2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Grasp-Anything: Large-scale Grasp Dataset from Foundation Models

A.D. Vuong[1], M.N. Vu*[,2,3], H. Le[1], B. Huang[4], H.T.T. Binh[5], T. Vo[6], A. Kugi[2,3], A. Nguyen[7]

***Abstract*— Foundation models such as ChatGPT have made significant strides in robotic tasks due to their universal representation of real-world domains. In this paper, we leverage foundation models to tackle grasp detection, a persistent challenge in robotics with broad industrial applications. Despite numerous grasp datasets, their object diversity remains limited compared to real-world figures. Fortunately, foundation models possess an extensive repository of real-world knowledge, including objects we encounter in our daily lives. As a consequence, a promising solution to the limited representation in previous grasp datasets is to harness the universal knowledge embedded in these foundation models. We present Grasp-Anything, a new large-scale grasp dataset synthesized from foundation models to implement this solution. Grasp-Anything excels in diversity and magnitude, boasting 1M samples with text descriptions and more than 3M objects, surpassing prior datasets. Empirically, we show that Grasp-Anything successfully facilitates zero-shot grasp detection on vision-based tasks and real-world robotic experiments. Our dataset and code are available at https://airvlab.github.io/grasp-anything/.**

I. INTRODUCTION

Grasp detection is a fundamental and long-standing research topic in robotics [1]. Establishing principles and techniques for grasp detection has enabled multiple applications such as manufacturing, logistics, and warehouse automation [2]. Recent advances in deep learning have introduced effective avenues for the development of data-driven systems for robotic grasping [3]. Numerous deep learning approaches have been proposed to address robotic grasping [4]–[7]; however, they primarily focus on improving the neural network (i.e., *model-centric* approach). While often excelling in the training progress, these model-centric strategies demonstrate unstable outcomes with different datasets, especially with real-world data [8]. Furthermore, Platt *et al.* [9] contend that the outcomes conducted on physical robots strongly depend on the training data, underscoring the pivotal role of grasp data. Consequently, our paper explores the *data-centric* approach, which aims to improve the quality of grasp data to achieve more robust generalization in grasp detection.

Over the years, many grasp datasets have been proposed (Table I). Nevertheless, current grasping datasets share common limitations. First, grasp datasets have been constrained by the limited number of objects [10]. The restriction of the number of objects leads to dissimilarities among existing datasets and may introduce inconsistencies when transferring to real-life robotic applications [8], [11]. Second, most of the current grasp datasets do not consider natural language descriptions for each scene arrangement, limiting human-robot interactions [12]. Finally, previous works make assumptions about scene arrangements resembling bin-like configurations [13] or lab-controlled environments [3], which diverges significantly from the complexity of natural settings [8]. To overcome the challenge, we aim to establish a new large-scale *language-driven grasp dataset* that ideally covers unlimited scene arrangements in our daily lives.

Fig. 1. We introduce Grasp-Anything, a new large-scale language-driven grasp dataset that universally covers objects in our daily lives by using knowledge from foundation models.

Recently, we have witnessed the applications of utilizing foundation models, such as large language model (LLM) [29] or text-to-image (T2I) model [30], across various domains of robotics research [31]. Foundation models have demonstrated remarkable encouragements in various tasks [32], for example, task and motion planning [33], manipulation [34], visual-and-language navigation [35], and scene understanding [36]. The utilization of large-scale pretrained foundation models has facilitated the integration of omniscient knowledge into robotic systems [37], overcoming the challenges faced by traditional methods in robustly modeling unstructured and novel environments [38]. Inspired by these phenomena, we hypothesize that it is possible to apply panoptic knowledge from large foundation models to ideally generate an unlimited number of objects and, therefore, serve as a foundation to synthesize a grasp dataset that universally covers possible objects and arrangements that come into existence.

We introduce **Grasp-Anything**, a new large-scale dataset for grasp detection. Unlike existing grasp datasets [1], [3], [5], [6], [11], [14], [15], [20]–[23], [25]–[27] that are limited to a predefined set of objects and arrangements, our dataset offers an extensive range of objects and closely replicates real-world scenarios in natural environments, thus alleviating

[1] FPT Software AI Center, Vietnam anvd2@fpt.com
[2] Automation & Control Institute, TU Wien, Vienna, Austria
[3] AIT Austrian Institute of Technology, Vienna, Austria
[4] Imperial College London, UK
[5] Hanoi University of Science and Technology, Vietnam
[6] Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[7] Department of Computer Science, University of Liverpool, UK
* Corresponding author minh.vu@ait.ac.at

TABLE I

SUMMARY OF THE PROPERTIES OF PUBLICLY AVAILABLE GRASP DATASETS.

