Embodied-R: Collaborative Framework for Activating Embodied Spatial Reasoning in Foundation Models via Reinforcement Learning

Baining Zhao[*]
Shenzhen International Graduate
School, Tsinghua University
Pengcheng Laboratory
Shenzhen, China

Ziyou Wang[*]
Northeastern University
Qinhuangdao, China

Jianjie Fang[*]
Northeastern University
Qinhuangdao, China

Chen Gao[†]
BNRist, Tsinghua University
Beijing, China
chgao96@gmail.com

Fanhang Man
Shenzhen International Graduate
School, Tsinghua University
Shenzhen, China

Jinqiang Cui
Pengcheng Laboratory
Shenzhen, China

Xin Wang
Department of Computer Science
and Technology, Tsinghua University
BNRist, Tsinghua University
Beijing, China

Xinlei Chen[†]
Shenzhen International Graduate
School, Tsinghua University
Shenzhen, China
chen.xinlei@sz.tsinghua.edu.cn

Yong Li
Department of Electronic
Engineering, Tsinghua University
BNRist, Tsinghua University
Beijing, China

Wenwu Zhu
Department of Computer Science
and Technology, Tsinghua University
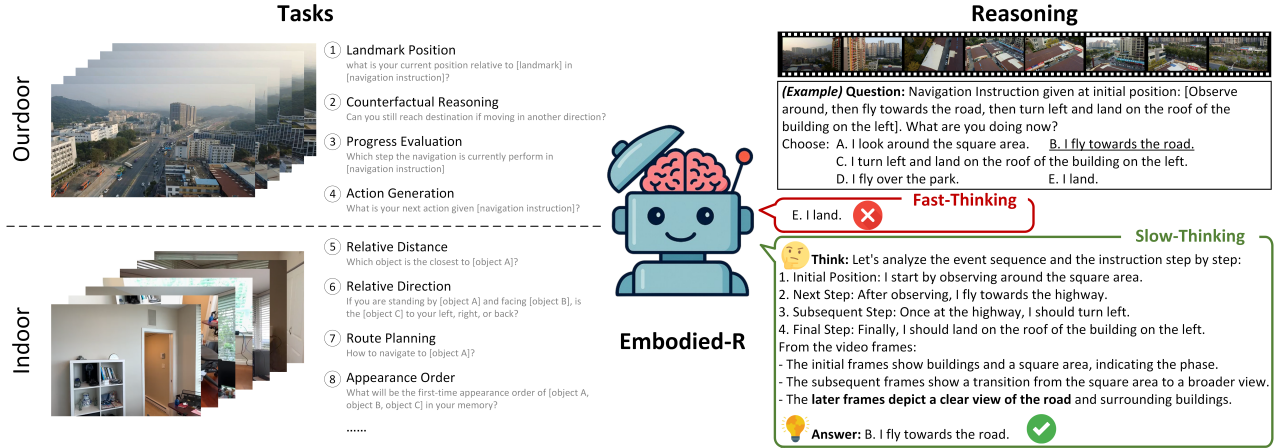BNRist, Tsinghua University
Beijing, China

Figure 1: Embodied spatial reasoning: tasks and thinking process. Challenging tasks from public embodied video datasets are identified, encompassing both indoor and outdoor scenarios. We introduce slow-thinking to improve reasoning performance.

Abstract

Humans can perceive and reason about spatial relationships from sequential visual observations, such as egocentric video streams. However, how pretrained models acquire such abilities, especially high-level reasoning, remains unclear. This paper introduces Embodied-R, a collaborative framework combining large-scale Vision-Language Models (VLMs) for perception and small-scale Language Models (LMs) for reasoning. Using Reinforcement Learning (RL) with a novel reward system considering think-answer logical consistency, the model achieves slow-thinking capabilities with limited computational resources. After training on only 5k embodied video samples, Embodied-R with a 3B LM matches state-of-the-art multimodal reasoning models (OpenAI-o1, Gemini-2.5-pro) on both in-distribution and out-of-distribution embodied spatial reasoning tasks. Embodied-R also exhibits emergent thinking patterns such as systematic analysis and contextual integration. We further explore research questions including response length, training on VLM, strategies for reward design, and differences in model generalization after SFT (Supervised Fine-Tuning) and RL training. The project page is available at: https://embodiedcity.github.io/Embodied-R/

---

[*]Both authors contributed equally to this research.
[†]Corresponding authors: Chen Gao, Xinlei Chen

CCS Concepts

• **Computing methodologies → Artificial intelligence**.

