Grasp-CLIP: Pick Up What You Want

1st Zhao Gao
Institute of Automation, Chinese Academy of Sciences
Beijing, China
gaozhao2022@ia.ac.cn

2nd Jieren Deng
Institute of Automation, Chinese Academy of Sciences
Beijing, China
dengjieren2019@ia.ac.cn

3rd Zihao Wan
Institute of Automation, Chinese Academy of Sciences
Beijing, China
wanzihao2020@ia.ac.cn

4th Haojian Zhang
Institute of Automation, Chinese Academy of Sciences
Beijing, China
zhanghaojian2014@ia.ac.cn

5th Yunkuan Wang
Institute of Automation, Chinese Academy of Sciences
Beijing, China
yunkuan.wang@ia.ac.cn

6th Jianhua Hu
Institute of Automation, Chinese Academy of Sciences
Beijing, China
jianhua.hu@ia.ac.cn

***Abstract*—The demand for intelligent grasping by robots in the field of robot pick-and-place is continuously increasing. Particularly in scenarios with specific requirements for the sequence of object grasping, robots need human guidance to grasp different objects in a specified order to complete manipulation tasks. To address the slow inference speed, high parameter usage, and low precision in current methods, this work investigates how robots can successfully grasp a target object in cluttered scenes based on natural language instruction, aiming to achieve faster inference speed, reduced parameter usage, and improved precision. We propose the "Grasp-CLIP" network, which extends the capabilities of GSNet [1] by designing the GPRNet (Grasp Point Re-filtering Network). Specially, we introduce the CLIP [2] model into our framework and leverage three modalities: point cloud, image, and text to re-filter grasp points, by which we extract the grasp poses specific to the target object from the set of six-degree-of-freedom grasp poses output by GSNet. Our network outperforms existing methods regarding the precision of grasp point selection, the success rate of grasp poses for target object grasping, inference speed, and parameter usage. Integrating the CLIP model improves the generalization performance of the network, granting it superior adaptability when faced with objects beyond the training dataset. Through robotic arm grasping experiments, we demonstrate that the Grasp-CLIP network successfully outputs grasp poses for the target object, accomplishing precise grasping tasks as specified.**

***Index Terms*—deep Learning, robot grasping, CLIP, textual instruction**

I. Introduction

In the realm of robot manipulation, the demand for enhanced intelligence and human-robot interaction capabilities is increasing. Robots are expected to be more responsive and adept at following human guidance. Consider a scenario where a book sits on a table with an apple atop it. To efficiently pick up both items without causing damage, the robot must grasp the apple first and then the book. Such precise grasping sequences, communicated as commands, require robots to understand and execute specific object-grasping instruction from complex scenes based on textual language input.

Current approaches face challenges related to slow inference speed, high parameter usage, and imprecise grasp pose output for the target object. This study aims to address these challenges by investigating how robots can grasp a target object in cluttered scenes based on natural language instruction, with a focus on achieving faster inference speed, reduced parameter usage, and improved precision.

To tackle these challenges, we introduce the "Grasp-CLIP" network, which extends the capabilities of GSNet (Graspness-based Sampling Network) by integrating the "Grasp Point Re-filtering Network" and leveraging the CLIP (Contrastive Language-Image Pre-training) model which enables the understanding of the relationship between images and text by jointly learning them. To aid the training of GPRNet (Grasp Point Re-filtering Network) for this task, we also create a tabletop scene dataset within Isaac Sim, a simulation platform developed by NVIDIA.

Experimental results demonstrate the effectiveness of our Grasp-CLIP network which outperforms existing networks including the two-stage approach that combines image segmentation and GSNet. Our network, GPRNet, showcases superior precision in selecting candidate grasp points solely from RGB and depth images during the inference phase, offering fast inference speed and reduced parameter usage.

In summary, our contributions encompass the development of GPRNet, an end-to-end grasp point re-filtering network that has significantly enhanced the precision and speed of robotic grasping for target object, guided by natural language instruction, with reduced parameter complexity. Notably, by incorporating the CLIP model, we have increased the capacity for generalization within the network, enhancing its ability to adapt to novel objects not encountered during training, thereby propelling advancements in robotic manipulation and intel-

ligent grasping. Additionally, we build a simulation dataset featuring a wide array of objects, streamlining scene configuration, and data acquisition processes.

