Reasoning Tuning Grasp: Adapting Multi-Modal Large Language Models for Robotic Grasping

Jinxuan Xu [*†] **Shiyu Jin** [‡] **Yutian Lei** [‡] **Yuqian Zhang** [†] **Liangjun Zhang** [‡]

Abstract:

Large language models (LLMs) have garnered increasing popularity owing to their remarkable reasoning capabilities. However, their primary utility within the field of robotics has predominantly been constrained to tasks related to manipulation planning, primarily due to their inherent text-based outputs. To overcome this limitation, this paper explores the potential of LLMs in the realm of numerical predictions in robotics, with a specific focus on the task of robotic grasping. We propose Reasoning Tuning, a novel approach that harnesses the extensive prior knowledge embedded within LLMs, optimizing them for tasks involving numerical prediction. This method empowers LLMs, notably with multi-modal capabilities, to generate precise numerical outputs, such as grasp poses for robot arms. The proposed method is extensively validated on the grasping benchmark and real-world grasping experiments, demonstrating that multi-modal LLMs can be adapted for numerical prediction tasks in robotics. This not only extends their applicability but also bridges the gap between text-based planning and direct robot control utilizing LLMs. More details and videos of this work are available on our project page: https://sites.google.com/view/rt-grasp.

Keywords: Robotic Grasping, Large Language Model

1 Introduction

The growth of artificial intelligence in recent years has been significantly driven by the emergence of large language models (LLMs). These models, with their immense knowledge, have been transforming how we approach various tasks, especially those involving language. In the world of robotics, the potent reasoning capabilities of LLMs have become indispensable for robot manipulation planning [1, 2, 3], enhancing adaptability across diverse real-world scenarios and fostering interactive engagements with humans. However, despite the potential that LLMs bring to the field of robotics, their application has predominantly been limited to planning tasks. A notable bottleneck lies in the textual nature of LLM outputs, which often presents challenges in integration with the numerical requisites essential for direct robot control.

Concurrently, multi-modal large language models have captured significant attention [4, 5]. These models equipped with the capability to understand not just text but also images, have redefined what we thought was possible with LLMs. In the realm of robotics, they bridge the gap between perception and planning, then address a variety of embodied reasoning tasks [6, 7]. However, their image understanding is mostly general, meaning they can tell what is the object in a picture but not where exactly it is located, as illustrated in Figure 1. Consequently, while robots can leverage the linguistic wisdom of LLMs in planning tasks, they have yet to capitalize on their potential for direct numerical manipulation.

[*]Work done while the author was an intern at Baidu Research, USA.

[†]Jinxuan Xu and Yuqian Zhang are with Rutgers University, Department of Electrical and Computer Engineering.

[‡]Shiyu Jin, Yutian Lei, and Liangjun Zhang are with Robotics and Autonomous Driving Lab (RAL), Baidu Research, USA.

On the other hand, traditional methods for robotic numerical prediction tasks face restrictions in real-world applications due to their absence of reasoning capabilities. Take the case of robotic grasping task: most existing methods [8, 9] use CNN-based architectures to predict the grasp pose. While some of them excel in experimental accuracy on benchmark datasets, their real-world applications remain constrained. For instance, traditional models might produce theoretically correct predictions, but these could be impractical for real scenarios like grippers with width constraints, as illustrated in Figure 1. Furthermore, these models also lack the capability to refine their predictions according to different situations because of their absence of reasoning faculties.

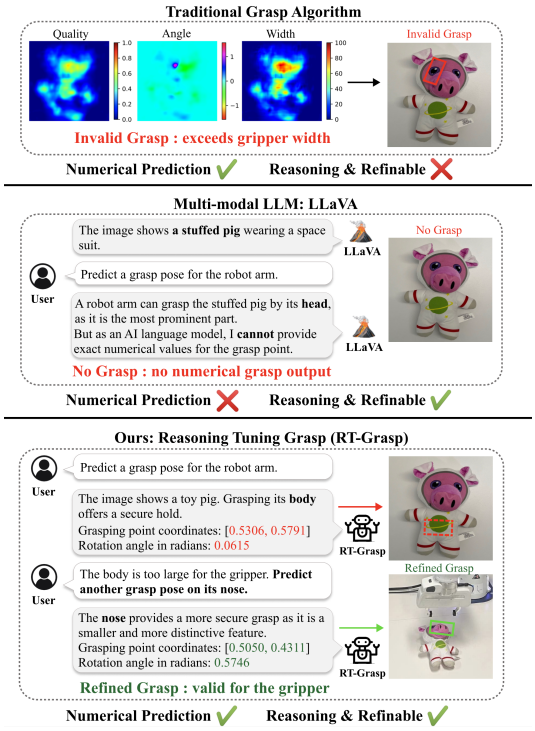

Figure 1: Comparing three robotic grasping approaches: 1) Traditional CNN-based algorithms produce fixed poses, which lack adaptability in practical situations. 2) Multi-model LLMs output adaptable grasping strategies but lack precise numerical predictions. 3) Ours combines the best of both, predicting adaptable numerical grasping informed by reasoned strategies.

