

https://doi.org/10.52202/079017-1487

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides sufficient information on the computer resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The paper discusses both potential positive societal impacts and negative societal impacts of the work in the conclusion and supplementary material.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The original owners of assets, used in the paper, are properly credited and the license and terms of use are explicitly mentioned and properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

https://doi.org/10.52202/079017-1487

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets introduced in the paper are well documented and the documentation is provided alongside the assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-1487