II. Related Work

A. Class-Agnostic Grasping in Multi-Object Scenes

In the domain of robot grasp pose detection, various methods employing an end-to-end approach have emerged, focusing on the direct estimation of robot grasp poses through the utilization of deep learning techniques. Based on the type of input data, these methods can be segregated into image-based grasp pose estimation and 3D point cloud-based grasp pose estimation strategies.

Efforts in image-based grasp pose estimation, such as those explored in DEXNET [3], have aimed to extend grasping to multi-object scenes. However, many of these methods are confined to the two-dimensional plane, which restricts the freedom of grasp poses. Conversely, grasp pose estimation based on point clouds involves extracting candidate grasp poses from point cloud inputs and assigning evaluation scores to these candidates. Notably, techniques such as PointNet-GPD [4], have been successful in generating grasps of 6-degree-of-freedom in multi-object settings. Despite this, their grasp success rates fall short compared to GSNet, which also utilizes point clouds as input and excels at producing 6-degree-of-freedom grasps with superior accuracy, efficiency, and inference speed. It is important to note that the image-based and point-cloud-based methodologies discussed earlier lack object category information in their output poses, leading to the inability to achieve intelligent grasping for specific targets.

B. Class-Specific Grasping in Multi-Object Scenes

CLIPort [5] combines Transporter Networks [6] with the CLIP model and learns the optimal grasp position and placement position for a target object from input RGBD image and textual instruction using a two-stream design. However, the grasps generated by this network are still limited to two-dimensional grasping. Other methods for specific object grasping mostly follow a two-stage approach, where the target object is first segmented from the scene and then a mature grasp network is used for grasping. Jianfeng Liao from Zhejiang University proposes using large language models [7] for robot grasp strategy formulation [8]. The method involves segmenting the image, matching the segmented objects with a textual description of the target object using the CLIP model for "positioning" and then selecting the grasp poses of the target object from the grasp poses output by Graspnet [9]. However, it does not provide an evaluation of the grasp success rate, and the inference speed may be slower due to the non-end-to-end nature of the approach. Peiqi Liu from Meta and New York University introduces a novel robot framework called OK-Robot [10], which utilizes the SAM [11] module to obtain a mask for the target object. This mask is then used to filter the grasp poses obtained from Any-Grasp (an upgraded version of GSNet, although not open-source [12]).

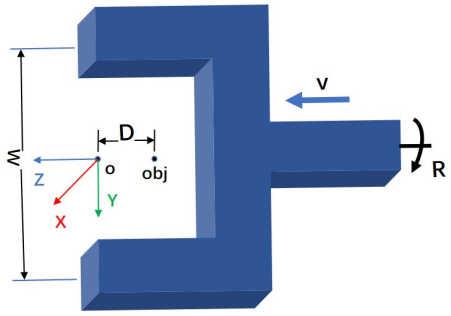

Fig. 1. Definition of a complete grasp pose for a gripper. "o" denotes the origin of the gripper frame, "obj" denotes the object point, "v" denotes the approaching vector, "R" denotes the in-plane rotation around the approaching axis, "D" denotes the approaching distance from grasp point to the origin of gripper frame, "W" denotes the gripper width.

It relies on both the segmentation module and grasp poses estimation module, sharing similar limitations as the robot grasping framework proposed by Jianfeng Liao.

In our pursuit of achieving more precise grasping for specific object within multi-object scenes with faster inference speed, reduced parameter usage, and improved precision, we introduce Grasp-CLIP. This framework aims to enable end-to-end 6-degree-of-freedom grasping for target object without relying on pre-existing image segmentation modules. To establish a robust foundation, we opt for GSNet, renowned for swiftly and accurately determining 6-degree-of-freedom grasps covering a diverse range of objects in multi-object scenarios, aligning seamlessly with our research objectives. Actually, GSNet is the evolution of GraspNet and introduces the Graspness [1] metric. The process in GSNet for detecting grasp poses involves filtering to identify 1024 grasp points with superior graspness values, subsequently deriving the optimal grasp pose for each grasp point alongside a corresponding grasp score [9] (as shown in Fig. 1).

III. Method

We present Grasp-CLIP, depicted in Fig. 2, an evolution of GSNet. Within the GPRNet framework, we have developed three key modules which will be discussed in sections III-A to III-C respectively.