Here a question is posed: *can the reasoning capabilities inherent in LLMs be utilized for numerical prediction tasks in robotics?* This paper offers a positive answer, showcasing an adaptation of multi-modal LLMs to robotic grasping tasks. Multi-modal LLMs provide more than just advanced image understanding; they also possess rich prior knowledge and reasoning capabilities. These traits make them excellent candidates for robotic tasks.

To efficiently utilize the reasoning capability of LLMs for numerical predictions, we introduce a novel approach, Reasoning Tuning. This approach introduces a crucial reasoning phase before the numerical prediction step during training, directing the model to base its predictions on sound logical reasoning, thereby unlocking the valuable information encapsulated within LLMs. Reasoning Tuning aims to amplify the existing wealth of knowledge on object concepts, shapes, structures, and materials already embedded within LLMs, thus enhancing their efficacy in numerical predictions for robotic tasks.

Furthermore, another benefit of adding a reasoning phase prior to numerical prediction is data efficiency. Typically, large-scale datasets are essential for training or fine-tuning LLMs, posing a challenge for adapting them to downstream tasks. However, a pre-trained LLM already contains high-level information about object attributes. Thus, a reasoning phase that capitalizes on this prior knowledge effectively reduces the need for expansive visual datasets. In this paper, we exemplify this approach using the robotic grasping task, demonstrating that the method not only can harness the prowess of multi-modal LLMs for precise numerical predictions but also requires only a limited dataset.

Moreover, we explore two economical training strategies: pre-training and Low-Rank Adaptation (LoRA) fine-tuning [10]. Our intent behind this investigation is to present a more resource-efficient method for transferring the capabilities of multi-modal LLMs to downstream robotic tasks.

In summary, we adapt multi-modal LLMs for numerical prediction tasks, with a focus on robotic grasping. Our approach not only enables reasoning capabilities but also allows for refinable predictions, as illustrated in Figure 1. The main contributions can be summarized as follows:

- We introduce Reasoning Tuning, a novel and data-efficient methodology that leverages the inherent prior knowledge of pre-trained LLMs, facilitating their adaptation to tasks requiring numerical predictions.
- We investigate two computationally efficient training strategies: pre-training and LoRA fine-tuning, designed to adapt multi-modal LLMs seamlessly into the downstream applications.
- We validate the efficacy of our proposed method on the grasp benchmark dataset and also conduct hardware experiments on real robots to demonstrate the performance.

2 RELATED WORK

2.1 Robotic Grasping

Traditionally, robotic grasping has leaned heavily on analytical approaches [11, 12, 13]. These techniques primarily focus on understanding the geometry of objects or analyzing the contact force to find a grasp that maximizes stability. However, these methods often struggle to generalize well to unseen objects and can fail when faced with irregular objects.

Data-driven methods, especially those using convolutional neural networks (CNNs), have shown promising results in recent years [14, 15, 16, 17, 9, 8, 18]. These approaches leverage large datasets of labeled grasping examples to train models capable of making grasp predictions. Despite their success, these models often suffer from overfitting. They cannot also reason the usage, category, material, and other properties of objects beyond their shape. This limitation restricts their effectiveness in real-world scenarios, particularly when grasping objects with unusual shapes or those requiring special handling due to their material properties or intended use.

2.2 Language Grounding for Robotics

Language-conditioned Robotic Manipulation. In recent years, the appeal of natural language has propelled research into language-conditioned robotic manipulation. Studies [19, 20, 21, 22] have explored grasp detection grounded following language instructions in scattered scenes. [23] performs grasping prediction based on language descriptions of an object's properties. And building on advancements in language models [24, 25], recent studies [26, 27, 28, 29, 30, 31] have successfully grounded more flexible language instructions into long-horizon manipulation tasks. However, these methods require lots of demonstrations to master image-based policies and additionally, they have to address the challenges of sim2real.

LLMs for Robotic Manipulation. With the rise of LLMs, there has been a surge in research exploring their capabilities for manipulation. Many studies [2, 3, 32] have integrated LLMs into closed-loop planning structures, decomposing language-conditioned long-horizon tasks into small steps. Yet, the gap between language instructions and actions still remains. Furthermore, some studies [33, 34, 35] have employed program-like specifications to prompt LLMs, melding planning and action using a predefined library of action functions. While intriguing, these methods are often constrained by the limitations of basic action functions and typically require extra perception models, reducing both system efficiency and flexibility. Recent studies [7] have made strides in bridging the planning-action gap using multi-modal LLMs, but the method has high data and computational requirements, which hinder its applicability in real-world scenarios. Different from the above, our work is able to utilize the prior knowledge embedded within pre-trained LLMs and achieve precise numerical predictions in the field of robotics.

