ManipVQA: Injecting Robotic Affordance and Physically Grounded Information into Multi-Modal Large Language Models

Siyuan Huang*[,1,3], Iaroslav Ponomarenko*[,2], Zhengkai Jiang[5], Xiaoqi Li[2], Xiaobin Hu[6]
Peng Gao[1], Hongsheng Li[1,4], Hao Dong[2]
[1]Shanghai AI Laboratory, [2]Peking University, [3]Shanghai Jiao Tong University, [4]CUHK, [5]UCAS, [6]TUM

MOTIVATION: The integration of Multimodal Large Language Models (MLLMs) with robotic systems has significantly enhanced the ability of robots to interpret and act upon natural language instructions. Despite these advancements, conventional MLLMs are typically trained on generic image-text pairs, lacking essential robotics knowledge such as affordances and physical knowledge, which hampers their efficacy in manipulation tasks.

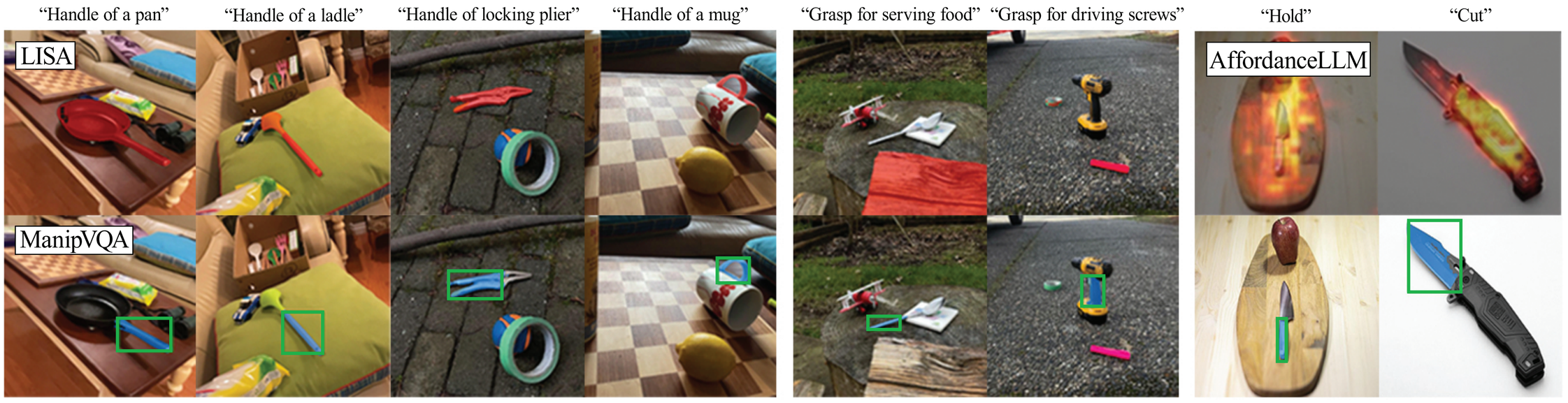
METHOD: To bridge this gap, we introduce ManipVQA, a novel framework designed to endow MLLMs with Manipulation-centric knowledge through a Visual Question-Answering format. This approach not only encompasses tool detection and affordance recognition but also extends to a comprehensive understanding of physical concepts. Our approach starts with collecting a varied set of images displaying interactive objects, which presents a broad range of challenges such as tool object detection, affordance, and physical concept understanding.

The ManipVQA training protocol integrates a pair of principal vision-language tasks: **Referring Expression Comprehension (REC)** and **Referring Expression Generation (REG)**. To advance ManipVQA's proficiency in recognizing robotic affordances and discerning object physical properties, we have augmented the task framework with:

- **REC-Grounding-Affordance**: This task aids the model in identifying functional parts of objects based on their usage descriptions. The objective is to localize these parts without explicitly naming the object or its components.
  REG-Physical: In this task, the model is tasked with identifying an object's physical property using the provided bounding box coordinates. This expands the model's capability to discern various physical attributes of objects within images.

These additional tasks enhance the core **REC** and **REG** tasks, together cultivating a robust skill set tailored for practical robotic deployment. To seamlessly integrate this robotic-specific knowledge with the inherent vision-reasoning capabilities of MLLMs, we adopt a unified VQA format and devise a fine-tuning strategy that preserves the original vision-reasoning abilities while incorporating a new robotic insights.

EXPERIMENTS: Empirical evaluations conducted across various vision task benchmarks demonstrate the robust performance of ManipVQA.



CONCLUSION: This study seeks to reconcile the disparity between the capabilities of existing Multimodal Large Language Models (MLLMs) and the demands of robotic systems. We present ManipVQA, a novel approach designed to equip MLLMs with manipulation-centric knowledge via a visual question-answering paradigm. Our approach involves the collection of a diverse set of images featuring interactive objects, thus encompassing a broad range of challenges related to object detection, affordance, and physical concept prediction. Empirical assessments performed in robotic simulators and across various vision task benchmarks substantiate the efficacy and resilience of ManipVQA. The code and dataset are publicly available at **https://github.com/SiyuanHuang95/ManipVQA**.

This work has been accepted for publication in the Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2024).

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY 北京大學 PEKING UNIVERSITY 香港中文大學 The Chinese University of Hong Kong