# Project Proposal

Contact
wang-zx23@mails.tsinghua.edu.cn

## Team Information

- **Track:**
- **Team Members:**
  - Zixuan Wang 2023011307
  - Member 2
  - Member 3
  - Member 4
  - Member 5

## Idea Justification

blablabla

## Problem Formulation

blablabla

## Literature Survey

1. **Embodied-Reasoner** (Zhang et al.) [1]: Develops a three-stage training pipeline (imitation learning, rejection sampling, and reflection tuning) that significantly outperforms baselines by reducing repetitive searches and improving logical consistency in embodied tasks.
2. **ASKTOACT** (Ramrakhya et al.) [2]: Proposes an RL framework that trains multimodal LLMs to resolve instruction ambiguity through minimal clarification questions, using LLM-generated rewards to eliminate manual reward engineering.
3. **Embodied-R** (Zhao et al.) [3]: Introduces a collaborative framework where large VLMs handle perception while small LMs perform reasoning, trained with RL using logical consistency rewards for spatial reasoning tasks.

**Pipeline Details** Embodied-Reasoner implements a comprehensive three-stage pipeline: initial imitation learning on synthetic trajectories, followed by self-exploration through high-temperature sampling with trajectory selection by a learned reward model, and finally reflection tuning where the model learns to identify and correct its own mistakes. Their training data consists of 9,300 synthesized trajectories containing 64,000 interactive images paired with 90,000 reasoning steps, supplemented by human-annotated test cases. ASK-TO-ACT firstly adapts a multimodal LLM architecture to process sequential visual observations through a Perceiver-based token downsampling method, then employ online RL training where the reward function is automatically generated by another LLM evaluating task progress and question quality. The training data comes from simulated home environments where agents interact with ambiguous object retrieval scenarios. Embodied-R takes a different approach by separating the perception and reasoning components - using a pretrained vision-language model for processing visual inputs while training a smaller language model with RL, where the key innovation is a logical consistency reward that aligns the model's reasoning process with its final answers. They train this system on a relatively small dataset of 5,000 embodied video samples with keyframe extraction to manage computational costs.

**Reward Design and Training Methodology** The ASKTOACT framework uses online reinforcement learning to fine-tune multimodal language models, utilizing LLM-generated sub-goals and question sequences as reward signals. Embodied-R enhances this approach by incorporating logical consistency rewards that evaluate the alignment between reasoning processes and final answers, enabling smaller language models to achieve sophisticated spatial reasoning capabilities.

**Datasets and Evaluation** ASKTOACT utilizes simulated home environments with ambiguous instructions, where rewards are derived from privileged environment states, and is evaluated on object fetching tasks against GPT-4 and supervised MLLMs. Embodied-R trains on 5,000 embodied video samples processed with keyframe extraction for efficiency, with testing conducted across both in-distribution and out-of-distribution spatial reasoning tasks compared to OpenAI and Gemini models. The Embodied-Reasoner system generates 9,300 synthetic trajectories containing 64,000 images and 90,000 reasoning steps, supplemented by 809 human-validated test cases, and demonstrates superior performance in real-world object search tasks when benchmarked against OpenAI and Claude models, particularly in success rates and logical consistency metrics.

## Preliminary Method

blablabla

## Expected Results

blablabla

## References


[1] W. Zhang, M. Wang, G. Liu, *et al.*, "Embodied-reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks," *arXiv preprint arXiv:2503.21696*, 2025.

[2] R. Ramrakhya, M. Chang, X. Puig, R. Desai, Z. Kira, and R. Mottaghi, "Grounding multimodal llms to embodied agents that ask for help with reinforcement learning," *arXiv preprint arXiv:2504.00907*, 2025.

[3] B. Zhao, Z. Wang, J. Fang, *et al.*, "Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning," *arXiv preprint arXiv:2504.12680*, 2025.