3 ROBOTIC GRASPING

In this work, the robotic grasping problem is defined as finding an antipodal grasp, perpendicular to a planar surface, given an n-channel image. Similar to [36, 9], the grasp pose can be parameterized as $g = \{x, y, \theta, w\}$, where (x, y) indicates the 2D coordinates signifying the center point of the grasp

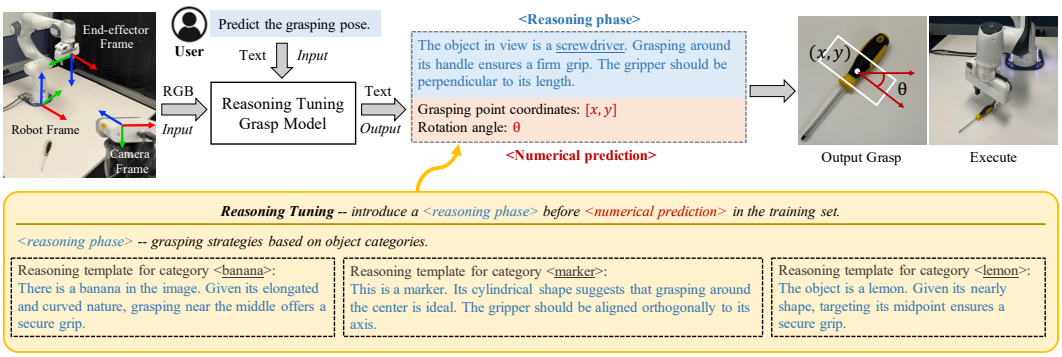

Figure 2: Overview. The proposed method processes RGB images and user instructions to yield text outputs, which comprise both a reasoning phase and a numerical grasp pose prediction $p = \{x, y, \theta\}$. The reasoning phase analyzes the object's shape and structure based on its category and generates corresponding grasping strategies. Figure illustrates examples of reasoning templates in the training set for three distinct categories.

pose; θ denotes the rotation angle of the gripper compared to the horizontal axis; w represents the width of the grasping box. In many related studies, the inclusion of w within the predicted grasp pose g is usually considered non-essential [37], due to the variations in grippers and camera configurations. To this end, our study, with its primary focus on probing the efficacy of LLMs in robotic grasping tasks, assumes w equals the maximum width of the gripper. Hence, the grasp pose considered in this paper can be defined by $p = \{x, y, \theta\}$. Here (x, y) coordinates are normalized by image width and image height respectively, and rotation angle θ is represented in radians scaling to $(-\frac{\pi}{2}, \frac{\pi}{2})$, as shown in Figure 2.

4 RT-Grasp

In this section, we introduce Reasoning Tuning for robotic grasping (RT-Grasp), a method specifically designed to bridge the gap between the inherent text-centric nature of LLMs and the precise numerical requirements of robotic tasks. Its primary objective is to exploit the vast prior knowledge encapsulated within LLMs to their maximum potential, optimizing it for numerical prediction. Through this method, the multi-modal LLM is able to generate precise numerical outputs, complemented by accompanying reasoning.

Our model builds upon the multi-modal LLM, specifically the Large Language and Vision Assistant (LLaVA) [5]. This model connects a visual encoder and LLaMA [38], an open-source LLM that matches the performance of GPT-3 [39]. LLaVA is originally trained on image-text paired datasets for general-purpose visual and language understanding. In alignment with this, we propose a novel method, Reasoning Tuning, and create our image-text dataset, named Reasoning Tuning VLM (Visual Language Model) Grasp dataset, aiming to utilize the intrinsic knowledge of LLMs. More details are introduced in section 4.1. In addition, to address the challenge of computational costs, we have ventured into two different frameworks, pre-training and LoRA fine-tuning, which are discussed in section 4.2.

4.1 Reasoning Tuning

An intuitive approach to adapting language models for numerical predictions is to directly train them using numerical values in textual form. However, our preliminary exploration of this approach yields limited performance, producing invalid outputs like exceedingly large values. We conjecture that the LLM finds it challenging to align numerical outcomes with their semantic significance. To this end, we introduce Reasoning Tuning, a methodology that integrates a reasoning phase before the model provides its numerical predictions. It compels the LLM to engage in reasoning about the image and task before producing any precise numbers. In our Reasoning Tuning VLM Grasp dataset, the target text was structured to include a reasoning segment, followed by the textual grasp pose, as illustrated in Figure 3 under the `With Reasoning` dataset variant.

This integrated reasoning phase is designed to leverage LLM's intrinsic knowledge and reasoning capability, thereby generating a grasp pose tailored to the specific object in the image. For example, consider cups, which may exhibit varied appearances in terms of color, design, or material. But a general grasping strategy for them – targeting the handle or the upper edge – is universal for robotic manipulation. Contemporary LLMs, such as LLaMA, are repositories of comprehensive knowledge about object properties such as shapes and structures. Utilizing this knowledge, the reasoning phase directs LLMs to first establish the grasping strategies, setting the stage for a more informed numerical prediction in the subsequent step.