A. CLIP-Based Image and Text Feature Fusion Module

We emphasize enhancing the generalization performance of the network, a primary objective achieved through the incorporation of the CLIP model. Leveraging extensive pre-training data, the CLIP model facilitates the alignment of text and image features. Drawing inspiration from the semantic flow design in CLIPORT, we leverage the image and text encoders of the frozen CLIP ResNet50 model to extract distinct image and text features. Within the decoder module, our methodology involves fusing text and image features via element-wise multiplication across multiple fusion steps. Ultimately, the output comprises pixel-level image features congruent in dimensions with the original input image, now

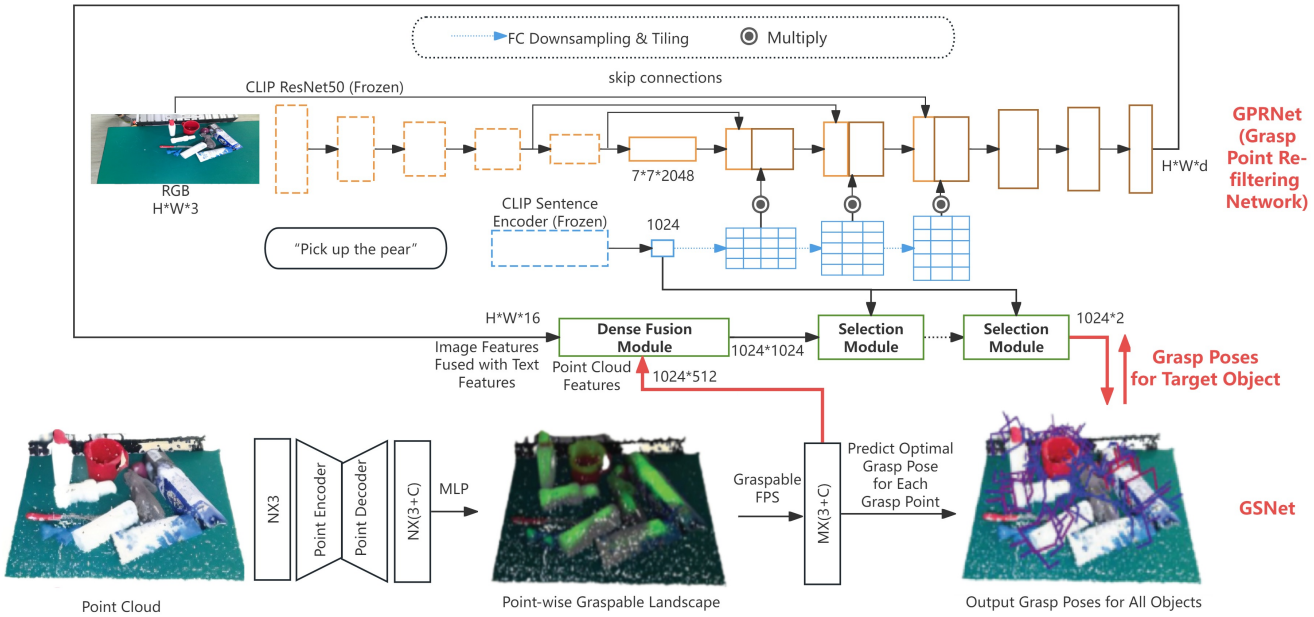

Fig. 2. Grasp-CLIP architecture. By leveraging the CLIP model, we obtain fused image features with textual features and merge them with the point cloud features output by GSNet through the "Dense Fusion Module". Subsequently, the fully fused multimodal features are processed through the "Selection Module" under textual instruction to accomplish the filtering of grasp points, by which we extract the grasp poses for the target object from all poses output by GSNet.

enriched with textual embeddings, reflecting our innovative approach.

B. Point Cloud and Image Feature Dense Fusion Module

The challenge of distinguishing visually similar objects solely relying on point cloud features necessitates the incorporation of color features for improved discrimination. To address this, we integrate the previously fused image features from section III-A with point cloud features extracted from GSNet. This fusion is not only motivated by the need for generalization but also crucial for precise grasp point selection.

The point cloud features are derived from 1024 candidate grasp points identified by GSNet through graspness metric filtering. These grasp points lack category information. The subsequent network of GSNet further derives the approach direction, optimal grasp rotation angle, grasp depth, gripper width, and corresponding grasp score for each grasp point. In essence, these 1024 points produce 1024 grasp poses. Here, the point cloud features generated by GSNet are our primary focus.