Keywords

Vision-Language Model; Embodied Intelligence; Reasoning

1 Introduction

On the path toward Artificial General Intelligence (AGI) [17], we hope that pre-trained foundation models can not only perform tasks such as dialogue and image understanding in the cyber world [2, 44] but also develop human-like embodied spatial cognition in the three-dimensional physical world, enabling them to perceive, think, and move [4, 32]. The fundamental way humans achieve spatial cognition is through continuous, dynamic visual observations, akin to video streams [26, 30]. For example, by observing their surroundings, humans can infer their position relative to nearby objects. Similarly, based on historical visual observations, humans can determine the actions they should take to reach a target destination.

Visual spatial cognition can be divided into two levels: perception and reasoning [51]. Perception refers to "what is seen", characterized by direct, low-level tasks such as object recognition, edge detection, or color differentiation [52]. Reasoning, on the other hand, involves "what is understood" and "what actions to take", which are indirect and higher-level tasks requiring logical inference and knowledge integration [62]. Examples of reasoning include "Where did I come from?" (e.g., recalling historical movement trajectories [36]), "Where am I?" (e.g., inferring the spatial relationships between nearby objects and distances [5]), and "Where do I want to go?" (e.g., planning actions and deciding movements to reach a destination [8]). While most existing research focuses on improving the perception capabilities of foundation models [6, 11], with notable progress, their spatial reasoning abilities remain limited [9, 58], and methods for enhancement are largely unexplored.

Specifically, video-based spatial reasoning poses several challenges, as follows:

- Reasoning is always built upon perception [19, 32]. For the studied problem, continuous visual observations impose higher demands on perception. Reasoning cannot be well achieved with faulty perceptions or hallucinations [53]. It is challenging to reason when it is already hard to perceive from the videos.
- Video data naturally involves complex spatio-temporal relationships, requiring the discovery of object associations across frames and the extraction of semantics relevant to the reasoning task [16]. For instance, to navigate to a destination outside the current field of view, one must infer their location from historical visual observations, build a mental map of the environment, develop a high-level plan to determine the direction, and finally decide on specific actions to execute. Existing supervised fine-tuning (SFT)

training methods lack supervision for the reasoning process, making it difficult to handle such reasoning tasks [62].
- Embodied visual observations have distinct characteristics. First, understanding disembodied videos, such as movies or TV shows, primarily emphasizes the content within the video, often from a broad and objective perspective [27]. In contrast, egocentric videos focus on understanding the relationship between the observer and the surrounding environment, often from a constrained first-person perspective [22]. Second, embodied continuous visual observations are generated over time, indicating that embodied perception should rely on sequential inputs rather than aggregating all visual observations for a single input after a prolonged period [31]. Finally, due to the continuity of motion in the physical world, egocentric visual observations also exhibit spatial continuity, meaning there is significant redundancy and repetition between frames. Consequently, directly applying existing multimodal large language models (MLLMs) to embodied videos leads to issues, including loss of generalization and input token limits caused by excessive redundant frames [1, 29].

Recently, the impressive performance of OpenAI's o1/o3 [38] and DeepSeek-R1 [24] in solving complex reasoning problems(e.g., mathematics, coding, science, etc.) has drawn attention to reinforcement learning (RL) techniques. By incorporating the chain-of-thought (CoT) reasoning process into post-training, large language models (LLMs) demonstrate a "slow-thinking" mode, where they reason thoroughly before generating responses [45, 55]. Inspired by this, we attempt to introduce "slow thinking" into embodied video-based spatial reasoning tasks, as shown in Figure 1.