Dataset	Grasp Label	Data Representation	Multi-object?	Grasps per object	**Num. objects**	Num. grasps	**Num. samples**	Data type	Label annotation	Scene description?
Cornell [14]	Rec.	Image	✗	33	240	8019	1035	Real	Human	✗
Pinto *et al.* [1]	Rec.	Image	✓	1	150	50K	50K	Real	Trial-and-error	✗
Levine *et al.* [3]	Rec.	Image	✓	-	-	800K	-	Real	Trial-and-error	✗
Dex-net [15]	Rec.	Image	✗	100	1500	6.7M	-	Sim	Analysis	✗
Jacquard [11]	Rec.	Image	✗	100	11K	1.1M	54K	Sim	Sim. (PyBullet [16])	✗
VMRD [17]	Rec.	Image	✓	20	15K	100K	4683	Real	Human	✗
OCID-grasp [18]	Rec.	Image	✓	6-7	∼40	75K	11K	Real	Human	✗
VR-Grasping-101 [19]	6-DoF	Image	✗	10-20	101	151K	-	Sim	Trial-and-error	✗
YCB-Video [20]	None	Image	✓	None	21	None	134K	Real	None	✗
GraspNet [21]	6-DoF	Image	✓	3-9M	88	1.2B	97K	Real	Analysis	✗
Kappler *et al* [22]	6-DoF	Point Cloud	✗	500	80	300K	700	Sim	Analysis	✗
6-DOF GraspNet [23]	6-DoF	Point Cloud	✗	34K	206	7M	206	Sim	Sim. (FleX [24])	✗
Epper *et al.* [25]	6-DoF	Point Cloud	✗	47.8M	21	1B	21	Sim	Sim. (FleX [24])	✗
EGAD! [26]	6-DoF	Point Cloud	✗	100	2231	233K	2231	Sim	Analysis	✗
ACRONYM [27]	6-DoF	Point Cloud	✓	2000	8872	17.7M	-	Sim	Sim. (FleX [24])	✗
MetaGraspNet [13]	6-DoF	Point Cloud	✓	5K	82	-	217K	Sim	Sim. (Isaac Sim [28])	✗
Grasp-Anything (Ours)	Rec.	Image	✓	∼200	∼3M	∼600M	1M	Synth.	Analysis	✓

the generalization issues [39]. We empirically demonstrate that Grasp-Anything facilitates zero-shot learning in both computer vision and robotic aspects of grasp detection. In the vision aspect, our findings confirm that the performances of grasp detection baselines trained by Grasp-Anything improve significantly over related datasets. In the robotic aspect, we demonstrate that our large-scale dataset can be directly applied to real robot systems to improve the grasping task. In summary, our contributions are as follows:

- We leverage knowledge from foundation models to introduce Grasp-Anything, a new large-scale dataset with 1M (one million) samples and 3M objects, substantially surpassing prior datasets in diversity and magnitude.
- We benchmark zero-shot grasp detection on various settings, including real-world robot experiments. The results indicate that Grasp-Anything effectively supports zero-shot grasp detection in light of its comprehensive representation of real-world scene arrangements.

II. RELATED WORK

Grasp Datasets. Several grasp datasets have been introduced recently [40] (see Table I). Many factors can be considered when designing a grasp dataset, such as data representation (RGB-D or 3D point clouds), grasp labels (rectangle-based or 6-DoF), and quantity [10]. Notably, a key distinction between our Grasp-Anything dataset and its counterparts lies in its *universality*. While a limited selection of objects constrains existing benchmarks, our dataset is designed to encompass a diverse spectrum of objects observed in our natural lives. In addition, our dataset incorporates natural settings for object arrangements, thereby distinguishing itself from prior works where object configurations are more strictly controlled [9]. Grasp-Anything outperforms other benchmarks in both the number of objects and the number of samples.

Grasp Detection. Deep learning techniques have demonstrated notable achievements in grasp detection. Lenz *et al.* [41] present one of the first works utilizing deep learning to detect grasp pose. Subsequently, learning-based approaches [4], [6], [18], [19], [42]–[44] have gained prominence as the most widely utilized solution for grasp detection. Despite the extensive research on deep learning methods for robotic grasp, it is still challenging to apply to real-world grasping applications [9]. This is primarily attributed to the limited size and diversity of existing datasets for robotic grasp [13]. On the other hand, zero-shot grasp detection is a promising approach for utilizing large-scale datasets. Liu *et al.* [45] consider the zero-shot grasp detection problem. However, they mainly focus on sim2real applications rather than benchmarking the zero-shot feasibility of existing datasets. Therefore, we are motivated to develop a large-scale benchmark for grasp detection, enabling the successful grasp of universal objects we encounter in real-life scenarios.

Foundation Models for Robotic Applications. Different attempts to incorporate foundation models into robotic applications have been proposed [35], [46]. For instance, Kapelyukh *et al.* [34] introduce a framework based on DALL-E [47] to solve practical rearrangement tasks. Pretrained image in-painting models have also been utilized in [48] for data augmentation. Although applying foundation models to robotic systems has become an inevitable trend [32], it still bears several uncertainties [37]. First, these models need to be improved in functional competence [37], impeding their ability to solve novel planning problems that require an understanding of how the world operates [49]. In addition, prior approaches often exhibit drawbacks in terms of scope, limited functionalities, or an open-loop nature that does not allow for fluid interactions and corrections based on user feedback [32]. Recognizing these limitations, we design a data-centric approach based on LLM and T2I models, focusing on generating large-scale data for robotic grasping within this paper.

III. THE GRASP-ANYTHING DATASET

Fig. 2 shows the overview of the procedure to generate our Grasp-Anything dataset. We first perform prompt engineering to generate scene descriptions and utilize foundation models to generate images from these text prompts. The grasp poses are then automatically generated and evaluated. We represent grasp poses as 2D rectangles [14] as in many

previous works due to the simplicity and compatibility with real-world parallel plate grippers [11].

