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

IROS '24

ABU DHABI

Iaroslav Ponomarenko
Peking University
Center on Frontiers of Computing Studies

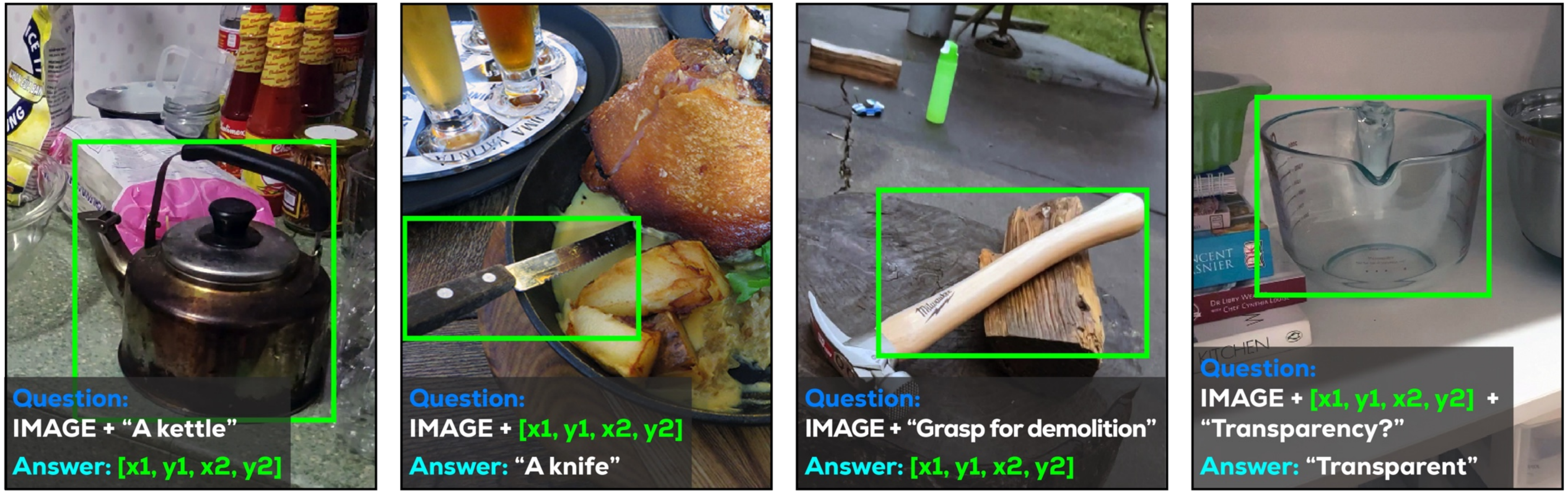
- Multi-Modal Large Language Models (MLLMs)
  - Excel at general vision tasks
  - Face challenges in robotic manipulation tasks

    Struggle to recognize affordances and physical properties of objects

- **ManipVQA** overcomes these limitations by
  - Infusing robotics-specific knowledge into MLLMs

    This empowers MLLMs to understand objects usage and physical properties, enhancing their ability to solve complex manipulation tasks

Unified VQA Format for General Vision and Robotics Specific Tasks



Referring Expression Comprehension (REC)

Referring Expression Generation (REG)

REC-Grounding-Affordance

REG-Physical

General Vision Tasks

Robotics Specific Tasks

Specialized Training on Affordance, Physically Grounded, and General Visual Reasoning Datasets

General Visual Reasoning Datasets

PACO, RefCOCO, and Visual Genome
Rich sources of information on parts and attributes of common objects

Physically Grounded Dataset

PhysObjects Dataset
We use annotations for liquid storage suitability, seal-ability, and transparency

