Embodied Question Answering under Generalization: Robotic Affordance with Manipulation Reasoning under Resource Constraints

Nahrin Jannat
nahrinnipun17@gmail.com

Introduction

Embodied AI is an advanced addition to artificial intelligence systems, represented as agents or robots in the physical world, that interact, act, and continuously learn from dynamic environments [1]. Although embodied AI has several tasks, such as visual exploration, navigation, and embodied question answering, it is quite challenging for agents to answer natural language questions. Recently, many researchers have integrated multi-model large language models (MLLMs) and trained the agent with image-text pairs, which noticeably improve the robot's ability to understand and execute natural language instructions [2, 3]. MLLMs showed impressive results on vision generalization tasks in addressing common reasons using pretrained samples, while being insufficient to provide affordance grounding and manipulation due to the absence of low-level action samples.

ManipVQA proposes a novel framework that augments multi-modal large language models with robotic affordance and physically grounded knowledge, while preserving their original reasoning capabilities [4]. The major concern of this work is its reliance on high computational resource availability to load models and train on large-scale datasets. Motivated by this observation, I want to explore embodied question answering in robotics with the affordance grounding and physical reasoning using a vision language model. Additionally, I propose a lightweight research direction with some prototypes to replicate a similar implementation of ManipVQA that demonstrates an affordance-aware EQA in a resource-constrained environment.

Generalized Task Definition

Embodied Question Answering is a generalized embodied AI task in which a robotic agent is initialized at a random location within a real world and is required to answer a natural language question by actively interacting with the environment.

With a question like "Can the robot hold a drill and turn it on?", the agent first needs to understand the concept of the question and search its environment to identify objects to analyze the question. Additionally, depending on the shape, material, and state of each object, the agent must determine what actions (its affordances) it permits, such as grasping, opening, or pouring.

Apart from affordances, reasoning deals with physical properties and constraints such as stability, holding ability, and whether the object is sealed, transparent, or supported. These properties assist the agent in determining whether the action is possible or more interaction is required. When interaction is not enough, the agent must manipulate the environment by moving objects or shifting objects to gather the required information to answer the question.

State-of-the-Art Method

Among recent approaches, ManipVQA [4] represents a state-of-the-art (SOTA) method that enhances MLLMs with affordance understanding and physical reason grounding. The key components of the method are summarized as follows:

- Integrates the SPHINX vision encoder with LLaMA to align the ensemble features with language embedding through projection layers.
- GPT-4V is used to generate physically grounded annotations to learn abstract physical properties.
- Used Physobejects dataset contains transparency, liquid storage, and sealability of an object.
- Used an RGB-D affordance dataset called HANDAL, which contained 7 distinct affordances (grasp, cut, scoop, contain, pour, support, and wrap-grasp).
- Used PACO, COCO, and Visual Genome datasets to provide a rich source of information on parts and the attributes of common objects.
- Used GPT-4 to generate complex instructions by augmentation.
- For a more specific affordance map, employed SAM-HQ for segmentation, then heuristic processing to extract actionable regions.
- Referring Expression Comprehension (REC) is used to predict bounding box coordinates, while Referring Expression Generation (REG) produces descriptive natural language explanations of physical properties.
- The AGD20K dataset (labeled with drink, sit-on) is used to evaluate zero-shot affordance understanding without task-specific fine-tuning.
- Evaluated manipulation using the PartNet-Mobility dataset within the SAPIEN simulator.
- Robotic manipulation executed using CLIPort.

Code Replication under Limited Resources

Full end-to-end replication of ManipVQA requires large-scale embodied simulators, multi-GPU training, and extensive annotated datasets, which is impossible for me to manage and execute code. Here is a functional replication of the SOTA method while understanding the core components of the ManipVQA.

- Using CLIP to align object regions with affordance-related textual prompts with visual reasoning.
- Region grounding using the Segment Anything Model (SAM) to identify candidate object parts relevant to manipulation.
- Physical property reasoning using BLIP-based visual question answering to attributes such as movability and safety.
- Using LLaVA to generate region-specific responses to affordance-related queries for vision language grounding.

All experiments were conducted on Kaggle using pretrained models without additional training, and here you can find the code for the task assignment: https://github.com/nahrin17136/EQA-Affordance-Grounding.