A. Scene Generation

Prompt Engineering. To tackle the challenge of generating a universal set of objects, we utilize ChatGPT [29] and perform a prompt engineering technique [50] to guide ChatGPT to generate diverse scene descriptions.

i) Directives Initialization. The concept of 'directives' refers to the configuration of the goal for ChatGPT to match the user preference [35]. In our context, the goal for ChatGPT is to generate a diverse set of scene descriptions that cover a large proportion of objects. Therefore, we initiate directives for ChatGPT using the following conversation

> **Q:** "Imagine you are helping me to generate a corpus of scene descriptions, each condensed to a single sentence. My goal is to generate as many diverse graspable objects as possible. Each sentence must be distinct and should contain at least two objects."
> **A:** "Sure, I'd be happy to help you generate a large corpus of scene descriptions with diverse objects."

We then set up the output template for ChatGPT as follows

> **Q:** "The template for each sentence contains two parts. The first part is the sentence with the structure as `<Obj_1><Obj_2>...<Verb><Container>`. The second part is the list of extracted objects from the sentence `[<Obj_1><Obj_2>...]`."
> **A:** "Understood! Let's generate the first sentence..."

This template guarantees that each generated prompt has two components: the text describing the scene arrangement and a list indicating graspable objects in the text.

ii) Context Augmentation. Directives guide ChatGPT at a high level for scene descriptions but do not guarantee long-term quality due to hallucination [32]. To ensure consistent quality over a long time horizon, we employ a context augmentation by creating a self-reinforcing loop for ChatGPT. We initialize a prompt buffer to store generated prompts, with the first 50 samples *manually* assigned. We sample 10-15 scene descriptions from the prompt buffer each time and input them to ChatGPT. An example of this process is given in the following example

> **Q:** "Based on sample scene descriptions, generate a new scene description with a similar structure."
> **A:** "Certainly! Here's a new scene description: A spoon and a plate on a desk. [spoon, plate]"

The new sample is then appended to the prompt buffer. We repeat the process until 1M scene descriptions are generated.

Image Synthesis. Given the scene descriptions generated by ChatGPT, we use Stable Diffusion 2.1 [30] to generate images that align with the scene descriptions. We then gather instance segmentation masks for every object that appeared in the grasp list using the state-of-the-art visual grounding and instance segmentation models (OFA [12] and Segment-Anything [51]). At the end of the image synthesis stage, we obtain a grounding mask for each referenced object.

B. Grasp Pose Annotation

To annotate grasp poses for each object in the grasp list of the scene description, we employ a pretrained model RAGT-3/3 [6]. Since the candidate poses may not always be accurate, we use a traditional method by Kamon *et al.* [52] to further evaluate grasp poses. More specifically, we determine the grasp quality of each pose by calculating the net torque, denoted as $\mathcal{T}$, associated with the grasp as follows

$$\mathcal{T} = \underbrace{(\tau_1 + \tau_2)}_{\text{Resistance}} - \underbrace{RMg}_{\text{Torque}} \ . \tag{1}$$

The resistance at each contact point, denoted as τ_i, can be computed by $\tau_i = K\mu_s F \cos\alpha_i, \forall i \in \{1, 2\}$. In Equation (1), the terms M, g, K, μ_s, F correspond to the object's mass, gravitational acceleration, geometrical characteristics of the contact area, coefficient of static friction, and the applied force, respectively. These parameters are assumed to be constant across all grasps. Consequently, each grasp pose is characterized by three variables: R, α_1, α_2, as depicted in Fig. 3(b). Due to the impracticality of explicitly determining the physical terms for each object [53], the computation of $\mathcal{T}$ becomes infeasible without knowing the physical parameters M, K, μ_s. We employ the following concept to address the challenges posed by physical constraints

$$\tilde{\mathcal{T}} = \frac{\cos\alpha_1 + \cos\alpha_2}{R} \ . \tag{2}$$

The term $\tilde{\mathcal{T}}$ indicates the ratio between the resistance and the torque. By examining Equations (1) and (2), we verify that the original net torque $\mathcal{T}$ is correlated to $\tilde{\mathcal{T}}$. In cases a grasp results in fewer than two contact points, we assign each missing term $\cos\alpha_i$ a value of -1. Antipodal grasps tend to yield greater positive values of $\tilde{\mathcal{T}}$ compared to non-antipodal grasps [54]. Thus, grasp qualities can be ranked based on $\tilde{\mathcal{T}}$. We define grasps with positive $\tilde{\mathcal{T}}$ values as positive grasps, while the remaining grasps are considered negative. We remark that our grasp evaluation procedure can only be applied when the segmentation mask for each object is available to compute its convex hull. Some examples of our dataset can be found in Fig. 5.

C. Grasp-Anything Statistics

Number of Categories. To assess the diversity of object categories in our dataset, we conduct the same methodology as in [56]. We leverage 300 categories of LVIS [57] dataset and identify 300 candidate objects from Grasp-Anything for each category using a pretrained model [58]. Subsequently, we select a subset of 90,000 objects and filter out objects not semantically aligned with their assigned categories. Each category is considered significant if it has more than 40 objects. The outcomes of this analysis are showed in Fig. 4(a). With this setup, we observe that our Grasp-Anything spreads over 236 categories of the LVIS dataset. We apply the same procedure to other datasets and show the comparison in Fig. 4(b). It is noteworthy that VMRD [17] dataset is the

