2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

RT-Grasp: Reasoning Tuning Robotic Grasping via Multi-modal Large Language Model

Jinxuan Xu[1*], Shiyu Jin[2], Yutian Lei[2], Yuqian Zhang[1] and Liangjun Zhang[2]

***Abstract*— Recent advances in Large Language Models (LLMs) have showcased their remarkable reasoning capabilities, making them influential across various fields. However, in robotics, their use has primarily been limited to manipulation planning tasks due to their inherent textual output. This paper addresses this limitation by investigating the potential of adopting the reasoning ability of LLMs for generating numerical predictions in robotics tasks, specifically for robotic grasping. We propose Reasoning Tuning, a novel method that integrates a reasoning phase before prediction during training, leveraging the extensive prior knowledge and advanced reasoning abilities of LLMs. This approach enables LLMs, notably with multi-modal capabilities, to generate accurate numerical outputs like grasp poses that are context-aware and adaptable through conversations. Additionally, we present the Reasoning Tuning VLM Grasp dataset, carefully curated to facilitate the adaptation of LLMs to robotic grasping. Extensive validation on both grasping datasets and real-world experiments underscores the adaptability of multi-modal LLMs for numerical prediction tasks in robotics. This not only expands their applicability but also bridges the gap between text-based planning and direct robot control, thereby maximizing the potential of LLMs in robotics. Our dataset will be released. More details and videos of this work are available on our project page: https://sites.google.com/view/rt-grasp.**

I. INTRODUCTION

The growth of artificial intelligence in recent years has been significantly driven by the emergence of large language models (LLMs). These models, packed with vast knowledge and advanced reasoning ability, have revolutionized our approach to various tasks, especially those involving language processing. In robotics, LLMs play a crucial role in facilitating direct interactions between robots and humans. For instance, in tasks such as robot manipulation planning, many studies [1], [2], [3] have utilized LLMs to interpret natural language commands from users and translate them into feasible multi-step plans for robots. However, despite their potential in robotics, LLMs' application has predominantly been limited to such planning tasks. A notable bottleneck lies in the textual nature of LLM outputs, which often pose challenges for tasks requiring precise numerical outputs.

Recently, multi-modal LLMs have expanded LLM capabilities by understanding both text and images. In robotics, they bridge the gap between perception and planning, addressing a variety of embodied reasoning tasks [4], [5]. However, their image understanding lacks precision, for example, they often struggle to accurately determine object locations, despite providing general descriptions. Although models like GPT-4 with vision [6] show promise in tasks like object detection, they encounter difficulties when tasked with making unique numerical predictions, such as grasp poses in robotic grasping (refer to Fig. 1). Another significant challenge in the application of multi-modal LLMs in robotics lies in the instability and verbosity of their textual outputs, which renders them unreliable for tasks requiring precise manipulation. While certain robotic tasks can benefit from the integration of multi-modal LLMs, their capacity for direct numerical prediction remains largely unexplored.

This paper investigates the potential use of multi-modal LLMs in numerical prediction tasks, specifically focusing on the domain of robotic grasping. Robotic grasping, considered one fundamental yet most challenging area in robotics, revolves around the generation of precise grasp poses essential for subsequent robot manipulation.

Traditional robotic grasping methods typically rely on deterministic predictions, which often fail in real-world scenarios due to their lack of reasoning capabilities. Most existing methods [7], [8], using CNN-based architectures, excel in experimental accuracy on benchmark datasets, but struggle in practical applications. For example, these traditional models may produce theoretically correct predictions that prove impractical in execution, as shown in Fig. 1 and labeled as invalid. Such predictions are hard to apply across robot arms due to varying gripper constraints. Additionally, some theoretically correct grasps may result in unsafe actions, such as targeting the sharp ends of screwdrivers during grasping.

Hence, adopting a non-deterministic approach equipped with reasoning ability is crucial. This capability not only allows the model to generate practical grasp poses applicable across various settings but also allows the refinement of predictions based on user commands. 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.