This brings a new challenge: the trade-off between model size and computational cost. Existing studies suggest a strong correlation between multimodal understanding/perception capabilities and model size [7, 20, 56]. Since reasoning builds on perception, larger vision-language foundation models should be used as the starting point for training. However, increasing model size leads to often unacceptable computational costs. Additionally, video inputs map to long token sequences, further raising computational demands. Is there a way to leverage the perception capabilities of large-scale models while developing embodied reasoning abilities at a lower computational cost?

Inspired by neuroscience [64], spatial perception and reasoning involve distinct brain regions: visual perception occurs in the visual areas of the occipital lobe [13], basic spatial understanding in the parietal lobe [18], and complex spatial reasoning in the prefrontal cortex [14]. This inspired the design of a collaborative framework with two main components: a large-scale vision-language model (VLM) for perception and a small-scale language model (LM) for reasoning. Based on the continuity of observations, we first propose a key-frame extractor to retain critical information while reducing computational costs. Using a VLM, we sequentially extract semantic information from the frames, which simulates real-world online reasoning while effectively managing the input token length of VLMs for long video inputs. Finally, the semantic information and reasoning question are fed into the small-scale language model, which outputs the reasoning process and final answers. The small-scale language model is trained with RL, where the reward modeling

Embodied-R: Collaborative Framework for Activating Embodied Spatial Reasoning
in Foundation Models via Reinforcement Learning

MM '25, October 27–31, 2025, Dublin, Ireland

not only incorporates rule-based rewards inspired by Deepseek-R1-Zero [24] but, more importantly, introduces a novel reward for the logical consistency of the reasoning process. In the experiments, we explore seven research questions, covering the framework's performance, RL's role in activating embodied spatial reasoning, and out-of-distribution generalization capabilities.

In general, the main contributions of this paper are as follows:

- We propose a **collaborative** framework for large-scale and small-scale foundation models to address spatial reasoning in the video modality. By decoupling perception and reasoning, the framework leverages the perceptual strength of large-scale foundation models while efficiently enhancing the reasoning capabilities of smaller models in a computationally resource-friendly manner.
- This is **the first work to employ reinforcement learning (RL) to enhance the embodied spatial reasoning abilities of foundation models**. Specifically, we introduce a novel **logical consistency reward**, which improves the alignment between reasoning processes and generated answers.
- Our proposed Embodied-R achieves performance **comparable to state-of-the-art multimodal large language models (e.g., OpenAI-o1/Gemini-2.5-Pro)** on both in-distribution and out-of-distribution benchmarks. We further investigate **research questions including the generalization comparison between models trained by SFT & RL, reward design strategies, etc**.

2 Related Work

Large Language Model Reasoning. Recently, enhancing reasoning capabilities has become a key focus in large model technologies, demonstrating remarkable performance on tasks such as mathematical and logical problem-solving [25, 47, 57]. Following the release of OpenAI's o1 [38], numerous studies have proposed various technical approaches to achieve similar functionalities, including Chain-of-Thought (CoT) [54], Monte Carlo Tree Search (MCTS) [23, 60], distillation [35], rejection sampling combined with supervised fine-tuning (SFT) or Direct Preference Optimization (DPO) [40], among others. Furthermore, Deepseek-r1 [24] introduced a method to foster the emergence of reasoning abilities in large language models (LLMs) through rule-based rewards combined with reinforcement learning. Similarly, Kimi k1.5 [45] proposed a comparable approach, presenting various training techniques, such as curriculum learning. This reinforcement learning paradigm has sparked significant interest, with subsequent works successfully reproducing related results [55, 59].

Embodied Spatial Reasoning with VLMs. Inspired by the generality of foundation models across various domains [2, 3], embodied intelligence aims to develop agents that utilize large multimodal models as their "brains" to achieve perception, navigation, and manipulation in the 3D physical world [15, 41]. In terms of input, human visual-spatial perception is more akin to continuous RGB observations, similar to video streams [12, 42], rather than static images [48] or point clouds [52]. Several embodied video benchmarks [58] demonstrate that, while perception tasks are relatively well-addressed, spatial reasoning tasks—such as spatial relationship inference, navigation, and planning—remain highly challenging. However, existing research [16, 43] on video reasoning primarily focuses on disembodied content reasoning, with little emphasis on scenarios involving embodied continuous visual inputs.