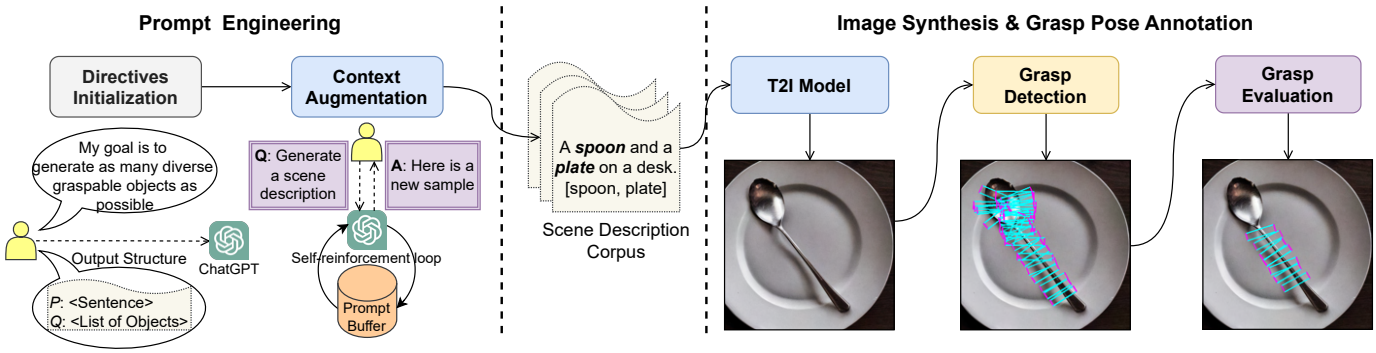
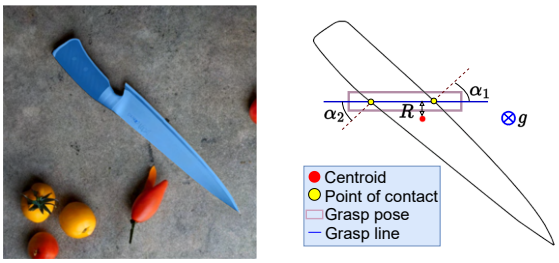


Fig. 2. **Dataset creation pipeline.**

TABLE II

BASE-TO-NEW GRASP DETECTION RESULT.

	Grasp-Anything (ours)			Jacquard [11]			Cornell [14]			VMRD [17]			OCID-grasp [18]		
Baseline	Base	New	H	Base	New	H	Base	New	H	Base	New	H	Base	New	H
GR-ConvNet [4]	**0.74**	**0.61**	**0.67**	**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 [18]	0.64	0.59	0.61	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 [55]	0.71	0.59	0.64	0.78	0.56	0.65	0.96	0.75	0.84	0.69	0.53	0.59	0.71	0.63	0.67



(a) Segmentation mask (b) Torque of a grasp pose

Fig. 3. **Grasp pose evaluation**. We construct a convex hull for each object segmentation mask and use the theory of [52] to determine feasible grasps.

runner-up and only has 115 LVIS categories, compared to 236 categories in our dataset.

Number of Objects. We visualize the number of objects (in log-scale) of Grasp-Anything and other grasp datasets in Fig. 4(b). From this figure, we can see that our Grasp-Anything dataset has a significantly larger number of objects than other datasets.

Number of POS tags. To categorize words in a text according to their grammatical roles and syntactic functions, we extract the POS tags [59] in our dataset and visualize them in Fig. 4(c). Our scene descriptions corpus utilizes a wide range of words to describe scene arrangements. We provide about 1.5M POS tags, with 35% being nouns, 20% being adjectives, 7% being verbs, and the remainder being other POS tags such as prepositions.

Object Shape Distribution. Fig. 6 compares object shape distributions between Grasp-Anything and Jacquard. We randomly select 5000 objects from each dataset and extract the (x, y) coordinates associated with pixels lying in the interior of each object. We then aggregate all coordinates and combine them to create heatmaps. These heatmaps are normalized by each dataset's image resolution. We can see that objects in Grasp-Anything span over a greater area than Jacquard's, indicating a greater degree of shape diversity.

D. *How will Grasp-Anything be helpful to the community?*

Given the large-scale nature and its multi-modalities, such as text prompts, images, and segmented masks, we hope that our new dataset can drive future research in these topics:

- *Grasp Detection*: Although many grasp datasets have been established [1], [3], [9], [11], [13], [14], we demonstrate that Grasp-Anything accommodates a broader range of objects and more natural scene settings; therefore, our dataset can advance more grasp detection research, especially on zero-shot grasp detection and domain adaptation.
- *Language-driven Grasping*: Language-driven grasping detection is a promising research area with industrial applications [46], [60], [61]. Grasp-Anything supports diverse scene descriptions, enabling large-scale training to align natural language with grasp detection.

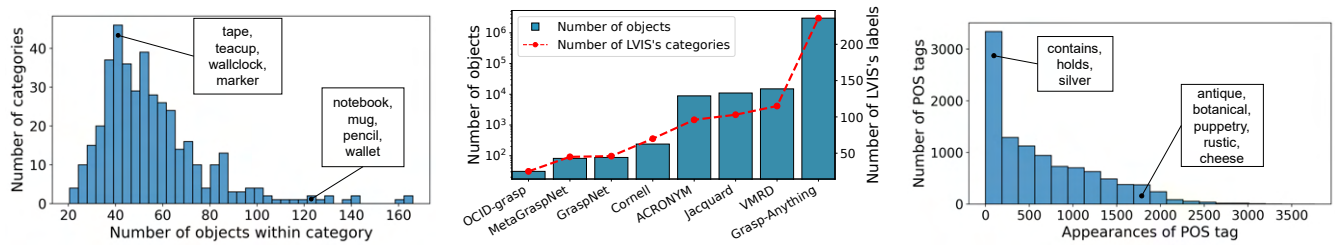
Furthermore, we believe Grasp-Anything can be useful in related tasks such as sim2real grasping [62], human-robot interaction [63], or language-driven mobile manipulation [39].