To efficiently utilize the reasoning capability of multi-modal LLMs for numerical predictions, we propose a novel methodology called Reasoning Tuning. This approach introduces a crucial reasoning phase preceding the numerical prediction step during training. The primary objective of this reasoning phase is to encourage the model to ground its predictions in logical reasoning principles. For instance,

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

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

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

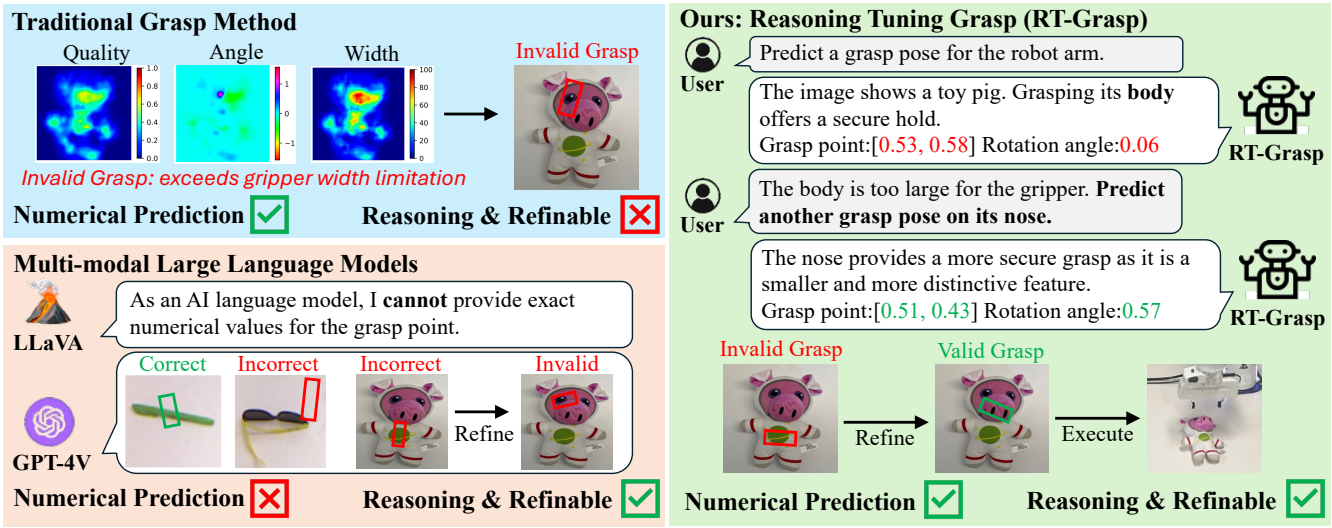

Fig. 1. Comparing three robotic grasping approaches: 1) Traditional CNN-based algorithms produce fixed poses, which lack adaptability in practical situations. 2) Multi-model LLMs output adaptable grasping strategies but lack precise numerical predictions. 3) Ours combines the best of both, predicting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

the model first logically infers attributes such as the object's type, shape, position, and a fundamental grasping principle. Subsequently, the numerical prediction is derived from this reasoning phase. This reasoning phase aims to unlock the valuable information encapsulated within multi-modal LLMs, leveraging their vast knowledge of general object attributes. In this paper, we empirically showcase that fine-tuning multi-modal LLMs with the integration of this reasoning phase enhances their efficacy in generating numerical predictions in robotic grasping.

We investigate two economical training strategies for the proposed Reasoning Tuning: pre-training and Low-Rank Adaptation (LoRA) fine-tuning [9]. 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, our work focuses on adapting multi-modal LLMs for numerical prediction tasks, specifically in the domain of robotic grasping. In contrast to deterministic traditional methods, our approach not only incorporates advanced reasoning capabilities but also introduces a novel paradigm for refining predictions, as illustrated in Fig. 1. The main contributions can be summarized as follows:

- We propose Reasoning Tuning, a novel methodology that utilizes the inherent prior knowledge of pre-trained multi-modal LLMs, facilitating their adaptation to tasks requiring numerical predictions.
- We present our dataset, Reasoning Tuning VLM Grasp dataset, designed for fine-tuning multi-modal LLMs for robotic grasping.
- We empirically validate the proposed method for robotic grasping using two computationally efficient training strategies and conduct real-world hardware experiments. Our results demonstrate its effectiveness and its ability to refine grasping predictions based on user commands.

II. RELATED WORK

A. Robotic Grasping

Traditionally, robotic grasping has heavily relied on analytical approaches [10], [11], [12]. These methods primarily focus on understanding object geometry or analyzing contact forces to determine a grasp that optimizes stability. However, these techniques often struggle to generalize well to unseen objects and can falter when confronted with irregularly shaped items.

In recent years, data-driven methods, particularly those leveraging convolutional neural networks (CNNs), have shown promising results [13], [14], [15], [16], [8], [7], [17]. These approaches leverage extensive datasets of labeled grasping examples to train models capable of predicting grasp poses. Despite their success, these models often suffer from overfitting. They also lack the ability to reason about 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.