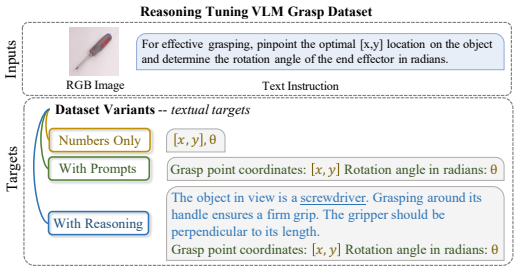

Figure 3: Reasoning Tuning VLM Grasp dataset. This image-text paired dataset contains three variants based on target texts and sources its RGB images from the Cornell Grasp dataset [14].

In our Reasoning Tuning VLM dataset, the reasoning texts commence by identifying the object's category and overarching shape. Then it outlines the general grasping strategy based on this specific category. For every object category in the dataset, we create a series of reasoning templates. To ensure the quality of these templates, we adopted a multi-step approach. First, we prompt ChatGPT [40] to generate a collection of templates tailored to each category. This was followed by instructing it to refine these drafts, excising redundant or irrelevant sentences. As a final quality check, we manually verify the correctness and relevance of generated templates. Some examples of our reasoning templates used in the training set are shown in Figure 2. The full collection and ChatGPT prompts can be found on our project page.

Additionally, we introduce two dataset variants for the ablation study: `Numbers Only` and `With Prompts`. Neither of these variants includes the reasoning phase in their target texts, as illustrated in Figure 3.

- `Numbers Only`: The target texts contain solely textual grasp poses p in this variant. While this approach is straightforward for adapting a language model to numerical prediction, its performance is unsatisfying.
- `With Prompts`: This variant enriches the textual grasp poses with added prompts. For instance, it specifies that (x,y) denotes center point coordinates, and θ indicates rotation angles. Although this modification led to an improvement in performance compared to `Numbers Only`, it remains suboptimal.

As for the input texts in our VLM dataset, we also employ ChatGPT to automatically generate a series of consistent instruction templates pertaining to the robotic grasping task, and an example template is shown in Figure 3. Notably, the methodology behind creating this Reasoning Tuning VLM dataset is adaptable to other numerical prediction tasks beyond robotic grasping. Adjusting the strategies in the reasoning phase can draw upon the appropriate prior knowledge embedded within LLMs tailored for different tasks.

4.2 Training Strategy

In our setup, for each image I, we have a single round conversation data form (S,A), where S represents the input instruction and A is the associated target answer. This paper performs two training strategies: pre-training and LoRA fine-tuning, as illustrated in Figure 4. Both strategies utilize an auto-regressive training objective following LLaVA. To elaborate, for a sequence of length l, the probability of producing the target answer A is formulated as

$$p(A|I,S) = \prod_{i=1}^{l} p_{\theta_m}(a_i|I,S,A_{<i}), \tag{1}$$

where θ_m is trainable parameters in the model; a_i represents the current prediction token; $A_{<i}$ indicate answer tokens before the current token a_i. During training, let $A = \{A_r, A_n\}$ represent the texts

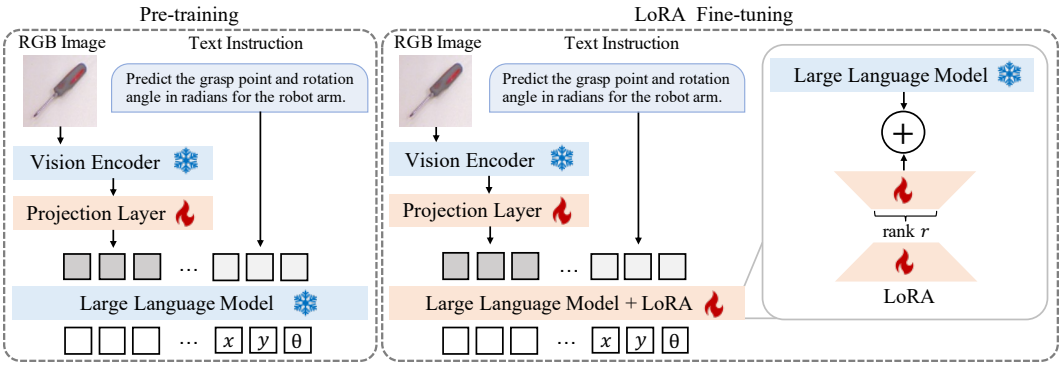

Figure 4: Two training strategies. 1) Pre-training: only parameters of the projection layer are trainable; 2) LoRA fine-tuning: only parameters of the projection layer and LoRA model are trainable.

for the reasoning phase A_r and the numerical prediction A_n. Then the equation (1) can be rewritten as

$$p(A|I,S) = p(A_r|I,S) \cdot p(A_n|I,S,A_r) = \prod_{i=1}^{|A_r|} p_{\theta_m}(a_i|I,S,A_{r_{<i}}) \cdot \prod_{j=1}^{|A_n|} p_{\theta_m}(a_j|I,S,A_r,A_{n_{<j}}) \quad (2)$$

where $p(A_r|I,S)$ denotes the probability of producing the reasoning texts, and $p(A_n|I,S,A_r)$ is the probability of producing numerical predictions conditioned on the input image I, instruction S, and reasoning phase texts A_r. And the entire answer length $l = |A_r| + |A_n|$.