During the candidate grasp point selection stage in the original GSNet, the indices of the candidate points within the entire point cloud are known, and each point in the point cloud corresponds to a pixel in the image (referring to the original input image) on a one-to-one basis. Consequently, it is straightforward to align each candidate grasp point in the point cloud with a specific image pixel. Inspired by Dense Fusion [13], we directly contact the corresponding point features and pixel features to generate the fused multimodal features, which will serve as the core of the "Grasp Point Selection Module".

C. Grasp Point Selection Module Based on Textual Instruction and Fused Multimodal Features

Having obtained fused features from the point cloud, image, and text data, we design the "Selection Module" (refer to Fig. 3) to identify the corresponding object points guided by textual language instruction. The text features, generated by the CLIP text encoder, are integrated with fused features of the 1024 grasp candidates, involving channel transformations, replication, and fusion using element-wise multiplication and addition. The resulting features pass through an MLP layer. Iterating this module multiple times (4 times in this study) yields binary classification decisions for each of the 1024 grasp candidates: 1 signifies belonging to the target object, while 0 indicates exclusion. Addressing the imbalance in positive and negative samples, we employ focal loss as the loss function (refer to (1)). These decisions guide the extraction of target grasp poses from the initial 1024 poses produced by GSNet for precise object manipulation.

$$FL = -\frac{1}{N}\sum_{i}^{N}\alpha_{y_i}(1-p_{iy_i})^{\gamma}log(p_{iy_i}), \tag{1}$$

where N represents the total number of samples, here referring to the total number of candidate grasp points in a scene. y_i denotes the true class of the i-th sample, α_{y_i} represents the class weight for class y_i, γ is used to adjust the weight of

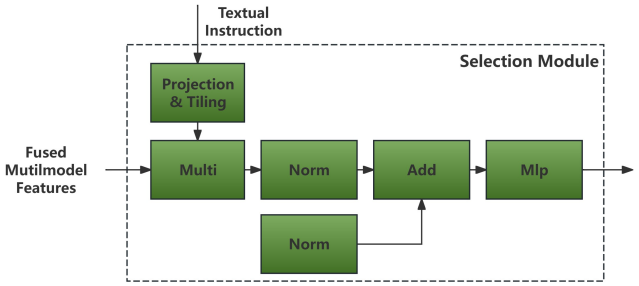

Fig. 3. Grasp point selection module based on textual instruction and fused multimodal features. In each selection module, the fused multimodal features are first multiplied by the text features. Subsequently, the multiplied result is added to the original multimodal features, followed by passing through an MLP layer.

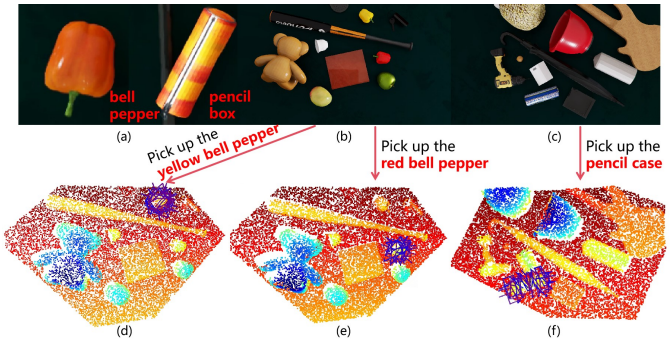

Fig. 4. Visualizations of specific target object grasping generated under textual guidance by Grasp-CLIP on the test unseen dataset. (a) showcases the **"bell pepper"** and **"pencil box"** from the training dataset. On the test unseen dataset, commands are issued to "Pick up the **yellow bell pepper**" and "Pick up the **red bell pepper**" as depicted in (b), with the resultant grasp poses generated by Grasp-CLIP illustrated in (d) and (e) respectively. Additionally, for the scenario illustrated in (c), the command "Pick up the **pencil case**" is given, and the corresponding grasp pose output by Grasp-CLIP is shown in (f).

difficult-to-classify samples, and p_{iy_i} indicates the probability of the i-th sample being predicted as class y_i.