B. Language Grounding for Robotics

1) Language-conditioned Robotic Manipulation: In recent years, the integration of natural language into robotic manipulation has garnered significant interest. Studies [18], [19], [20], [21] have explored grasp detection grounded following language instructions within cluttered scenes. [22] performs grasping prediction based on language descriptions of object properties. Building upon advancements in language models [23], [24], recent studies [25], [26], [27], [28], [29], [30] have successfully grounded more flexible language instructions into long-horizon manipulation tasks. However, these approaches often require extensive demonstrations to master image-based policies.

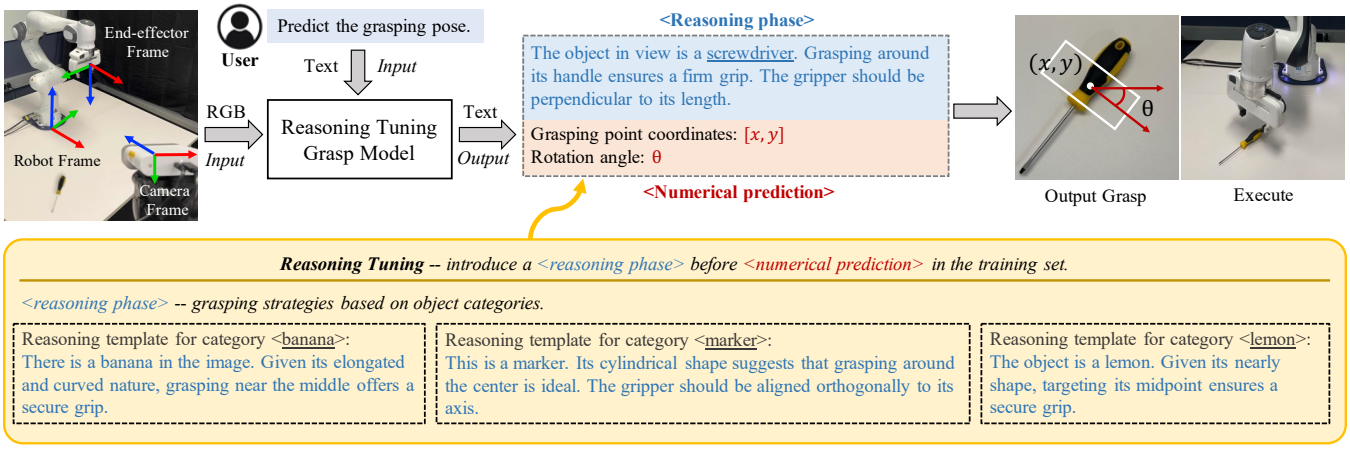

Fig. 2. Overview. The proposed method processes RGB images and user instructions to yield text outputs, which comprise both a reasoning phase and a numerical grasp pose prediction $p = \{x, y, \theta\}$. The reasoning phase analyzes the object's shape and structure based on its category and generates corresponding grasping strategies.

2) LLMs for Robotic Manipulation: With the rise of LLMs, there has been a surge in research exploring their capabilities for robotic manipulation. Many studies [2], [3], [31] have integrated LLMs into closed-loop planning frameworks, decomposing language-conditioned long-horizon tasks into multiple manageable steps. However, bridging the gap between language instructions and actions in robotics remains a challenge. Additionally, some studies [32], [33], [34] have employed program-like specifications to prompt LLMs, melding planning and action using a predefined library of action functions. While intriguing, these methods often face limitations stemming from basic action functions and typically rely on additional perception models, leading to reduced system efficiency and flexibility. Recent studies [5] have made progress in narrowing the planning-action gap by leveraging multi-modal LLMs. However, the method has high data and computational requirements, limiting their feasibility in real-world applications. In contrast, our approach capitalizes on the inherent knowledge embedded within LLMs to achieve precise numerical predictions in the realm of robotics, offering a promising alternative to existing methodologies.

III. 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 and accompanying textual instructions. Similar to [35], [8], 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 pose; θ denotes the rotation angle of the gripper compared to the horizontal axis; w represents the width of the rectangular grasping box, corresponding to the width of the gripper. However, in many studies, the inclusion of w within the predicted grasp pose g is usually considered non-essential [36], due to variations in gripper width limitations.

To this end, our study, with its primary focus on probing the efficacy of LLMs in numerical prediction tasks, assumes w equals the maximum width of the gripper. This paper defines the grasp pose as:

$$p = \{x, y, \theta\}, \tag{1}$$

where (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 illustrated in Fig. 2.

IV. RT-GRASP

In this section, we introduce Reasoning Tuning for robotic grasping (RT-Grasp), a novel method 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 facilitate multi-modal LLMs for numerical prediction by leveraging their extensive encapsulated prior knowledge.