4.2.1 Pre-training

Within this training strategy, both the visual encoder and weights of the LLM are maintained in a frozen state. Only weights of the projection layer, which aligns image features with the word embedding space of the LLM, are updated.

4.2.2 LoRA Fine-tuning

To further enhance the performance, we adopt LoRA [10] fine-tuning, a computationally efficient technique that adds an external model to the existing LLM. Specifically, we inject LoRA into all linear layers within the LLMs. Notably, both the vision encoder and the original LLM remain frozen. Only weights of added LoRA and the projection layer are set as trainable parameters.

5 Experiments

In this section, we assess the performance of the proposed approach using both grasping datasets (Section 5.1) and household test objects on real robots (Section 5.2). Moreover, we have developed three variants of the Reasoning Tuning VLM Grasp dataset for an ablation study. This study underscores the enhanced performance achieved by introducing the reasoning phase.

5.1 Evaluation on Reasoning Tuning VLM grasp datasets

5.1.1 Setup

For all experiments, we utilize LLaVA-7B-v0 [5] as the base model, which is derived from the large language model LLaMA-7B [38]. For the vision encoder, we employ the CLIP ViT-L/14 [25] to extract image features. During the pre-training, we set the batch size to 32 with a learning rate of 2×10^{-3}. During the LoRA fine-tuning, the batch remains 32 and the learning rate is 5×10^{-4}. And we choose a rank $r = 64$ and $\alpha = 32$ for LoRA configurations.

5.1.2 Datasets

We evaluate the proposed method using the Reasoning Tuning VLM Grasp dataset that we developed. This dataset sources its RGB images from the benchmark Cornell Grasp dataset [14], which consists of 885 color images representing 240 distinct objects. We've manually divided these objects into 74 different categories, formulating specific grasping strategies for each category as introduced in Section 4.1. Given the relatively limited number of images, we have implemented data augmentation techniques such as image rotation, zooming, and random cropping, by following related studies [9, 41, 42]. Consequently, we have crafted three variants of the Reasoning Tuning VLM grasp dataset, with each containing 76k image-text paired grasp samples. And only positively labeled grasps were included during training.

5.1.3 Evaluation metrics

We follow a cross-validation setup as in previous works and partition the datasets into 5 folds. Both image-wise and object-wise splits are utilized for evaluation. Performance is reported using the rectangle metric [14]. And a grasp pose is deemed valid if fulfills the following two conditions: 1) The Intersection over Union (IoU) score between the predicted and target rectangles exceeds 25%. 2) The angular deviation between the orientations of the predicted and target rectangles is less than 30 degrees. This metric requires a grasp rectangle representation, while our method predicts the grasp pose without the width w. Thus, to evaluate the accuracy, we convert the pose p combined with the ground truth w into the rectangle representation.

5.1.4 Results

In Table 1, we present the grasp prediction accuracy of our method on the Reasoning Tuning VLM Grasp dataset, including three variants for an ablation study. Given that our VLM dataset originates from images of the Cornell Grasp dataset, we also include results from traditional grasping algorithms on this dataset. Notably, our method directly generates precise numerical values and is trained on RGB images only, while traditional grasping algorithms output heatmaps which require post-processing to obtain grasp poses and are trained on RGB-D images. As Table 1 shows, our method offers a promising grasping accuracy without extra depth information, demonstrating the potential of multi-modal LLMs in numerical prediction. Moreover, variant `With Reasoning` enhances accuracy by $9-26\%$ across all settings compared to variant `Numbers Only`, highlighting the effectiveness of our proposed Reasoning Tuning method. Our method bridges the gap between planning and action in robotics, coupled with LLM's reasoning ability.

Table 1: Results on grasping datasets.

Method	Modality	Grasp Accu (%)	
		Image-Wise (IW)	Object-Wise (OW)
Traditional grasping algorithms on *Cornell Grasp dataset*			
SAE, struct [15]	RGB-D	73.90	75.60
GG-CNN2 [8]	RGB-D	84.00	82.00
GR-ConvNet [9]	RGB-D	97.70	96.60
Our Pre-training on *RT VLM Grasp dataset* (mean±std)			
Numbers Only	RGB, text	65.70±0.87	61.55±1.32
With Prompts	RGB, text	72.94±2.08	67.04±3.46
With Reasoning	RGB, text	74.41±0.88	72.61±2.78
Our LoRA Fine-tuning on *RT VLM Grasp dataset* (mean±std)			
Numbers Only	RGB, text	58.44±6.04	50.31±14.34
With Prompts	RGB, text	69.15±11.00	67.44±9.99
With Reasoning	RGB, text	**84.05**±0.78	**77.02**±0.93

5.2 Evaluation on real-world experiments

5.2.1 Setup

In the real-world experiments, the traditional method GR-ConvNet [9] and our proposed method were thoroughly assessed. A 7-DoF Franka Emika Panda equipped with a Franka Hand parallel gripper was utilized for grasping. To perceive the object, one Azure Kinect was positioned in front of the robot. A cropped 400×400 RGB image of the workspace was utilized as the input to each method to predict the grasp pose.

