

Language-driven Grasp Detection

An Vuong¹ Minh Nhat Vu^{2, 3} Baoru Huang⁴ Nghia Nguyen¹
Hieu Le¹ Thieu Vo⁵ Anh Nguyen⁶

¹FPT Software AI Center ²ACIN - TU Wien ³Austrian Institute of Technology

⁴Imperial College London ⁵Ton Duc Thang University ⁶University of Liverpool

March 27, 2024



Contents

- 1 Introduction
- 2 Methodology
- 3 Results

Introduction

Introduction



Figure: We present Grasp-Anything++, a new *language-driven grasp dataset* for executing grasps through linguistic commands featuring 1M samples, over 3M objects, and upwards of 10M grasping instructions.

Key Contributions

Our contributions are three-fold:

- We propose Grasp-Anything++, a large-scale language-driven dataset for grasp detection tasks.
- We propose a diffusion model with a training objective that explicitly contributes to the denoising process to detect the grasp poses.
- We demonstrate that our Grasp-Anything++ dataset and the proposed method outperform other approaches and enable successful robotic applications.





Methodology

Dataset Generation - Overview

We utilize large-scale foundation models to create the Grasp-Anything++ dataset. There are three key steps in establishing our dataset:

- 1 Prompting ChatGPT for a corpus of scene descriptions.
- 2 Synthesizing images from descriptions and annotating grasp poses.
- 3 Evaluating grasp poses and reducing biases.

Dataset Generation - Details

Step	Description		
Scene Generation	User	Please help me generate scene descriptions for natural arrangements of daily objects. Each description has the following form: <code><Object.1><Object.2>...<Verb><Container.Object></code> . Please also ensure the incorporation of a rich and varied lexicon in the scene descriptions.	
	Sample	A steel knife, a polished fork and a pristine ceramic plate on a wooden table.	
	Text-to-Image	We use Stable Diffusion [57] to proceed text-to-image generation.	
Object Masking	User	For object part-level description, given an input list <code>{<Object.1>, <Object.2>, ...}</code> , the output will be a list that describes the parts of objects as: <code>{<Object.1>: [<Part.1.1>, <Part.1.2>, ...], <Object.2>: [<Part.2.1>, <Part.2.2>, ...]}</code> .	
	Sample	<code>{knife: [handle, blade], fork: [handle, neck, stem, tines], plate: [rim, base]}</code>	
	Post Process	We use OFA [74] and SAM [34] to locate the region describing the objects.	
Part Masking	User	Given the object list and part lists of each scene description, you will generate for me all prompts with the following format: <code>{<Manipulation.Action><Object.ID><Part.ID>}</code> . The part that is more suitable for human grasping is positioned at the start of the list to represent the grasping actions.	
	Sample	Give me the steel knife; Grasp the knife at its handle.	
	Post Process	We leverage VLPART [65] to locate the region describing the parts of objects.	
Grasp Generation	User	Generate for me a scene description with grasp instructions following the templates.	
	Sample	Scene description: A steel knife, a polished fork and a pristine ceramic plate on a wooden table. Object list: <code>{knife, fork, plate}</code> . Part lists: <code>{knife: [handle, blade], fork: [handle, neck, stem, tines], plate: [rim, base]}</code> . Prompts: Give me the steel knife; Grasp the knife at its handle.	
	Grasp Labelling	We utilize a pretrained RAGT-3/3 [8] to generate grasp poses corresponding to the located region.	

Method Overview

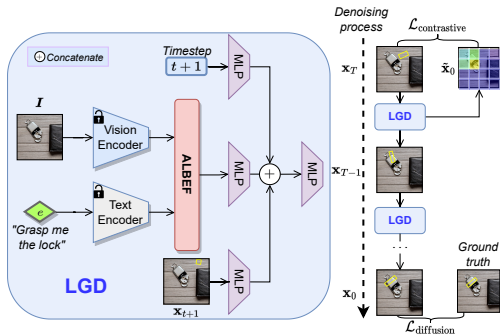


Figure: Network architecture.

We leverage a denoising diffusion probabilistic model [2] to generate grasp poses. Through ALBEF module [4], we integrate embeddings from both text and image. The attention mask produced by ALBEF serves as our estimate, and we apply a contrastive loss to this mask against x_T . We generate the denoised grasp pose by passing ALBEF's features through MLP layers.

Theoretical Findings

Proposition

Suppose that \tilde{x}_0 , x_0 and ϵ are independent, and that

$$\left\| \frac{\sqrt{\bar{\alpha}_T} \tilde{x}_0 - x_T}{\sqrt{1 - \bar{\alpha}_T}} \right\|_2^2 \geq M$$

Then there exists $C > 0$ such that: for arbitrary $\delta > 0$, if $\mathcal{L}_{\text{contrastive}} < \delta$, then

$$\mathbb{E} [\|\tilde{x}_0 - x_0\|_2^2] < C\delta$$

The proposition suggests that if the contrastive loss $\mathcal{L}_{\text{contrastive}}$ tends to zero, then the prediction \tilde{x}_0 will approach the ground truth x_0 .