A pre-trained multi-model LLM, such as LLaVA [37], can be directly fine-tuned in a fully supervised manner when given the image and the text instruction. The model is trained by predicting each token in the text output sequentially. The proposed Reasoning Tuning introduces a structured text output, which includes a reasoning phase and a subsequent numerical prediction. We created our image-text dataset for robotic grasping, named Reasoning Tuning VLM (Visual Language Model) Grasp dataset, used for fine-tuning multi-modal LLMs. Additionally, we introduce a method that automatically generates such image-text datasets using GPT-3.5 [38], which can be applied to datasets for tasks beyond robotic grasping. Further details are presented in Section IV-A. Furthermore, we discuss two cost-efficient training strategies employed in our experiments in Section IV-B.

A. Reasoning Tuning

In this section, we introduce Reasoning Tuning, a method that fine-tunes multi-modal LLMs using image-text pairs as inputs and generating structured text outputs. This structured output comprises an initial reasoning phase followed by a subsequent numerical prediction, as illustrated in Fig. 2.

Notably, the entire output is in textual form, and the model is trained to predict corresponding tokens sequentially. By incorporating a reasoning phase at the outset of the output, we encourage the model to generate precise predictions based on logical reasoning specific to the task.

For robotic grasping, we created a new dataset for fine-tuning multi-modal LLMs, called the Reasoning Tuning VLM Grasp dataset. Each data sample includes an RGB image and a text instruction prompting the model to predict the grasp pose (refer to Fig. 3). Additionally, the structured target text in this dataset contains a reasoning phase for the object within the input image, followed by a ground truth grasp pose. The reasoning phase provides a general description of the object, covering aspects such as shape and position, and suggests a corresponding grasping strategy. For instance, consider cups, which may vary in color, design, or material, but a general grasping strategy for them is universal by targeting the handle or the upper edge. Integrating such a reasoning phase guides the model to establish a broad understanding of the object and relevant grasping strategies, thereby facilitating a more informed numerical prediction in subsequent steps.

Existing datasets for robotic grasping typically comprise solely images and numerical ground truth grasp poses. In contrast, our Reasoning Tuning VLM dataset provides image-text pairs tailored specifically for integrating multi-modal LLMs into robotic grasping. In this dataset, images are sourced from the benchmark Cornell Grasp dataset [13], while the accompanying structured texts consist of a reasoning phase followed by ground truth grasp poses presented in textual format. Next, we detail the methodology employed to automatically generate corresponding texts in our dataset.

For the reasoning phase in the structured text, we generated templates based on object categories, as grasping strategies for objects of the same type are usually similar. For each category, we create a series of different reasoning templates. In the structured text of each data sample, one reasoning template is randomly selected based on the object category, followed by the appending of the ground truth grasp pose in textual form (refer to Fig. 3).

To ensure the quality of these reasoning templates, we adopt a multi-step approach. Initially, we prompt GPT-3.5 [38] to generate a collection of templates tailored to each category. Subsequently, we instruct it to refine these

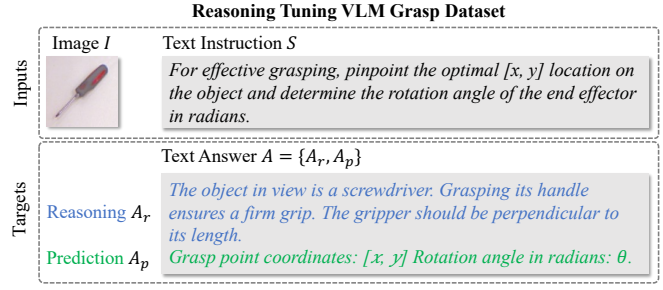

Fig. 3. Illustration of a data sample from the Reasoning Tuning VLM Grasp dataset. The structured text answer contains both a reasoning phase and the ground truth grasp pose.

"object category": ["Description of the object type and general shape. Grasping strategy.", ...]

reasoning_templates = {

"baseball": ["The image shows a baseball, which is spherical. Grasping it as its center will ensure optimal balance." ,
"The object is a baseball. Its round shape requires a center grip for stability." , ...] ,

"cup": ["The object is a cup, which is generally cylindrical. An secure way to grasp a cup is by its edge from the top and the rotation angle should be such that the gripper is orthogonal to the edge of the cup." ,
"Recognized object is a cup. Its best grasped by targeting its top edge, ensuring the gripper is perpendicular to its circular opening." , ...] ,

"sunglasses": ["The object is a pair of sunglasses, with lenses and a frame. Grasping the frame, away from the lenses, is safest." ,
"Recognized sunglasses. It's essential to grasp its frame while avoiding the delicate lenses." , ...] ,

... }

Fig. 4. Examples of reasoning templates within the Reasoning Tuning VLM dataset.

drafts, removing redundant or irrelevant sentences. Finally, as a quality check, we manually verify the correctness and relevance of the generated templates. These reasoning templates typically describe the shape of the object and offer a general grasping strategy. We present some examples of reasoning templates in Fig. 4, and the full collection and GPT-3.5 prompts can be found on our project page.