IV. EXPERIMENTS

We conduct experiments to validate the usefulness of our Grasp-Anything dataset and answer *two questions*: *i)* Can Grasp-Anything serve as a challenging dataset for grasp detection? and *ii)* Since our dataset is generated by the foundation models, will it work on real robot experiments?

A. *Zero-shot Grasp Detection*

Setup. We train three deep-learning grasp networks: GR-ConvNet [4], Det-Seg-Refine [18], and GG-CNN [55] on five datasets: Grasp-Anything, Jacquard [11], Cornell [14], VMRD [17], and OCID-grasp [18]. The primary metric is the success rate, defined similarly to [4], requiring an IoU score above 0.25 with the ground truth grasp and an offset angle less than 30°. For zero-shot learning's base and new labels [64], we utilize the classified LVIS's labels from Section III-C. Initially, we merge LVIS labels from all five



(a) Grasp-Anything sorted by LVIS categories. (b) Number of objects (in log-scale) comparison. (c) Grasp-Anything's POS tags.

Fig. 4. **Dataset Statistics**. We provide statistics on (a) the number of categories, (b) the number of POS tags, and (c) a comparison of number of objects.

Fig. 5. **Samples from Grasp-Anything.** For simplicity, we only display the grasp with the highest $\hat{\mathcal{T}}$ for each graspable object (marked in bold).

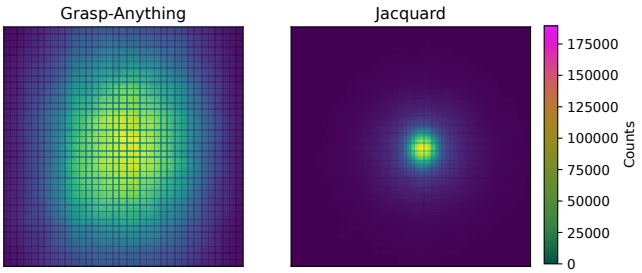


Fig. 6. **Object shape heatmap visualization.**

datasets and then identify the top 70% labels by occurrence. These labels form the 'Base' classes, while the remaining 30% become 'New' classes. We use the harmonic mean ('H') to measure the overall zero-shot success rates as in [64].

Base-to-New Generalization. We report the base-to-new grasp detection results in Table II. There are two central observations from the results. First, three baselines GR-ConvNet, Det-Seg-Refine, and GG-CNN exhibit satisfactory performances over five datasets, implying there is less room for model-centric approaches to improve the grasp detection results on each separated dataset. Second, Grasp-Anything is more challenging to train as our detection results are lower than related datasets using the same approaches due to the greater coverage of unseen objects in the testing phase.

Cross-dataset Transfer Learning. Table III presents the results of training GR-ConvNet on a dataset (row) and testing on another dataset (column). For instance, when a GR-ConvNet is trained on Jacquard and tested on Cornell, an accuracy of 0.51 is achieved. Our dataset improves in about $9-29\%$ over other datasets. Particularly, when testing on Jacquard, the performance of GR-ConvNet trained on Grasp-Anything is four times better than other datasets.

TABLE III

CROSS-DATASET RESULT USING GR-CONVNET [4] AS THE BASELINE.

Train \ Test	Jacquard	Cornell	VMRD	OCID-grasp	Grasp-Anything
Jacquard [11]	0.89	0.51	0.13	0.21	**0.17**
Cornell [14]	0.06	0.98	0.20	0.12	0.15
VMRD [17]	0.06	0.21	0.79	0.11	0.12
OCID-grasp [18]	0.09	0.12	0.20	0.74	0.12
Grasp-Anything (ours)	**0.38**	**0.60**	**0.32**	**0.37**	0.67

TABLE IV

GENERALIZED ZERO-SHOT GRASP DETECTION

		Cornell			VMRD			OCID-Grasp		
Baseline	Dataset Used	Base	New	H	Base	New	H	Base	New	H
GR-ConvNet [4]	Jacquard	0.53	0.50	0.51	0.18	0.10	0.13	0.23	0.19	0.21
GR-ConvNet [4]	Grasp-Anything	**0.68**	**0.57**	**0.62**	**0.33**	**0.29**	**0.31**	**0.42**	**0.37**	**0.39**
Det-Seg-Refine [18]	Jacquard	0.55	0.53	0.54	0.15	0.06	0.09	0.23	0.16	0.19
Det-Seg-Refine [18]	Grasp-Anything	**0.72**	**0.61**	**0.66**	**0.30**	**0.26**	**0.28**	**0.39**	**0.31**	**0.35**
GG-CNN [55]	Jacquard	0.45	0.39	0.42	0.10	0.07	0.08	0.18	0.10	0.12
GG-CNN [55]	Grasp-Anything	**0.64**	**0.56**	**0.60**	**0.22**	**0.18**	**0.19**	**0.36**	**0.27**	**0.31**

Generalized Zero-shot Learning. We conduct a generalized zero-shot grasp detection in Table IV. We compare our dataset with Jacquard, as it shows competitive results in Tables II and III. The other three datasets are used for testing. In all scenarios, Grasp-Anything considerably improves the performances of all baselines.