D. Building Our Dataset

This work aims to empower the network to filter grasp poses for novel objects previously unseen by the robot, necessitating a diverse training dataset while the GraspNet-1Billion dataset used by GSNet is limited in object categories (only 88 with a subset for training).

In this work, we leverage the NVIDIA Isaac Sim robot simulation environment to establish a data augmentation platform and create a unified paradigm for data generation, achieving efficient one-click data generation. Initially, we manually select 266 objects suitable for tabletop grasping from thousands of objects provided by the "Behavior-1K" dataset [14], constructing 142 training scenes with 10 randomly chosen objects each. For each scene, we randomly initialize the relative poses of the objects on the tabletop and capture 255 RGB images, depth images, and instance segmentation labels, totaling 36,210 training images from all scenes for Grasp-CLIP.

IV. Expriments and Results

A. Generalization Performance of Grasp-CLIP

Grasp-CLIP is implemented with PyTorch and trained on the training dataset for 10 epochs using one Nvidia RTX 3090 GPU. The Adam optimizer is employed with a batch size of 4. The initial learning rate is set to 0.001 and is reduced by 5% after every epoch. We also evaluate the effectiveness of our model by reconstructing 27 scenes ("test seen") using the same 266 objects employed in the training dataset. Additionally, to assess the generalization performance of Grasp-CLIP, we collect an additional 100 objects from the Behavior-1K dataset to establish a "test unseen" dataset comprising 10 scenes. The placement of objects in these 10 scenes undergoes minimal manual adjustments to reduce occlusion and prevent objects from moving out of the view of the camera due to excessive randomness. These 100 objects were not seen by the network during training and include variations like different appearances and specific descriptions: (i) objects of the same category as those in the training dataset but with different appearances. (ii) objects with additional specific descriptions such as color, size, and shape. This evaluation aims to assess whether the network can recognize novel-colored bell peppers and distinguish between bell peppers of different colors. (iii) We also collect some objects of the same category as those in the training dataset but assign them new category names. Visualizations of object grasping under textual guidance by Grasp-CLIP on the test unseen dataset are depicted in Fig. 4. The grasp poses of different bell peppers with novel colors can be successfully distinguished, as can the novelly described pencil case, showcasing the robust generalization performance of the network. This achievement is credited to the integration of CLIP, facilitating effective alignment between textual descriptions and images.

B. Comparing with Representative Methods

We benchmark the superiority of our Grasp-CLIP by comparing its 6-DoF grasp success rate with other state-of-the-art grasping networks. Our comparisons include CLIPORT and a two-stage approach which first segments the mask of the target object by using "Grounded SAM" module [15] based on textual instruction and then feeds the corresponding point cloud into GSNet to output the grasp pose for the target object. The tasks of CLIPORT involve "packing seen Google objects group", allowing for grasping a specified object.

To ensure fairness, we retrain CLIPORT on our training dataset and evaluate it on both the test seen and test unseen datasets. As CLIPORT outputs a single optimal grasp position, we convert its generated 2D grasp into a 6-DoF grasp.

For each method, the final comparison involves calculating the average grasp success rate of the best 6-DoF grasp across different friction coefficients, assessed using the tool of GraspNet [9]. We also evaluate the precision and recall of grasp pose (grasp point) filtering separately on the test seen and test unseen datasets. Furthermore, we assess the theoretical upper

TABLE I

EVALUATION RESULTS ON THE TEST SEEN AND TEST UNSEEN DATASETS. "PRE": THE PRECISION OF GRASP POSES FILTERING; "REC": THE RECALL OF GRASP POSES FILTERING; "GSR": THE AVERAGE GRASP SUCCESS RATE OF THE TOP-1 GRASP POSE; "REAL GSR": THE THEORETICAL UPPER LIMIT OF THE AVERAGE GRASP SUCCESS RATE OF THE TOP-1 GRASP POSE BASED ON GSNET OUTPUTS; "PARAMS": THE TOTAL NUMBER OF MODEL PARAMETERS FOR DIFFERENT METHODS; "INFERENCE TIME": THE TIME TAKEN FOR A SINGLE INFERENCE. IN THE INTEREST OF FAIRNESS, ALL THE INFERENCE IS CONDUCTED ON A SINGLE NVIDIA RTX 3090 GPU.