For the input text instruction in our dataset, we also employ GPT-3.5 to generate a series of consistent instruction templates pertaining to the robotic grasping task, and an example template is presented in Fig. 3. Notably, the methodology behind creating this image-text 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.

B. Training Strategy

In our dataset, 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 Fig. 5. Both strategies utilize an auto-regressive training objective following LLaVA [37]. To elaborate, for a sequence of length l, the probability of producing the target answer A is formulated as

$$p(A|I,S) = \prod_{i=1}^{l} p_{\theta_m}(a_i|I,S,A_{<i}), \tag{2}$$

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.

In our Reasoning Tuning VLM dataset, we define the target answer $A = \{A_r, A_p\}$, which consists of two phases: A_r representing the texts for the reasoning phase, and A_p

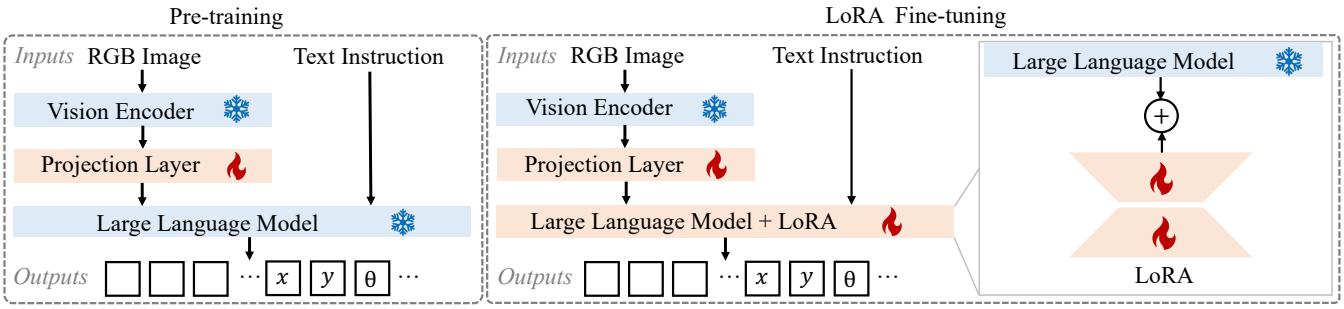

Fig. 5. Two training strategies. 1) Pre-training: only parameters of the projection layer are trainable; 2) LoRA fine-tuning: only parameters of the projection layer and LoRA model are trainable.

indicating the predictions of grasp poses including coordinates $[x, y]$ and rotation angle θ. It is worth noting that since these numerical predictions A_p essentially are also in text format, they are generated as tokens first by LLMs and then converted to textual numbers. Then Equation 2 can be rewritten as

$$\begin{aligned} p(A|I,S) &= p(A_r|I,S) \cdot p(A_p|I,S,A_r) \\ &= \prod_{i=1}^{|A_r|} p_{\theta_m}(a_i|I,S,A_{r_{<i}}) \cdot \prod_{j=1}^{|A_p|} p_{\theta_m}(a_j|I,S,A_r,A_{p_{<j}}) \end{aligned} \tag{3}$$

where $p(A_r|I,S)$ denotes the probability of producing the reasoning texts, and $p(A_p|I,S,A_r)$ is the probability of producing grasp pose predictions conditioned on the input image I, instruction S, and reasoning phase texts A_r. The total length of the entire textual answer A is $l = |A_r| + |A_p|$.

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.

2) LoRA Fine-tuning: To further enhance the performance, we adopt LoRA [9] 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.

V. Experiments

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

A. Evaluation on Reasoning Tuning VLM Grasp datasets

1) Setup: For all experiments, we utilize LLaVA-7B-v0 [37] as the base model, which is derived from the large language model LLaMA-7B [39]. For the vision encoder, we employ the CLIP ViT-L/14 [24] 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 size remains 32 and the learning rate is 5×10^{-4}. And we choose a rank $r = 64$ and $\alpha = 32$ for LoRA configurations.