Propose New Ideas

After analyzing ManipVQA and other recent research work, here I propose some new ideas:

- **Lightweight ManipVQA Framework**
  Since ManipVQA depends on large-scale models, a lightweight alternative can be explored using smaller pretrained models such as CLIP, BLIP, SAM, and LLaVA or using other medium sized pretrained models that can operate in low-computational environments. This direction broader the accessibility and faster experimentation, with the possibility of extending to larger models when sufficient resources become available for exploring larger models.

- **Expanded Affordance Training Data**
  ManipVQA is only trained on robotic affordance datasets (HANDAL). In future work, if a feasible training environment is available, the framework can be extended by incorporating the AGD20K dataset and more domain-specific datasets, such as industrial manipulation datasets, to improve generalization across various real-world scenarios.

- **Memory-Augmented Visual Question Answering**
  A memory-based VQA module can be integrated to retain previously inferred affordances and physical properties. This would allow the system to perform multi-step reasoning and maintain contextual consistency across multiple interactions with the same object.

- **Improved Localization and Physical Reasoning**
  The localization limitations observed in LLaMA-based models from insufficient visual physical reasoning capacity. By integrating SoM with GPT-4V, a stronger localization performance can be achieved. In future, we can focus on enhancing region-level physical reasoning to bridge this gap.

Implement one Specific Idea prototype

Among above proposed ideas, I implemented a lightweight prototype of ManipVQA that can operate under limited computational resources. Since the original ManipVQA framework requires loading large-scale vision language models and extensive training, my implementation focuses on replicating its core components using pretrained, publicly available models that are compatible with Kaggle environments.

Specifically, CLIP is used for affordance visual reasoning by aligning object images and candidate regions with affordance-related textual prompts in a zero-shot manner. This enables the system to find high-level manipulation affordances without additional training. To localize actionable regions, the Segment Anything Model (SAM) is used to generate candidate object parts, which are then evaluated using CLIP to associate each region with the most relevant affordance.

For physical property reasoning, a BLIP-based visual question answering model is utilized to collect attributes such as movability, safety, and object usage through natural language queries. This component approximates the physical reasoning module of ManipVQA while remaining computationally efficient. Furthermore, LLaVA is used to perform vision language grounding by generating region-specific responses to affordance-related queries.

All experiments are conducted using pretrained models without fine-tuning. While this implementation does not aim to reproduce the full performance of ManipVQA, it successfully demonstrates affordance reasoning, region grounding, physical property inference, and language-based querying in a unified and lightweight framework. This prototype serves as a scalable foundation for future extensions involving larger models, embodied simulators, and end-to-end training. The implementation of this prototype is available here: https://github.com/nahrin17136/EQA-Affordance-Grounding

Conclusion

This work presents a lightweight baseline of affordance grounding and manipulation reasoning for EQA inspired by ManipVQA. By functionally replicating key components under resource constraints, proposing new research ideas, and implementing a prototype, this study can be useful for further completion as a full research work. Although, I skip the evaluation part beacause training is not possible for resources constraints, the overall outcome highlight that meaningful embodied reasoning can be achieved whilw adding training, and they point toward promising directions for future work on generalizable and resource-efficient embodied AI systems.

References

[1] K. Borazjani, P. Abdisarabshali, F. Nadimi, N. Khosravan, M. Liwang, X. Wang, Y. Hong, and S. Hosseinalipour, "Multi-modal multi-task (m3t) federated foundation models for embodied ai: Potentials and challenges for edge integration," *arXiv preprint arXiv:2505.11191*, 2025.

[2] 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.

[3] D. Liu, R. Zhang, L. Qiu, S. Huang, W. Lin, S. Zhao, S. Geng, Z. Lin, P. Jin, K. Zhang *et al.*, "Sphinx-x: Scaling data and parameters for a family of multi-modal large language models," *arXiv preprint arXiv:2402.05935*, 2024.

[4] S. Huang, I. Ponomarenko, Z. Jiang, X. Li, X. Hu, P. Gao, H. Li, and H. Dong, "Manipvqa: Injecting robotic affordance and physically grounded information into multi-modal large language models," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7580–7587.