Methods	Test Seen				Test Unseen				Params	Inference Time
	pre	rec	gsr	real gsr	pre	rec	gsr	real gsr		
CLIPORT	-	-	3.47%	-	-	-	2.96%	-	0.21B	0.401s
Grounded SAM + GSNet	50.94%	50.47%	8.40%	16.29%	62.04%	**61.35%**	10.40%	16.31%	0.83B	0.625s
Ours(Grasp-CLIP)	**84.79%**	**85.54%**	**14.53%**	16.29%	**69.00%**	47.26%	**11.11%**	16.31%	**0.18B**	**0.238s**

TABLE II

ABLATION RESULTS ON THE TEST SEEN AND TEST UNSEEN DATASETS. THE "-" SYMBOL INDICATES "REMOVING" OR "BEING REPLACED WITH ANOTHER MODULE".

Methods	Test Seen		Test Unseen	
	pre	rec	pre	rec
Grasp-CLIP	**84.79%**	85.54%	**69.00%**	47.26%
-Selection Module	**85.44%**	89.75%	**64.00%**	45.52%
-Dense Fusion Module	41.05%	35.50%	25.91%	19.42%
-CLIP ResNet50	83.85%	93.37%	60.42%	48.97%

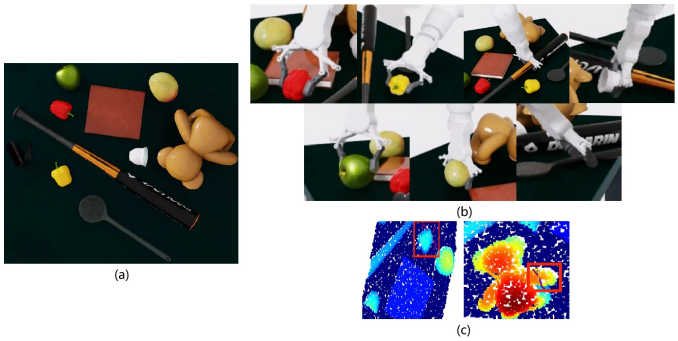

Fig. 5. Results of robot grasping experiment. (a) Scenario 2 from the test unseen dataset. (b) The robot successfully completes the grasp after issuing the pick-up commands for seven distinct objects separately. (c) When issuing "Pick up the red hardback", Grasp-CLIP erroneously outputs the grasp pose for a red bell pepper. While Grasp-CLIP successfully outputs the grasp pose for the large brown teddy bear, its unique shape leads to collisions during the subsequent grasping process.

limit of the grasp success rate of GSNet under the assumption of perfect grasp pose filtering for the target object. See Table I for detailed results.

Results and Analysis When it comes to the output optimal grasp position from CLIPORT, considering the transition from 2D grasp to 6-DoF grasp, we only calculate the success rate of the transformed 6-DoF grasp. As shown in Table I, our method demonstrates superior performance compared to the other two approaches across various aspects, including grasp success rate, number of parameters, and inference time on both the test seen and test unseen datasets. Moreover, our method achieves a grasp success rate that approaches the theoretical upper limit outstandingly. Particularly on the test unseen dataset, our method achieves an impressive grasp pose filtering precision of 69%, highlighting the exceptional generalization capabilities of our network, facilitated by leveraging the CLIP model. Compared to the traditional two-stage object-specific grasping method, our approach effectively reduces parameters by 80%, resulting in a 1.7-fold increase in inference speed while maintaining higher precision in grasp pose filtering. This improvement stems from our decision to avoid relying on an image segmentation module, thus avoiding the potential risks of inaccurately segmenting the target object. Instead, we have adopted an end-to-end design, which not only speeds up inference but also reduces parameter usage. The utilization of multimodal features has significantly contributed to the enhancement of precision in our grasp pose filtering.

C. Ablation Studies

We conduct three ablation experiments to gain insights into the individual contribution and necessity of each component in our approach. When we design GPRNet, we place greater emphasis on the precision of grasp point filtering rather than ensuring that all grasp points of the target object are successfully selected. This is because in practical grasping scenarios, our goal is to accurately pick up the object rather than generating as many grasp poses as possible. The improvement in recall rate often comes at the cost of decreased precision, which is not desirable in our case. Additionally, we anticipate the network to exhibit improved generalization performance. Therefore, our main emphasis lies on the precision of grasp point selection, particularly when dealing with the test unseen dataset. As shown in Table II: when removing the "Selection Module", precision slightly rises for the test seen dataset but notably drops for the test unseen dataset, which emphasizes the effectiveness of the "Selection Module". Removing the "Dense Fusion Module" and relying solely on the point cloud modality significantly reduces performance across all metrics, highlighting the indispensable nature of the image modality. Substituting the CLIP ResNet50 image encoder with a regular ResNet50 image encoder results in decreased precision on both

the test seen and unseen datasets, demonstrating the strong and effective capabilities of the CLIP model in enhancing the generalization ability of the network. In summary, each module we design in GPRNet is effective, contributing to the superior performance of our network compared to other state-of-the-art networks.