2) Datasets: We evaluate the proposed method using our Reasoning Tuning VLM Grasp dataset. This dataset sources its RGB images from the benchmark Cornell Grasp dataset [13], which consists of 885 images representing 240 distinct objects. We have manually divided these objects into 74 different categories, formulating specific grasping strategies for each category as introduced in Section IV-A. 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 [8], [40], [41]. Consequently, we have 76k image-text paired grasp samples, and only positively labeled grasps were included.

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 [13]. A grasp pose is deemed valid if fulfills the following two conditions:

- The Intersection over Union (IoU) score between the predicted and target rectangles exceeds 25%.
- 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 grasp pose p combined with the ground truth w into the rectangle representation.

4) Ablation studies: To demonstrate the effectiveness of integrating the reasoning phase, we have developed two variants of the structured text answer in our Reasoning Tuning VLM Grasp dataset. Neither variant includes the reasoning phase in their text answers, as shown in Fig. 6. These variants maintain the image and input text instructions identical to those in the Reasoning Tuning VLM Grasp dataset, differing only in the text answers:

- `No Reasoning-A`: This variant solely contains the textual grasp pose p.

- `No Reasoning-B`: This variant enhances the textual grasp pose with simple prompts. For instance, it specifies that (x, y) denotes the center point coordinates, and θ indicates the rotation angle.

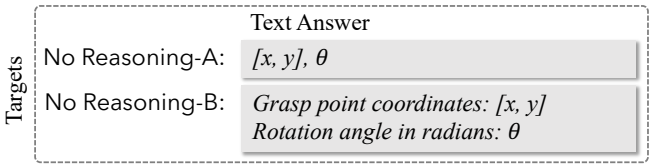

Fig. 6. Illustration of two variants in ablation studies.

5) Results: In Table I, we present the grasp prediction accuracy of two training strategies on our Reasoning Tuning VLM Grasp dataset. Given that our dataset originates from images of the Cornell Grasp dataset, we also include results from traditional grasping algorithms applied to this dataset. Notably, our reasoning-tuned model directly generates precise numerical values, whereas the traditional grasping method outputs heatmaps requiring post-processing to derive grasp poses.

Our primary objective is to develop a method leveraging the reasoning ability of multi-modal LLMs for numerical robotic tasks. As presented in Table I, our method demonstrates promising grasping accuracy by incorporating the reasoning phase, underscoring the potential of multi-modal LLMs in numerical prediction.

TABLE I

RESULTS ON GRASPING DATASETS. *GR-ConvNet results are from [8].*

Method	Grasp Accuracy (%)	
	Image-Wise (IW)	Object-Wise (OW)
Deterministic traditional method on Cornell Grasp dataset [13]		
GR-ConvNet [8]	96.60	95.50
Pre-training on RT VLM Grasp dataset (mean±std)		
No Reasoning-A	65.70±0.87	61.55±1.32
No Reasoning-B	72.94±2.08	67.04±3.46
Ours (RT-Grasp)	74.41±0.88	72.61±2.78
LoRA Fine-tuning on RT VLM Grasp dataset (mean±std)		
No Reasoning-A	58.44±6.04	50.31±14.34
No Reasoning-B	69.15±11.00	67.44±9.99
Ours (RT-Grasp)	**84.05**±0.78	**77.02**±0.93

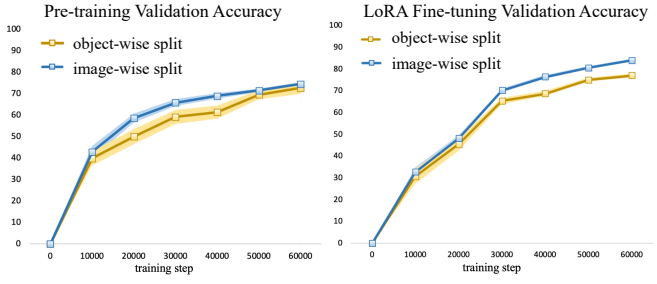

Fig. 7. Validation accuracy of our method (RT-Grasp) for two training strategies.

Furthermore, the performance of `No Reasoning-A`, representing a straightforward adaptation of a language model to numerical prediction, is subpar. While the other variant `No Reasoning-B`, which incorporates simple prompts, shows an improvement over `No Reasoning-A`, its performance remains unsatisfactory. Our reasoning tuning method enhances accuracy by $8.7-26.7\%$ across all settings compared to `No Reasoning-A` and by $1.5-14.9\%$ compared to `No Reasoning-B`, highlighting the effectiveness of integrating the reasoning phase. The validation accuracy curve of our method (RT-Grasp) throughout the training process is illustrated in Fig. 7. Our approach effectively bridges the gap between planning and action in robotics, coupled with multi-modal LLMs' reasoning ability.