5.2.2 Household test objects

A diverse set of 27 household objects (Figure. 5) was selected to examine the grasping performance. These objects exhibited variability in several attributes including category, size, shape, and material. Notably, all of these objects are unseen instances from the training dataset. To ensure a comprehensive analysis, each object was tested under 5 distinct positions and orientations, resulting in 135 grasps in total for each method.

Figure 5: Household test objects for real-world grasping experiments.

5.2.3 Results

Here we draw attention to the performance gap between grasping dataset accuracy (Table 1) and real-world experimental results (Table 2). Specifically, in real-world experiments the traditional method GR-ConvNet achieved an 85.2% success rate, marking a noticeable degradation in performance compared to its results on the Cornell Grasp dataset. This underscores its limitations in effectively reasoning about attributes of unseen objects. On the other hand, our proposed RT-Grasp achieved success rates of 80.0% and 83.7% for the pre-training and LoRA fine-tuning strategy, respectively, aligning closely with its performance on VLM grasping datasets. These results not only demonstrate competitive accuracy in grasping household objects when compared to traditional algorithms but also emphasize the model's ability to generalize its capabilities on unseen objects. It is noteworthy that some of the objects tested here belong to unseen categories not present in the training dataset. Additionally, in this testing, we utilize the initial predictions from RT-Grasp without any subsequent refinement, highlighting the effectiveness of adapting multi-modal LLMs to numerical robotic grasping tasks.

Table 2: Results on real-world evaluation.

Method	Grasp Accu (%)
GR-ConvNet [9]	85.2 (115/135)
RT-Grasp Pre-training	80.0 (108/135)
RT-Grasp LoRA Fine-tuning	83.7 (113/135)

5.3 Interactive refinement and reasoning

In this section, we showcase the proficiency of our model in reasoning, as it outputs novel grasping strategies for unseen categories. In addition, it possesses the flexibility to refine the grasp poses through real-time interaction with users. The case in Figure 1 highlights our model's capacity to generate a novel grasping strategy for an object belonging to a category not encountered during training. Figure 6 (in Appendix) illustrates two cases showing the dynamic interaction between users and the model, enabling adaptable and refinable grasp predictions through multiple rounds of conversation.

6 CONCLUSIONS

This research underscores the potential of LLMs beyond their traditional text-centric domains. The proposed method utilizes the vast prior knowledge of LLMs for numerical predictions. We verified its efficacy on robotic grasping tasks through extensive experiments on both benchmark datasets and real-world scenarios. For future work, we aim to further validate this method by applying it to grasping datasets featuring a broader array of objects, such as the Jacquard dataset [43]. Moreover, the adaptation of multi-modal LLMs for numerical predictions in various other robotic manipulation tasks is also a promising research direction.

Acknowledgments

We would like to express our heartfelt gratitude to our colleagues at Baidu Research, USA for their support and assistance throughout the research process. Their expertise, insights, and assistance were instrumental in shaping the direction of this research. We also would like to extend our sincere appreciation to the reviewers for their invaluable comments and constructive feedback, which greatly contributed to the refinement of this paper.

References

[1] K. Lin, C. Agia, T. Migimatsu, M. Pavone, and J. Bohg. Text2motion: From natural language instructions to feasible plans. *arXiv preprint arXiv:2303.12153*, 2023.

[2] 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.

[3] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

[4] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[5] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[6] 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.

[7] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[8] D. Morrison, P. Corke, and J. Leitner. Learning robust, real-time, reactive robotic grasping. *The International journal of robotics research*, 39(2-3):183–201, 2020.

[9] S. Kumra, S. Joshi, and F. Sahin. Antipodal robotic grasping using generative residual convolutional neural network. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9626–9633. IEEE, 2020.

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[11] J. Maitin-Shepard, M. Cusumano-Towner, J. Lei, and P. Abbeel. Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding. In *2010 IEEE International Conference on Robotics and Automation*, pages 2308–2315. IEEE, 2010.

[12] Y. Domae, H. Okuda, Y. Taguchi, K. Sumi, and T. Hirai. Fast graspability evaluation on single depth maps for bin picking with general grippers. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1997–2004. IEEE, 2014.

[13] M. A. Roa and R. Suárez. Grasp quality measures: review and performance. *Autonomous robots*, 38:65–88, 2015.