Robotic Affordance Datasets

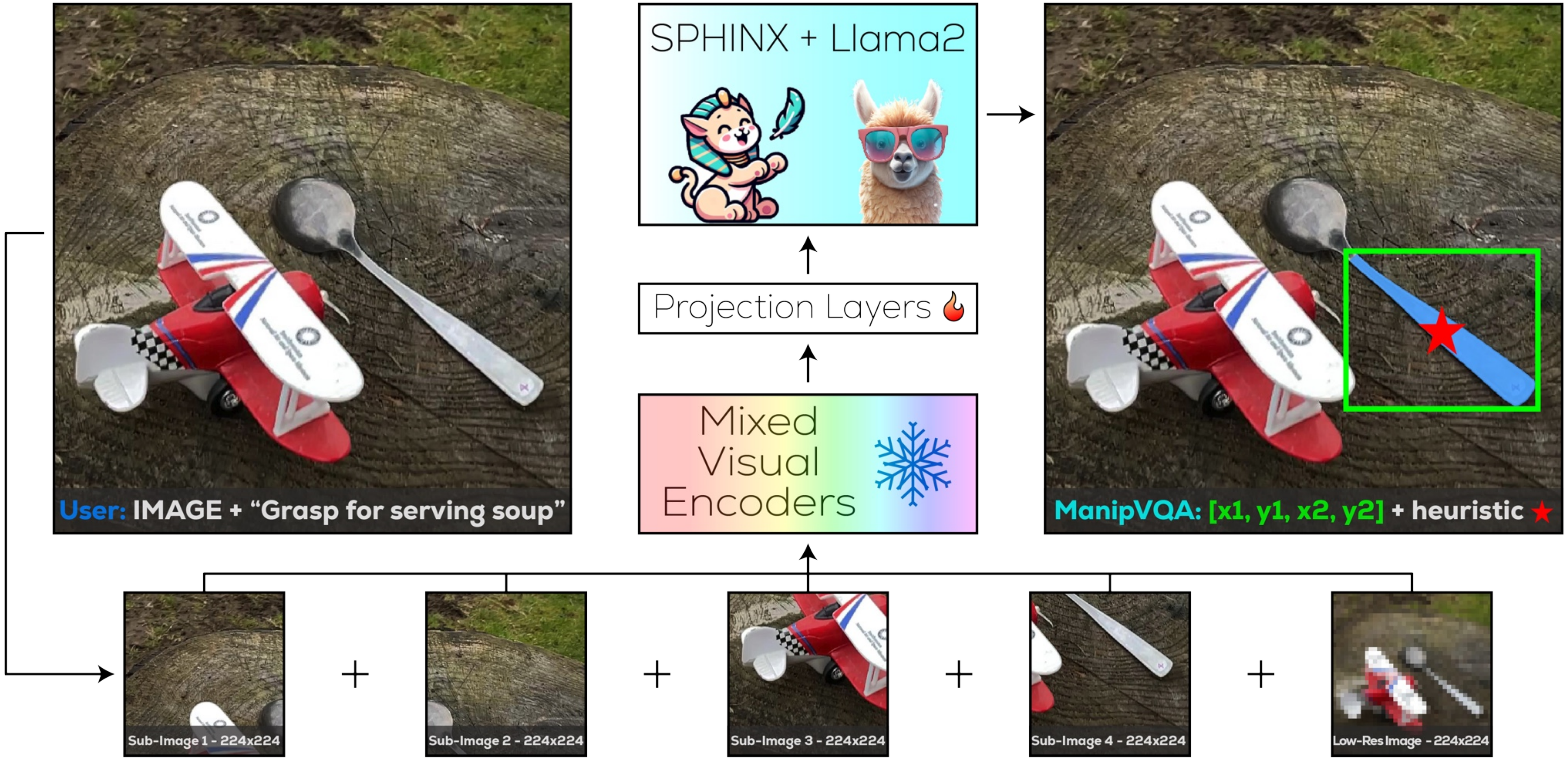
HANDAL Dataset
212 hardware and kitchen tools with annotated handle locations

RGB-D Part Affordance Dataset
105 kitchen, workshop, and garden tools with 7 pre-defined affordances