B. Evaluation on real-world experiments

1) Setup: In our real-world experiments, we conducted a comprehensive evaluation of both the traditional method GR-ConvNet [8] and our proposed approaches. We utilized a 7-DoF Franka Emika Panda equipped with a Franka Hand parallel gripper along with one Azure Kinect camera for grasping tasks. Visual input for all methods was a 400×400 image crop of the workspace.

2) Household test objects: Fig. 8 displays a diverse set of 27 household objects selected for real-world experiments. Notably, all objects were unseen during training. Each object underwent testing in 5 distinct positions and orientations in our experiments, totaling 135 grasp tests for each method.

Fig. 8. Household test objects for real-world grasping experiments.

3) Results: We present grasp accuracy for both the traditional grasping method and our methods trained using two different strategies in Table II. The results showcase the competitive performance of our approach in real-world experiments compared to traditional methods. Notably, our method (RT-Grasp Pre-training) performs even better in real-world scenarios than on the grasping dataset, highlighting the effectiveness of our approach in adapting to unseen objects. Additionally, in Fig. 9, we present grasp success rates for 27 household test objects, illustrating that our methods successfully grasp all objects at least once during testing. This demonstrates the robustness of our approach across various object shapes.

Here we draw attention to the performance gap between grasping dataset accuracy (Table I) and real-world experimental results (Table II). In real-world experiments, GR-ConvNet achieved an 85.19% success rate, marking a noticeable degradation compared to its performance on the Cornell

TABLE II

RESULTS ON REAL-WORLD EVALUATION.

Method	Grasp Accuracy (%)	
GR-ConvNet [8]	85.19	(115/135)
RT-Grasp Pre-training	80.00	(108/135)
RT-Grasp LoRA Fine-tuning	83.70	(113/135)

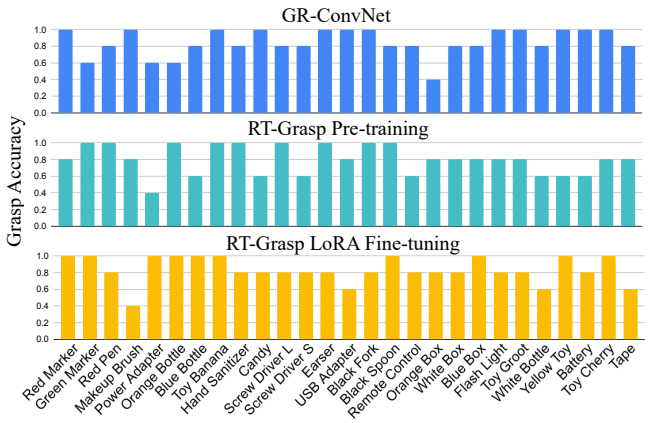

Fig. 9. Grasp Accuracy on household test objects.

Grasp dataset. This underscores its limitations in effectively reasoning about unseen object attributes. Conversely, our proposed RT-Grasp achieved success rates of 80.00% and 83.70% for two training strategies respectively, closely aligning with its performance on VLM grasping datasets. These results demonstrate competitive accuracy in grasping household objects and emphasize the model's ability to generalize to unseen objects, including those from unseen categories such as unique toys. 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.

C. Interactive refinement and reasoning

In contrast to deterministic traditional methods in robotic grasping, one notable advantage of adopting multi-modal LLMs is their capacity for the refinement of numerical outputs based on user instructions. Traditional CNN-based methods typically generate predictions solely based on input images, lacking the flexibility for refinement, thus limiting their applicability in real-world scenarios. However, our approach, enhanced with reasoning capabilities, offers the flexibility to generate different predictions in response to user commands.

In this section, we demonstrate the effectiveness of our model in reasoning by showcasing its ability to propose novel grasping strategies for unseen object categories. Additionally, our model exhibits the flexibility to refine grasp poses through real-time interaction with users. The case depicted in Fig. 1 exemplifies our model's proficiency in generating innovative grasping strategies for objects from categories not encountered during training. Furthermore, Fig. 10 illustrates two instances of dynamic interaction between users and the