B. Robotic Evaluation

Setup. Our robotic evaluation on a KUKA robot is shown in Figure 7. We use GR-ConvNet [4] as the grasp detection network. The grasp detection results are converted into a 6DOF grasp pose using the depth image from a RealSense camera, as in [4]. The trajectory planner [65], [66] is used to execute the grasp. The evaluation is conducted for a single object and a cluttered scenario utilizing a set of 15 objects. We conduct experiments for each scenario 60 times with network weights of GR-ConvNet [4] trained on the proposed Grasp-Anything, OCID-grasp, VMRD, and Jacquard datasets, respectively.

Results. Table V statically shows the effectiveness of our proposed dataset, which helps the GR-ConvNet achieve the highest performance (93.3% for the single object and 91.6% for the cluttered scene) compared to other datasets. Several demonstrations can be found in our Demonstration Video.

TABLE V

REAL-WORLD GRASPING EXPERIMENTS.

Baseline	Dataset Used	Single Object	Cluttered Scene
GR-ConvNet [4]	Cornell	81.6%	68.3%
	Jacquard	85.0%	88.3%
	VMRD	78.3%	70.0%
	OCID-grasp	80.0%	71.6%
	Grasp-Anything (ours)	**93.3%**	**91.6%**

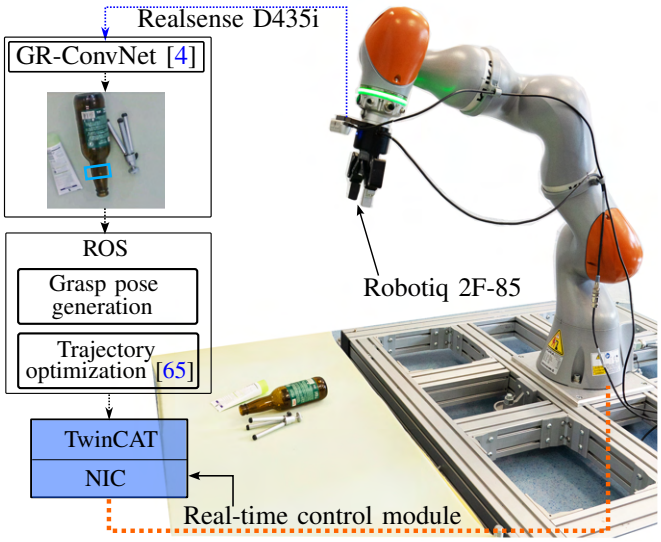


Fig. 7. **Overview of the robotic experiment setup.**

C. Grasp Detection in the Wild

We visualize grasp detection of a daily office arrangement image from GR-ConvNet trained by different datasets in Fig. 8. This figure shows that Grasp-Anything can improve the grasp detection quality over related datasets when the same baseline network is used. Fig. 9 showcases grasp detection examples utilizing a pretrained GR-ConvNet on the Grasp-Anything dataset on random images from the internet and different datasets. We can see that the detected grasp poses are adequate in quality and quantity.

D. Discussion

Through the zero-shot experiment (Section IV-A), we can answer our first question and conclude that our Grasp-Anything can be used as a challenging dataset for grasp detection as networks trained on our dataset achieve lower accuracy compared to other datasets (Table II). Furthermore, the cross-dataset experiment in Table III implies that utilizing our dataset on other datasets brings significant improvement. Also, the robotic experiments (Section IV-B) demonstrate that the model trained on our synthesis dataset outperforms the same model trained on different datasets with real-world images. This answers our second question and validates the usefulness of our dataset in real-world robotic experiments.

While promising results have been achieved with our dataset, we see several important improvement points. First, we remark that our dataset's creation is time-consuming and relies on access to the commercial ChatGPT API. Specifically, it took approximately three months to generate and process 1M scene descriptions on a cluster of three NVIDIA Quadro 8000. Fortunately, future research can reuse our provided assets (images, prompts, etc.) without starting from scratch. Second, our dataset currently lacks 3D point clouds. This is primarily because text-to-point-cloud or image-to-point-cloud foundation models have yet to achieve convincing results [67]. Therefore, creating point clouds from our prompts and images would make our dataset more useful in robotic tasks. Finally, since our dataset includes text prompts, we believe it will foster interesting research directions, such as language-driven grasping and human-robot interaction.

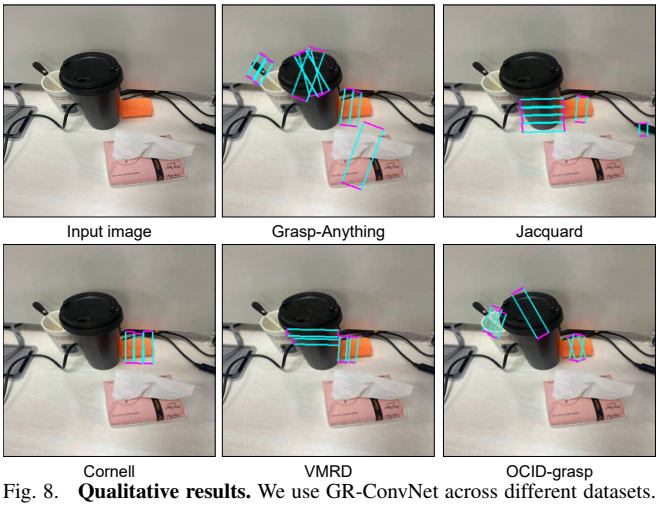


Fig. 8. **Qualitative results.** We use GR-ConvNet across different datasets.