Collaboration between large and small models. Existing research primarily focuses on addressing the resource consumption and privacy risks associated with large models, as well as the efficiency and performance advantages of small models in specific scenarios [50]. Small models can assist large models in data selection, prompt optimization, and reasoning enhancement [28, 61]. The use of small models to detect hallucinations and privacy leakage is explored in [49, 63], improving overall system reliability. While our work shares the goal of reducing computational resource demands, it differs by emphasizing the complementary roles of large-scale VLMs in perception and small-scale LMs in enhancing embodied spatial reasoning.

3 The Embodied-R Method

We first define the problem of embodied spatial reasoning. Subsequently, we introduce the VLM-based perception module and the LM-based reasoning module. The collaborative framework is shown in Figure 2.

3.1 Problem Formulation

In the physical world, an agent moves through space, generating a sequence of video frames (continuous visual observations) $\mathbf{f} = [f_0, f_1, \ldots, f_T]$. Suppose a spatial reasoning problem is denoted as q. Our goal is to build a model that takes q and $\mathbf{f}$ as inputs and outputs an answer a. The answer a is considered correct if it is semantically consistent with the ground truth g; otherwise, it is deemed incorrect.

3.2 Large-Scale VLM-based Perception

3.2.1 ***Key-Frame Extractor***. As the agent moves continuously in space, high sampling frequencies result in significant overlap between consecutive frames. On one hand, the VLM relies on changes in the static objects within the environment across frames to infer the agent's pose variation. On the other hand, excessive overlap between frames leads to increased inference costs for both the VLM and LLM. To address this, we designed a key-frame extractor tailored to the characteristics of embodied videos, selecting key frames that retain overlap while ensuring sufficient information gain between them.

The extraction of key-frames is based on the overlap of visual fields caused by motion continuity. When the agent moves forward, the visual content in the latter frame is expected to overlap with a portion of the former frame, and the reverse is true when moving backward. Similarly, during left or right rotations, the latter frame should partially overlap with the former frame in the horizontal direction, and during upward or downward rotations, the overlap occurs in the vertical direction. Given that the sampling frequency of visual observations is typically much higher than the agent's motion speed, frames generally exhibit significant overlap.