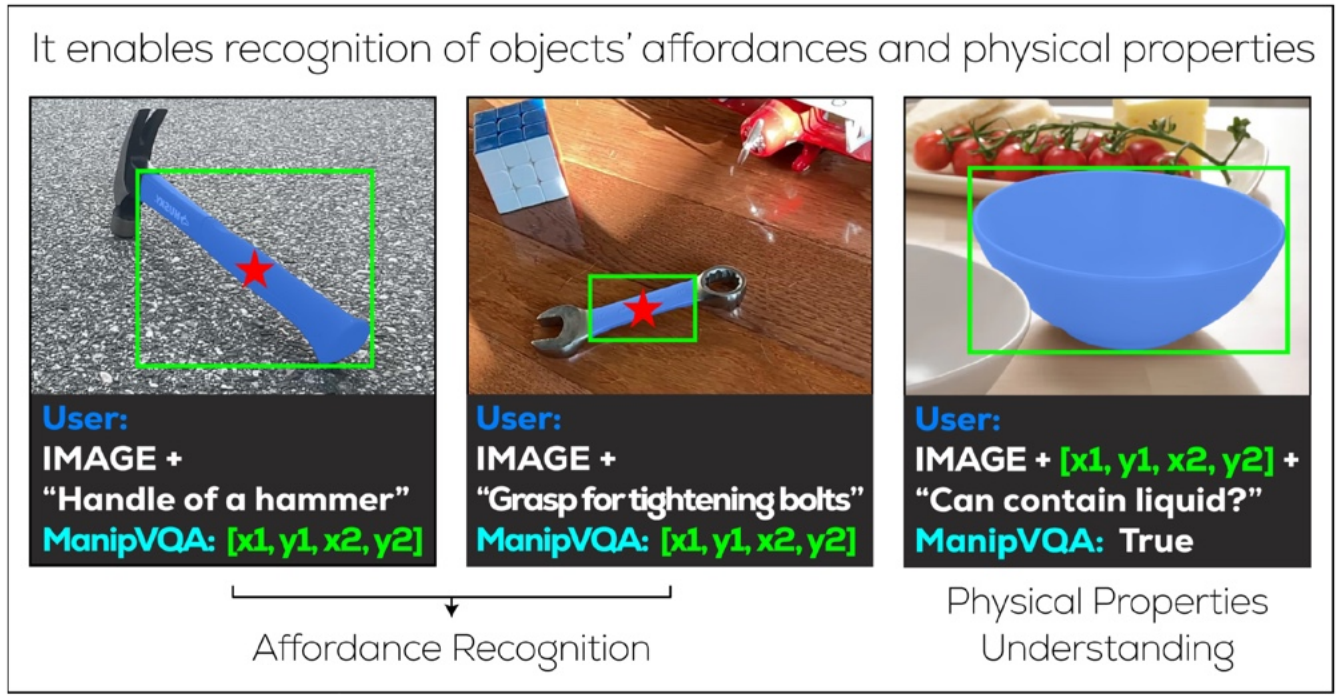
×

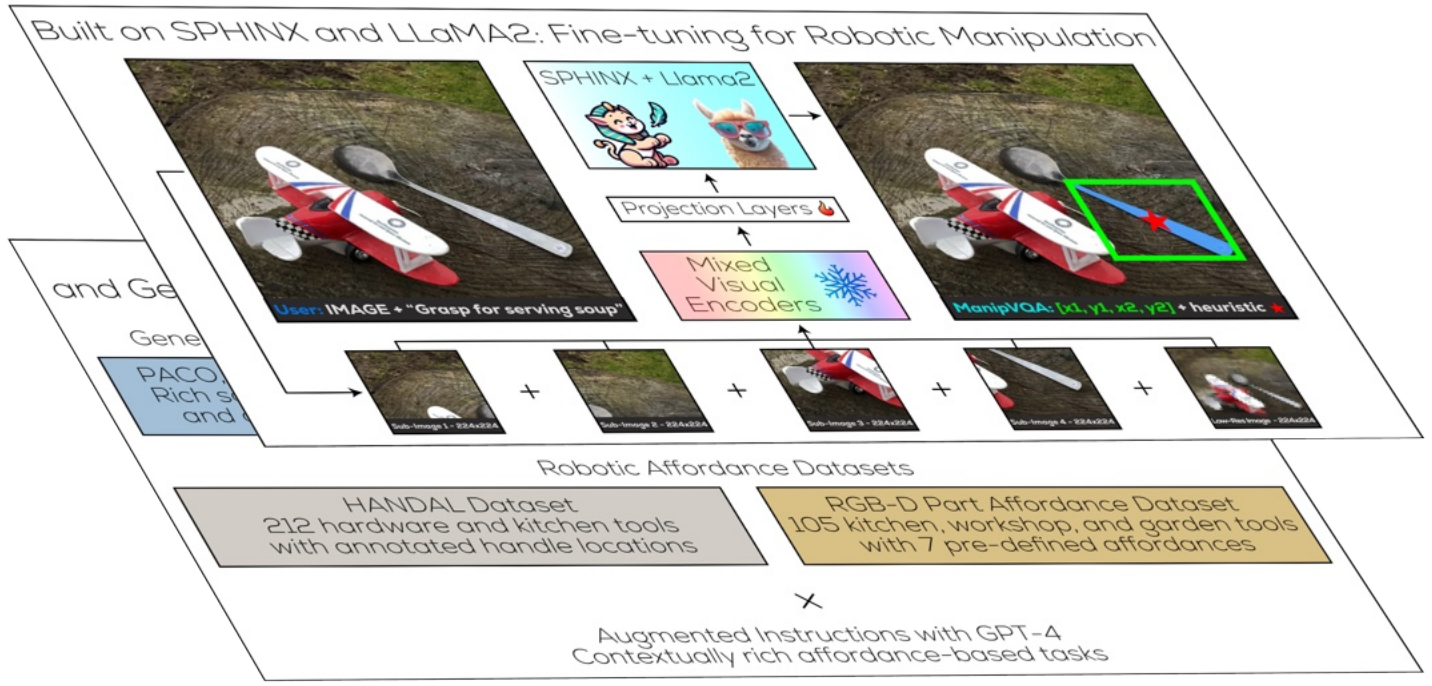
Augmented Instructions with GPT-4
Contextually rich affordance-based tasks

Built on SPHINX and Llama2: Fine-tuning for Robotic Manipulation



- ManipVQA enhances MLLMs for robotics manipulation
  - Expands existing datasets with GPT-4 augmentation
  - Fine-tuning for balance between general vision and manipulation-specific tasks
  - It enables recognition of objects affordances and physical properties





ManipVQA Outperforms Previous Models in Robotic Specific Vision Tasks

Performance Evaluation within the SAPIEN Simulator using PartNet-Mobility Dataset

Method	Training Categories														
Where2Act [38]	0.26	0.36	0.19	0.27	0.23	0.11	0.15	0.47	0.14	0.24	0.12	0.56	0.68	0.07	0.40
FlowBot3D [39]	0.67	0.55	0.20	0.32	0.27	0.31	0.61	0.68	0.15	0.28	0.18	0.21	0.70	0.18	0.26