[14] Y. Jiang, S. Moseson, and A. Saxena. Efficient grasping from rgbd images: Learning using a new rectangle representation. In *2011 IEEE International conference on robotics and automation*, pages 3304–3311. IEEE, 2011.

[15] I. Lenz, H. Lee, and A. Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.

[16] L. Pinto and A. Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 3406–3413. IEEE, 2016.

[17] 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.

[18] X. Zhu, Y. Zhou, Y. Fan, L. Sun, J. Chen, and M. Tomizuka. Learn to grasp with less supervision: A data-efficient maximum likelihood grasp sampling loss. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 721–727. IEEE, 2022.

[19] 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.

[20] Y. Chen, R. Xu, Y. Lin, and P. A. Vela. A joint network for grasp detection conditioned on natural language commands. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4576–4582. IEEE, 2021.

[21] H. Ito, H. Ichiwara, K. Yamamoto, H. Mori, and T. Ogata. Integrated learning of robot motion and sentences: Real-time prediction of grasping motion and attention based on language instructions. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 5404–5410. IEEE, 2022.

[22] J. Hatori, Y. Kikuchi, S. Kobayashi, K. Takahashi, Y. Tsuboi, Y. Unno, W. Ko, and J. Tan. Interactively picking real-world objects with unconstrained spoken language instructions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3774–3781. IEEE, 2018.

[23] A. B. Rao, K. Krishnan, and H. He. Learning robotic grasping strategy based on natural-language object descriptions. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 882–887. IEEE, 2018.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[25] 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*, pages 8748–8763. PMLR, 2021.

[26] S. Stepputtis, J. Campbell, M. Phielipp, S. Lee, C. Baral, and H. Ben Amor. Language-conditioned imitation learning for robot manipulation tasks. *Advances in Neural Information Processing Systems*, 33:13139–13150, 2020.

[27] K. Zheng, X. Chen, O. C. Jenkins, and X. Wang. Vlmbench: A compositional benchmark for vision-and-language manipulation. *Advances in Neural Information Processing Systems*, 35: 665–678, 2022.

[28] M. Shridhar, L. Manuelli, and D. Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022.

[29] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, and D. Fox. Rvt: Robotic view transformer for 3d object manipulation. *arXiv preprint arXiv:2306.14896*, 2023.

[30] Y. J. Ma, W. Liang, V. Som, V. Kumar, A. Zhang, O. Bastani, and D. Jayaraman. Liv: Language-image representations and rewards for robotic control. *arXiv preprint arXiv:2306.00958*, 2023.

[31] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[32] C. Jin, W. Tan, J. Yang, B. Liu, R. Song, L. Wang, and J. Fu. Alphablock: Embodied finetuning for vision-language reasoning in robot manipulation. *arXiv preprint arXiv:2305.18898*, 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. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE, 2023.

[34] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor. Chatgpt for robotics: Design principles and model abilities. *Microsoft Auton. Syst. Robot. Res*, 2:20, 2023.

[35] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

[36] 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.

[37] S. Kumra and C. Kanan. Robotic grasp detection using deep convolutional neural networks. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 769–776. IEEE, 2017.

[38] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[39] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[40] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[41] S. Kumra, S. Joshi, and F. Sahin. Gr-convnet v2: A real-time multi-grasp detection network for robotic grasping. *Sensors*, 22(16):6208, 2022.

[42] Q. Zhang, J. Zhu, X. Sun, and M. Liu. Htc-grasp: A hybrid transformer-cnn architecture for robotic grasp detection. *Electronics*, 12(6):1505, 2023.

[43] A. Depierre, E. Dellandréa, and L. Chen. Jacquard: A large scale dataset for robotic grasp detection. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3511–3516. IEEE, 2018.

7 Appendix

7.1 Interactive refinement and reasoning

We illustrate two test cases on unseen categories (Figure 6) to show the proficiency of our model in reasoning. Predicted grasping results can be refined according to real-world scenarios via dynamic interaction between users and the model.

For the case of the screwdriver, initially, our model outputs a grasp point located on the metallic part of the screwdriver, which is considered a valid grasp for the robot arm based on the training dataset. However, recognizing the potential insecurity of this grasp, the user instructs the model to refine its grasping prediction, specifically to avoid the metallic tip. Leveraging its reasoning ability, our model successfully generates a distinct grasp pose in response to the user's new instruction.

In the case of grasping a cup, the model initially suggests a grasp pose on the handle. Nevertheless, this pose may be unstable when grippers attempt to grasp the cup from the top. Responding to the user's guidance, the model refines the grasp pose by identifying the cup's edge, resulting in a more stable grasp when approached from the top.