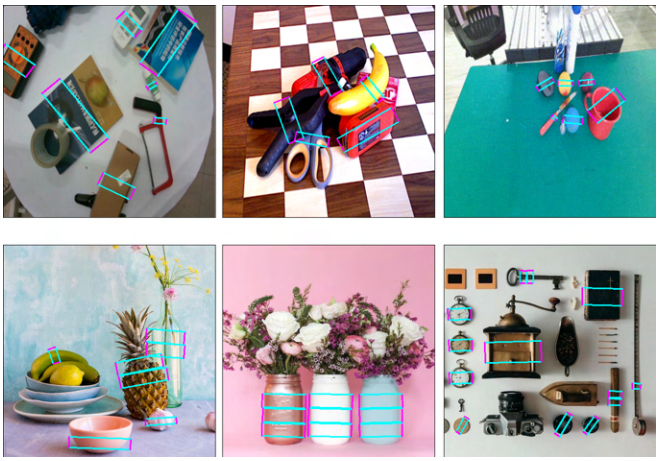


Fig. 9. **In the wild grasp detection with a model trained on our dataset.** The top row are images from other datasets: NBMOD [6], YCB-Video [20], GraspNet [21]; while the bottom row includes internet images.

V. CONCLUSIONS

We have presented Grasp-Anything, a new large-scale language-driven dataset for robotic grasp detection. Our analyses demonstrate that Grasp-Anything encompasses many objects and natural scene arrangements. The experiments on different networks and datasets with the real robot reveal that our Grasp-Anything dataset improves significantly over related datasets. By incorporating natural scene descriptions, we hope our dataset can serve as a foundation dataset for language-driven grasp detection.

REFERENCES

[1] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *ICRA*, 2016.

[2] A. Sahbani, S. El-Khoury, and P. Bidaud, "An overview of 3d object grasp synthesis algorithms," *Robotics and Autonomous Systems*, 2012.

[3] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *IJRR*, 2018.

[4] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in *IROS*, 2020.

[5] S. Song, A. Zeng, J. Lee, and T. Funkhouser, "Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations," *RA-L*, 2020.

[6] B. Cao, X. Zhou, C. Guo, B. Zhang, Y. Liu, and Q. Tan, "Nbmod: Find it and grasp it in noisy background," *arXiv preprint arXiv:2306.10265*, 2023.

[7] T. Bernardi, Y. Fleytoux, J.-B. Mouret, and S. Ivaldi, "Learning height for top-down grasps with the digit sensor," in *ICRA*, 2023.

[8] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *T-RO*, 2023.

[9] R. Platt, "Grasp learning: Models, methods, and performance," *Annual Review of Control, Robotics, and Autonomous Systems*, 2023.

[10] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. J. Wu, "A review of generalized zero-shot learning methods," *TPAMI*, 2022.

[11] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *IROS*, 2018.

[12] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *ICML*, 2022.

[13] M. Gilles, Y. Chen, T. R. Winter, E. Z. Zeng, and A. Wong, "Metagraspnet: A large-scale benchmark dataset for scene-aware ambidextrous bin picking via physics-based metaverse synthesis," in *CASE*, 2022.

[14] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation," in *ICRA*, 2011.

[15] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *arXiv preprint arXiv:1703.09312*, 2017.

[16] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," 2019, accessed: April 25th 2023. [Online]. Available: https://pybullet.org/wordpress/.

[17] H. Zhang, X. Lan, S. Bai, X. Zhou, Z. Tian, and N. Zheng, "Roi-based robotic grasp detection for object overlapping scenes," in *IROS*, 2019.

[18] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from rgb," in *ICRA*, 2021.

[19] X. Yan, J. Hsu, M. Khansari, Y. Bai, A. Pathak, A. Gupta, J. Davidson, and H. Lee, "Learning 6-dof grasping interaction via deep geometry-aware 3d representations," in *ICRA*, 2018.

[20] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[21] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *CVPR*, 2020.

[22] D. Kappler, J. Bohg, and S. Schaal, "Leveraging big data for grasp planning," in *ICRA*, 2015.

[23] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[24] M. Macklin, M. Müller, N. Chentanez, and T.-Y. Kim, "Unified particle physics for real-time applications," *TOG*, 2014.

[25] C. Eppner, A. Mousavian, and D. Fox, "A billion ways to grasp: An evaluation of grasp sampling schemes on a dense, physics-based grasp data set," in *ISRR*, 2019.

[26] D. Morrison, P. Corke, and J. Leitner, "Egad! an evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation," *RA-L*, 2020.

[27] C. Eppner, A. Mousavian, and D. Fox, "Acronym: A large-scale grasp dataset based on simulation," in *ICRA*, 2021.

[28] NVIDIA, "Nvidia omniverse platform for creating and operating metaverse applications," 2023, accessed: May 15th 2023. Available: https://www.nvidia.com/en-us/omniverse/. [Online]. Available: https://www.nvidia.com/en-us/design-visualization/omniverse/

[29] OpenAI, "Introducing ChatGPT," Software, accessed: July 6th 2023. [Online]. Available: https://openai.com/blog/chatgpt/

[30] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[31] A. Bucker, L. Figueredo, S. Haddadinl, A. Kapoor, S. Ma, and R. Bonatti, "Reshaping robot trajectories using natural language commands: A study of multi-modal data alignment using transformers," in *IROS*, 2022.

[32] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, "Chatgpt for robotics: Design principles and model abilities," *Microsoft Auton. Syst. Robot. Res*, 2023.

[33] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Generating situated robot task plans using large language models," *arXiv preprint arXiv:2209.11302*, 2022.