Results

Language-driven Grasp Detection

We compare our language-driven grasp detection method (LGD) with the linguistically supported versions of GR-CNN [3], Det-Seg-Refine [1], GG-CNN [5], CLIPORT [7] and CLIP-Fusion [8]. In all cases, we employ a pretrained CLIP or BERT as the text embedding.

	Baseline	Seen	Unseen	H
GR-ConvNet [3] + CLIP [6]		0.37	0.18	0.24
Det-Seg-Refine [1] + CLIP [6]		0.30	0.15	0.20
GG-CNN [5] + CLIP [6]		0.12	0.08	0.10
CLIPORT [7]		0.36	0.26	0.29
CLIP-Fusion [8]		0.40	0.29	0.33
LGD (ours) + BERT		0.44	0.38	0.41
LGD (ours) + CLIP		0.48	0.42	0.45

Table: Language-driven grasp detection results.

Qualitative Visualizations

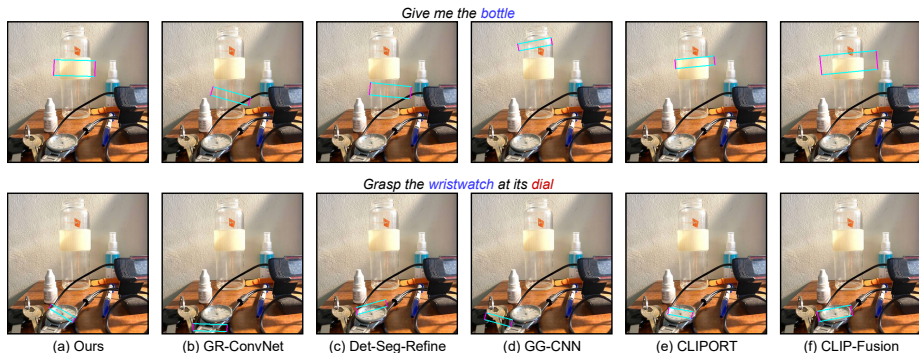


Figure: Language-driven grasp detection results visualization.

We provide qualitative results of the language-driven grasp detection task in Fig. 3. The outcomes suggest that our method LGD generates more semantically plausible than other baselines.

Zero-shot Grasp Detection

This experiment evaluates the performance of Grasp-Anything++ and LGD in traditional grasp detection tasks against existing datasets and methods.

Baseline	Grasp-Anything++ (ours)			Jacquard			Cornell			VMRD			OCID-grasp		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H	Base	New	H
GR-ConvNet [3]	0.71	0.59	0.64	0.88	0.66	0.75	0.98	0.74	0.84	0.77	0.64	0.70	0.86	0.67	0.75
Det-Seg-Refine [1]	0.62	0.57	0.59	0.86	0.60	0.71	0.99	0.76	0.86	0.75	0.60	0.66	0.80	0.62	0.70
GG-CNN [5]	0.68	0.57	0.62	0.78	0.56	0.65	0.96	0.75	0.84	0.69	0.53	0.59	0.71	0.63	0.67
LGD (no text) (ours)	0.74	0.63	0.68	0.89	0.69	0.77	0.97	0.76	0.85	0.79	0.66	0.72	0.88	0.68	0.76

Table: Base-to-new zero-shot grasp detection results.

The results reveal that Grasp-Anything++ poses a greater challenge due to its extensive inclusion of unseen objects during testing, as evidenced by lower detection results compared to similar approaches on related datasets.

Robotic Evaluation

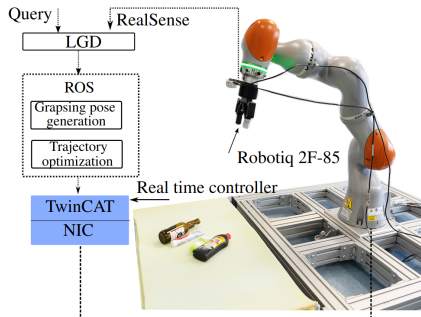


Figure: Robotic experiment setup.

	Baseline	Single	Cluttered
GR-ConvNet [3] + CLIP [6]		0.33	0.30
Det-Seg-Refine [1] + CLIP [6]		0.30	0.23
GG-CNN [5] + CLIP [6]		0.10	0.07
CLIPORT [7]		0.27	0.30
CLIP-Fusion [8]		0.40	0.40
LGD (ours)		0.43	0.42

Table: Robotic language-driven grasp detection results.

Reference

- [1] Stefan Ainetter and Friedrich Fraundorfer. End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from rgb. In *ICRA*, 2021.
- [2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [3] Sulabh Kumra, Shirin Joshi, and Ferat Sahin. Antipodal robotic grasping using generative residual convolutional neural network. In *IROS*, 2020.
- [4] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 2021.
- [5] Douglas Morrison, Peter Corke, and Jürgen Leitner. Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach. *arXiv preprint arXiv:1804.05172*, 2018.
- [6] 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 *ICML*, 2021.
- [7] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *CoRL*, 2022.
- [8] 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. *arXiv preprint arXiv:2302.12610*, 2023.

Thank you!