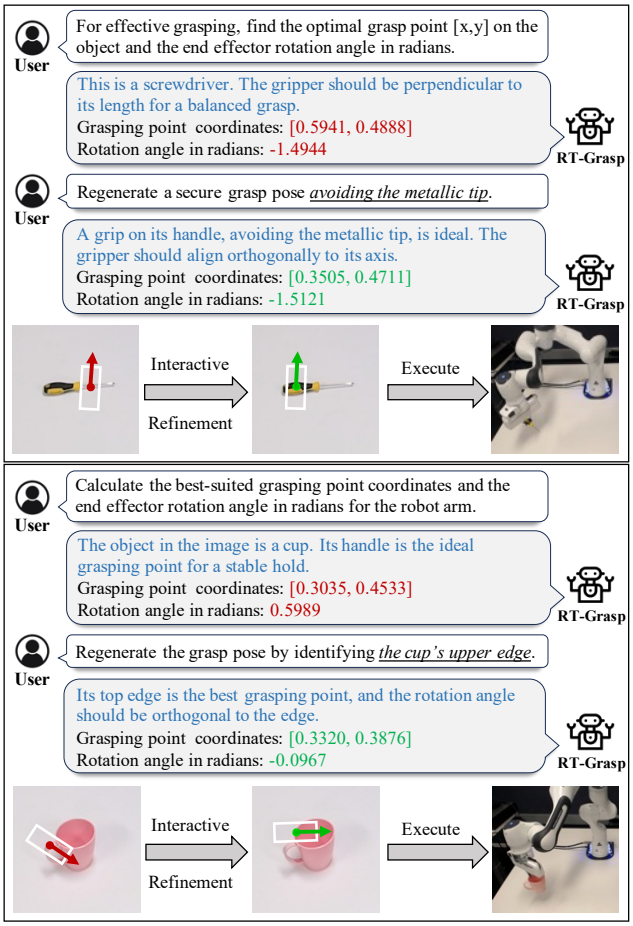

Figure 6: Reasoning and interactive refinement. Outputs from RT-Grasp include a reasoning phase (in blue) and a numerical grasp pose. The initial predicted grasp is indicated in red, while the grasp after refinement is denoted in green.

7.2 Validation accuracy

We present a visualization of the validation accuracy along with the training steps for the `With Reasoning` variant in Figure 7, which encompasses both pre-training and LoRA fine-tuning strategies. Our aim is to offer a clear and insightful visualization of the validation accuracy dynamics throughout the training process, shedding light on the effectiveness of our model. The consistent ascent in validation accuracy serves as a testament to the robustness and adaptability of our model.

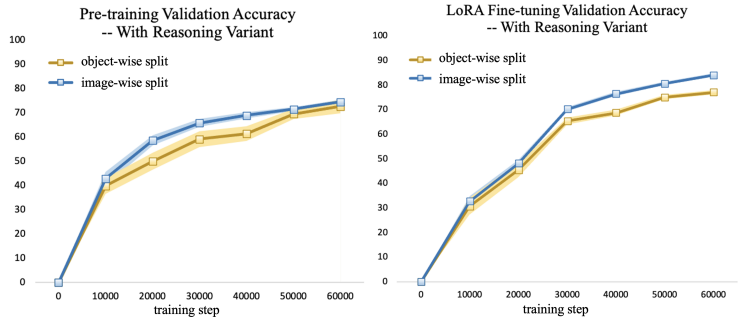

Figure 7: Validation accuracy for two training strategies on With Reasoning.

7.3 Reasoning template examples

We present a selection of reasoning templates employed during the training of our model. These reasoning templates were meticulously generated through a multi-step pipeline outlined in Section 4, initially created by ChatGPT, and subsequently subjected to rigorous manual verification. It is crucial to emphasize that these reasoning templates were tailored to specific categories, as they may exhibit distinct geometric properties and characteristics. The entire collection can be found on our project page.

Category: <*banana*>

Templates: -- There is a banana in the image. Given its elongated and curved nature, grasping near the middle offers a secure grip. The gripper rotation should be orthogonal to the curve direction.

-- Identified object in the image is a banana. Its unique curved shape suggests grasping it in a way that aligns with its length for stability. The ideal gripper orientation would be perpendicular to its longitudinal curve.

Category: <*bottle*>

Templates: -- There is a bottle in the near center of the image, usually cylindrical with a narrow neck. The midsection offers the most stability for a grip. The gripper should align with the bottle's axis.

-- Recognized object in the image is a bottle. Its cylindrical shape suggests that a grip around its body is most stable. the grippers rotation should be perpendicular to its length.

Category: <*toothbrush*>

Templates: -- The object is a toothbrush, long with bristles on one end. Grasping the handle, away from the bristles, ensures a firm grip. The gripper should be perpendicular to its length.

-- This is a toothbrush. The handle, avoiding the bristled end, is the best grasping point. The gripper should align orthogonally to its main axis.

Category: <*lime*>

Templates: -- The object is a lime, typically spherical. A grasp near the center would be stable.

-- This is a lime. Its rounded nature means targeting the midpoint provides a balanced grip.

Figure 8: Examples of reasoning templates.