D. Robotic Arm Grasping Simulation Experiments

We conduct practical robot grasping tests in NVIDIA Isaac Sim to showcase the effectiveness and generalization of Grasp-CLIP. Using the "COBOTTA PRO 900" robotic arm, we perform actual grasping experiments on a scene from the test unseen dataset. After specifying the grasping target through instruction, we test the grasping performance using the optimal grasp pose provided by Grasp-CLIP. Successful grasping is achieved for most objects except the large brown teddy bear, black key chain, and red hardback, despite Grasp-CLIP generating precise grasp poses for all of them. The inability to grasp the large brown teddy bear or black key chain is solely due to their complex shapes, unique placement positions, or orientations, causing the robot to collide with the target object before reaching the grasp pose, resulting in failed attempts. Implementing certain robotic arm trajectory controls could prevent such occurrences. As for the red hardback, Grasp-CLIP confuses it with the red bell pepper, which shares the same color as the red hardback. This confusion arises from the the failure of the network to effectively learn or distinguish the features of the "hardback". Even if Grasp-CLIP successfully generates precise grasp poses for it, the hardback lying completely flat on the table makes grasping significantly more challenging. Regardless, Grasp-CLIP accurately distinguishes and grasps seven objects, none of which are encountered during training, highlighting its strong generalization and adaptability. Please refer to Fig. 5 for specific results.

V. Conclusion

We propose Grasp-CLIP, an end-to-end framework comprising GSNet and GPRNet which leverages the CLIP model to achieve natural language-guided target object grasping by effectively utilizing information from point cloud, image, and text modalities, resulting in faster, more precise, and highly generalizable inference capabilities compared to existing methods while utilizing fewer parameters, even when encountering novel objects. Furthermore, our work includes the establishment of a simulation dataset and the development of a standardized data generation process which will all be publicly accessible, providing valuable support for tasks such as 3D vision. Moving forward, our future research will explore the potential of GPNet for few-shot learning and its plug-and-play nature, aiming to further enhance its scalability and applicability in various robotic scenarios. Overall, Grasp-CLIP showcases the seamless integration of language guidance and multi-modal information for advanced robotic grasping, paving the way for more intelligent and efficient robotic manipulation.

References

[1] C. Wang, H.-S. Fang, M. Gou, H. Fang, J. Gao, and C. Lu, "Graspness discovery in clutters for fast and accurate grasp detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15964–15973, 2021.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[3] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, p. eaau4984, 2019.

[4] 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 *2019 International Conference on Robotics and Automation (ICRA)*, pp. 3629–3635, IEEE, 2019.

[5] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on robot learning*, pp. 894–906, PMLR, 2022.

[6] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani, *et al.*, "Transporter networks: Rearranging the visual world for robotic manipulation," in *Conference on Robot Learning*, pp. 726–747, PMLR, 2021.

[7] 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.

[8] J. Liao, H. Zhang, H. Qian, Q. Meng, Y. Sun, Y. Sun, W. Song, S. Zhu, and J. Gu, "Decision-making in robotic grasping with large language models," in *International Conference on Intelligent Robotics and Applications*, pp. 424–433, Springer, 2023.

[9] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11444–11453, 2020.

[10] P. Liu, Y. Orru, C. Paxton, N. M. M. Shafiullah, and L. Pinto, "Ok-robot: What really matters in integrating open-knowledge models for robotics," *arXiv preprint arXiv:2401.12202*, 2024.

[11] 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," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

[12] 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," *IEEE Transactions on Robotics*, 2023.

[13] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3343–3352, 2019.

[14] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun, *et al.*, "Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation," in *Conference on Robot Learning*, pp. 80–93, PMLR, 2023.

[15] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.