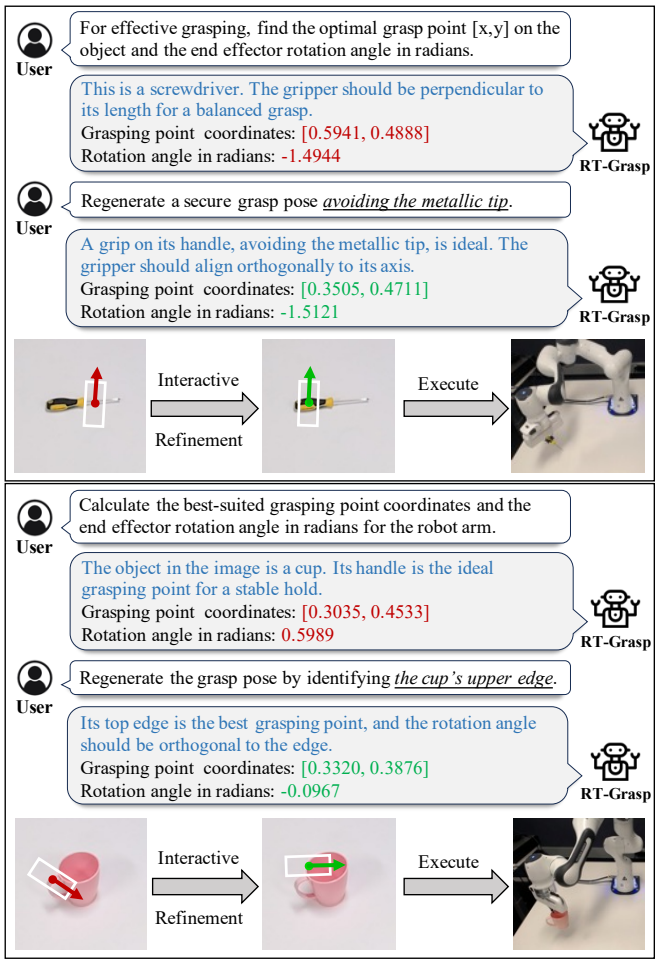

Fig. 10. Reasoning and interactive refinement. Outputs from RT-Grasp include a reasoning phase (in blue) and a numerical grasp pose. The initial predicted grasp is indicated in red, while the grasp after refinement is denoted in green.

model, showcasing the model's ability to provide adaptable and refinable grasp predictions through multiple rounds of conversation.

VI. CONCLUSIONS

This research underscores the potential of LLMs beyond their conventional text-centric applications. Our proposed method utilizes the extensive prior knowledge of LLMs for numerical predictions, specifically in robotic grasping. Through comprehensive experiments conducted on both benchmark datasets and real-world scenarios, we have demonstrated the efficacy of our approach. For future work, we plan to extend the validation of our method by applying it to grasping datasets featuring a broader array of objects, such as the Jacquard dataset [42]. Moreover, the adaptation of multi-modal LLMs for numerical predictions in other robotic manipulation tasks is also a promising research direction.

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. 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.

[5] 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.

[6] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[7] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *The International journal of robotics research*, vol. 39, no. 2-3, pp. 183–201, 2020.

[8] 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)*. IEEE, 2020, pp. 9626–9633.

[9] 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.

[10] 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*. IEEE, 2010, pp. 2308–2315.

[11] 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)*. IEEE, 2014, pp. 1997–2004.

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

[13] 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*. IEEE, 2011, pp. 3304–3311.

[14] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.

[15] 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)*. IEEE, 2016, pp. 3406–3413.

[16] 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.

[17] 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)*. IEEE, 2022, pp. 721–727.

[18] 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.

[19] 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)*. IEEE, 2021, pp. 4576–4582.

[20] 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)*. IEEE, 2022, pp. 5404–5410.

[21] 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)*. IEEE, 2018, pp. 3774–3781.

[22] 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)*. IEEE, 2018, pp. 882–887.

[23] 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.

[24] 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*. PMLR, 2021, pp. 8748–8763.

[25] 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*, vol. 33, pp. 13 139–13 150, 2020.

[26] 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*, vol. 35, pp. 665–678, 2022.

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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)*. IEEE, 2023, pp. 11 523–11 530.

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

[34] 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.

[35] 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.

[36] 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)*. IEEE, 2017, pp. 769–776.

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

[38] 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*, vol. 35, pp. 27 730–27 744, 2022.

[39] 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.

[40] S. Kumra, S. Joshi, and F. Sahin, "Gr-convnet v2: A real-time multi-grasp detection network for robotic grasping," *Sensors*, vol. 22, no. 16, p. 6208, 2022.

[41] Q. Zhang, J. Zhu, X. Sun, and M. Liu, "Htc-grasp: A hybrid transformer-cnn architecture for robotic grasp detection," *Electronics*, vol. 12, no. 6, p. 1505, 2023.

[42] 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)*. IEEE, 2018, pp. 3511–3516.