ManipLLM [4]	**0.68**	0.64	0.36	0.77	0.43	**0.62**	0.65	0.61	**0.65**	0.52	0.40	0.64	**0.71**	**0.60**	**0.64**
Ours	0.67	**0.87**	**0.46**	**0.91**	**0.56**	0.42	**0.69**	**0.79**	0.41	0.53	**0.69**	**1.00**	0.53	0.17	0.58

Method	Training Categories					Testing Categories									AVG
Where2Act [38]	0.13	0.18	0.13	0.40	0.18	0.35	0.38	0.28	0.05	0.21	0.17	0.20	0.15	0.15	0.25
FlowBot3D [39]	0.17	0.53	0.29	0.42	0.23	0.10	0.60	0.39	0.27	0.42	0.28	0.51	0.13	0.23	0.35
ManipLLM [4]	**0.41**	**0.75**	0.44	0.67	0.38	0.22	0.81	**0.86**	**0.38**	**0.85**	0.42	**0.83**	0.26	0.38	0.57
Ours	0.20	0.56	**0.47**	**0.75**	**0.68**	**0.93**	**0.92**	0.82	0.32	0.58	**0.71**	0.81	**0.69**	**0.51**	**0.63**

Our model achieves robust performance without fine-tuning on samples from the simulator

Thank you!

ManipVQA

For more technical details and evaluation results, please refer to the original paper

IROS '24
ABU DHABI