[34] I. Kapelyukh, V. Vosylius, and E. Johns, "Dall-e-bot: Introducing web-scale diffusion models to robotics," *RA-L*, 2023.

[35] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar, "Voyager: An open-ended embodied agent with large language models," *arXiv preprint arXiv:2305.16291*, 2023.

[36] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser, *et al.*, "Openscene: 3d scene understanding with open vocabularies," in *CVPR*, 2023.

[37] B. Liu, Y. Jiang, X. Zhang, Q. Liu, S. Zhang, J. Biswas, and P. Stone, "Llm+ p: Empowering large language models with optimal planning proficiency," *arXiv preprint arXiv:2304.11477*, 2023.

[38] A. Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. Varley, *et al.*, "Robots that ask for help: Uncertainty alignment for large language model planners," *arXiv preprint arXiv:2307.01928*, 2023.

[39] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[40] R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, J. Bohg, A. Morales, T. Asfour, D. Kragic, *et al.*, "Deep learning approaches to grasp synthesis: A review," *T-RO*, 2023.

[41] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *IJRR*, 2015.

[42] H. Liang, X. Ma, S. Li, M. Görner, S. Tang, B. Fang, F. Sun, and J. Zhang, "Pointnetgpd: Detecting grasp configurations from point sets," in *ICRA*, 2019.

[43] Z. Jiang, Y. Zhu, M. Svetlik, K. Fang, and Y. Zhu, "Synergies between affordance and geometry: 6-dof grasp detection via implicit representations," *arXiv preprint arXiv:2104.01542*, 2021.

[44] B. Wen, W. Lian, K. Bekris, and S. Schaal, "Catgrasp: Learning category-level task-relevant grasping in clutter from simulation," in *ICRA*, 2022.

[45] D. Liu, Y. Chen, and Z. Wu, "Digital twin (dt)-cyclegan: Enabling zero-shot sim-to-real transfer of visual grasping models," *RA-L*, 2023.

[46] J. Yang, W. Tan, C. Jin, B. Liu, J. Fu, R. Song, and L. Wang, "Pave the way to grasp anything: Transferring foundation models for universal pick-place robots," *arXiv preprint arXiv:2306.05716*, 2023.

[47] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[48] T. Yu, T. Xiao, A. Stone, J. Tompson, A. Brohan, S. Wang, J. Singh, C. Tan, J. Peralta, B. Ichter, *et al.*, "Scaling robot learning with semantically imagined experience," *arXiv preprint arXiv:2302.11550*, 2023.

[49] K. Mahowald, A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko, "Dissociating language and thought in large language models: a cognitive perspective," *arXiv preprint arXiv:2301.06627*, 2023.

[50] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," in *ACM CHI*, 2021.

[51] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[52] I. Kamon, T. Flash, and S. Edelman, "Learning to grasp using visual information," in *ICRA*, 1996.

[53] S. Caldera, A. Rassau, and D. Chai, "Review of deep learning methods in robotic grasp detection," *MTI*, 2018.

[54] I.-M. Chen and J. W. Burdick, "Finding antipodal point grasps on irregularly shaped objects," *IEEE transactions on Robotics and Automation*, vol. 9, no. 4, pp. 507–512, 1993.

[55] D. Morrison, P. Corke, and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," *arXiv preprint arXiv:1804.05172*, 2018.

[56] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3d objects," in *CVPR*, 2023.

[57] A. Gupta, P. Dollar, and R. Girshick, "Lvis: A dataset for large vocabulary instance segmentation," in *CVPR*, 2019.

[58] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li, *et al.*, "Regionclip: Region-based language-image pretraining," in *CVPR*, 2022.

[59] A. Tehseen, T. Ehsan, H. B. Liaqat, A. Ali, and A. Al-Fuqaha, "Neural pos tagging of shahmukhi by using contextualized word representations," *Journal of King Saud University-Computer and Information Sciences*, 2023.

[60] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[61] K. Xu, S. Zhao, Z. Zhou, Z. Li, H. Pi, Y. Zhu, Y. Wang, and R. Xiong, "A joint modeling of vision-language-action for target-oriented grasping in clutter," *arXiv preprint arXiv:2302.12610*, 2023.

[62] D. Horváth, K. Bocsi, G. Erdős, and Z. Istenes, "Sim2real grasp pose estimation for adaptive robotic applications," *arXiv preprint arXiv:2211.01048*, 2022.

[63] J. Mišeikis, P. Caroni, P. Duchamp, A. Gasser, R. Marko, N. Mišeikienė, F. Zwilling, C. De Castelbajac, L. Eicher, M. Früh, *et al.*, "Lio-a personal robot assistant for human-robot interaction and care applications," *RA-L*, 2020.

[64] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *CVPR*, 2022.

[65] M. N. Vu, F. Beck, M. Schwegel, C. Hartl-Nesic, A. Nguyen, and A. Kugi, "Machine learning-based framework for optimally solving the analytical inverse kinematics for redundant manipulators," *Mechatronics*, 2023.

[66] F. Beck, M. N. Vu, C. Hartl-Nesic, and A. Kugi, "Singularity avoidance with application to online trajectory optimization for serial manipulators," *arXiv preprint arXiv:2211.02516*, 2022.

[67] M. Deitke, R. Liu, M. Wallingford, H. Ngo, O. Michel, A. Kusupati, A. Fan, C. Laforte, V. Voleti, S. Y. Gadre, *et al.*, "Objaverse-xl: A universe of 10m+ 3d objects," *arXiv preprint arXiv:2307.05663*, 2023.