Grasp as You Say: Language-guided Dexterous Grasp Generation

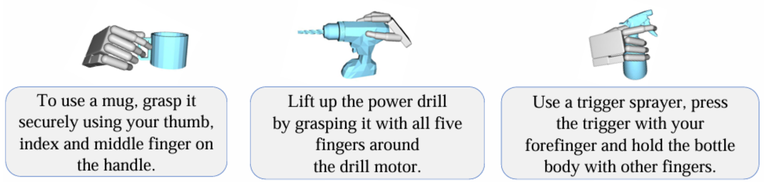
Yi-Lin Wei[1], Jian-Jian Jiang[1], Chengyi Xing[2], Xian-Tuo Tan[1], Xiao-Ming Wu[1], Hao Li[2], Mark Cutkosky[2], Wei-Shi Zheng[1†]

[1] Sun Yat-sen University [2] Stanford University

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

This paper explores a novel task "Dexterous Grasp as You Say" (DexGYS), enabling robots to perform dexterous grasping based on human natural language commands.



First, 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.

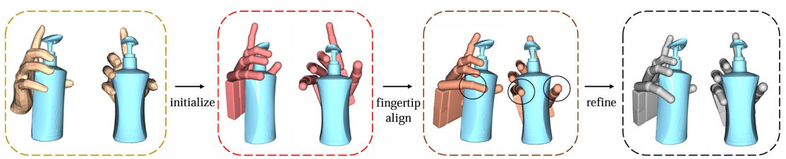
Second, we introduce the **DexGYSGrasp** framework for generating dexterous grasps based on human language instructions, with the capability of producing grasps that are intent-aligned, high quality and diversity.

DexGYSNet Dataset

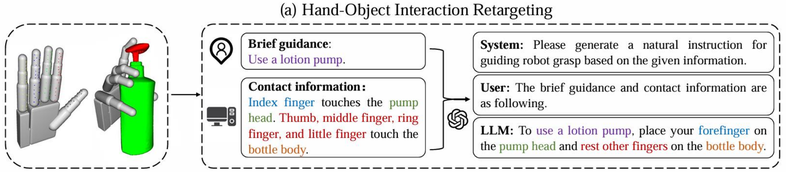
The construction process of the DexGYSNet dataset.

(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.



(a) Hand-Object Interaction Retargeting

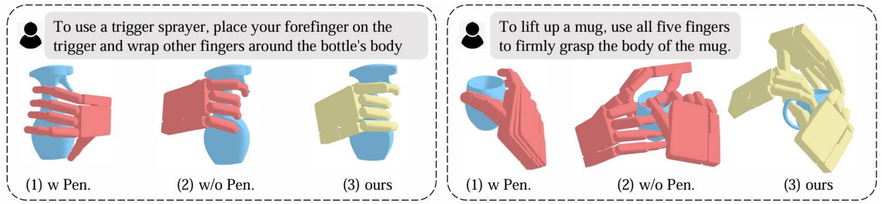


(b) LLM-assisted Language Guidance Annotation

DexGYSGrasp framework

Problem Formulation. Given object point cloud $\mathcal{O} \in \mathbb{R}^{M*3}$ and the language guidance, the framework generates the dexterous grasp $g = (r, t, q)$, where $\mathrm{r} \in SO(3)$ and $\mathrm{t} \in \mathbb{R}^3$ are the global rotation and translation in the world coordinate, and $\mathrm{q} \in \mathbb{R}^J$ is the joint angles of the J-DoF dexterous hand.

Learning Challenge in DexGYS. We find that a single model struggles to generate grasp with intention alignment, high diversity and high quality simultaneously, due to the optimization challenge caused by the commonly used object penetration loss.

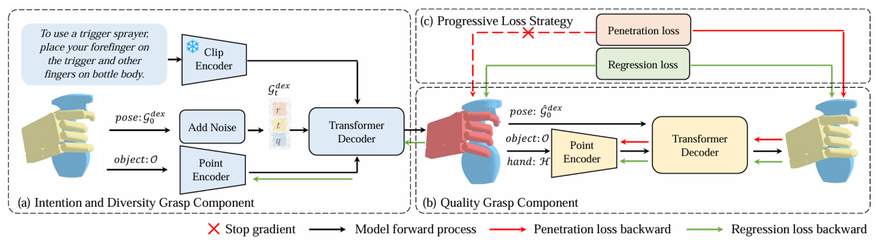


(a) Comparison for Intention Consistency and Quality (b) Comparison for Grasp Diversity and Quality.

Progressive Grasp Components. We propose a diffusion-based generative model for language-object conditioned coarse grasp generation and a transformer-based discriminative model for grasp refinement under different supervision.

(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 istrained to refine the coarse pose improve the grasp quality while maintain intension consistency.



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

Contact: weiylin5@mail2.sysu.edu.cn

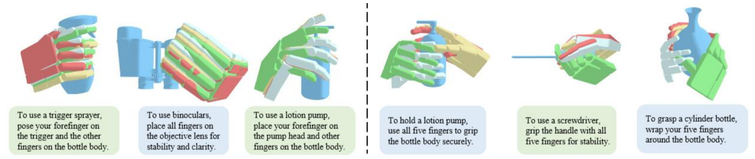
My Homepage: https://wyl2077.github.io/

Experiments

Results on DexGYSNet. The overall results confirm that our framework achieves SOTA performance in generating intention-aligned, high-quality and diverse grasps.

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**

Visualization of generated dexterous grasp. The left shows that the generated grasp are consistent with clear and specific guidance, while the right shows that the diversity achieved under relatively ambiguous instructions.



Visualization of real world experiments. The real robot experiments shows that our methos can well generalize and deploy to real world.