Specifically, a perspective transformation is used to model the geometric relationship between frames. Assuming f_t is a key-frame, to determine whether f_{t+1} should also be considered a keyframe, keypoints and descriptors are calculated from f_t and f_{t+1} using the Oriented FAST and Rotated BRIEF (ORB) algorithm. Next, a feature
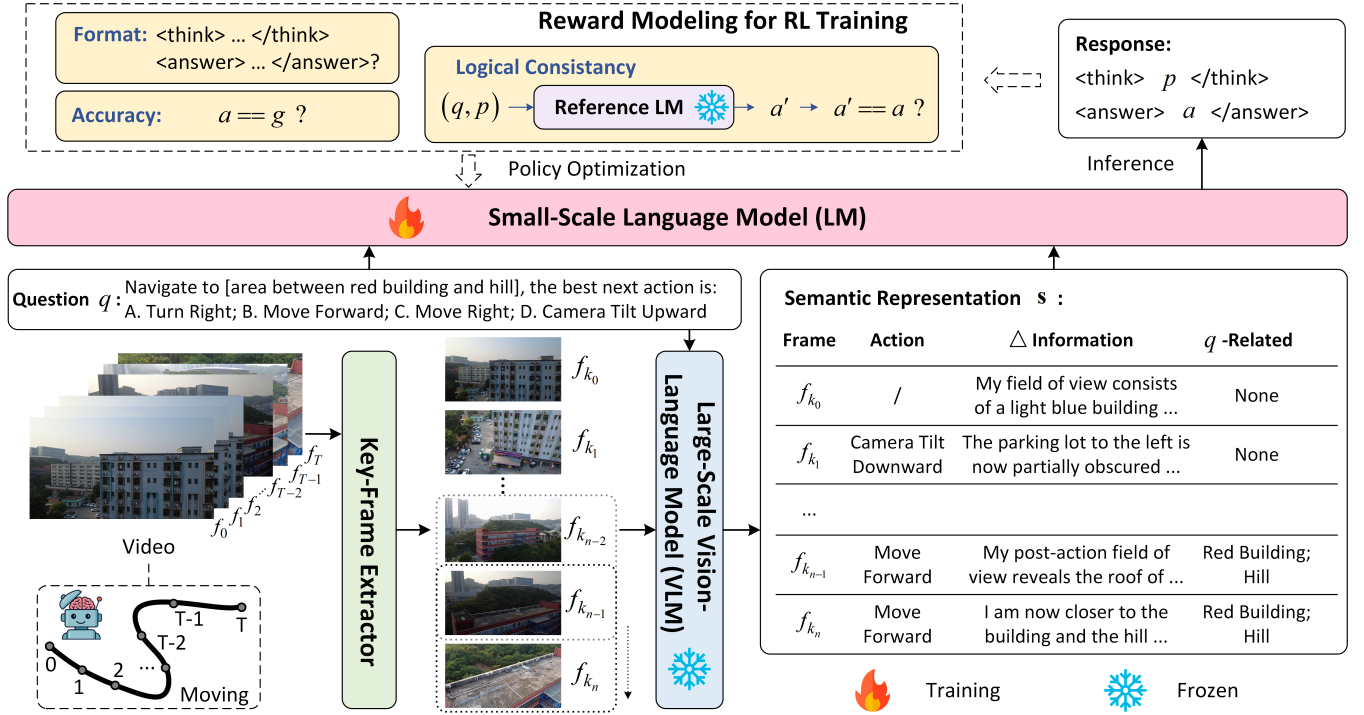
Figure 2: The proposed Embodied-R is a collaborative embodied spatial reasoning framework integrating a Vision-Language Model (VLM) and a Language Model (LM). The separation of perception and reasoning enables us to leverage the perceptual capabilities of large-scale VLMs while training a resource-efficient small-scale LM to activate embodied reasoning through RL. Notably, we introduce a novel logical consistency reward to guide the LM in producing logically coherent reasoning and answer.

matching algorithm, such as the Brute-Force Matcher, is applied to match the descriptors between the two frames and the Random Sample Consensus (RANSAC) algorithm is employed to estimate the homography matrix. The overlap ratio between two frames is then computed. If overlap ratio is less than a predefined threshold, it indicates significant visual changes between the frames, and f_{t+1} is marked as a key-frame. Otherwise, the algorithm proceeds to calculate the overlap ratio between f_t and f_{t+2}. This process continues until a new key-frame is identified, which then becomes the reference for subsequent frames. Considering the effect of viewpoint changes, rotations (both horizontal and vertical) result in larger field-of-view variations, leading to more frames being recorded during these movements. If the indices of the extracted keyframes are denoted as $\mathbf{f}' = \left[f_{k_0}, f_{k_1}, \ldots, f_{k_n} \right]$, the keyframe extraction process can be summarized as:

$$\mathbf{f}' = \text{K-Extractor}(\mathbf{f}). \tag{1}$$

*3.2.2 **Embodied Semantic Representation**. Since perceptual capability is positively correlated with model size [27, 58, 62], we employ a large-scale VLM to process visual inputs to ensure high-quality perception. The differential information of each key frame is described sequentially. This approach provides two key benefits: 1) The sequential and dynamic processing aligns better with the characteristics of embodied scenarios, where visual observations are continuously generated over time. At each moment, the model

should integrate historical semantic representations with the latest visual observations, rapidly updating the semantic understanding of spatial perception. 2) It facilitates the handling of long videos by avoiding the input token limitations that arise when all frames are processed simultaneously by the VLM.

Specifically, for the first frame, the VLM identifies the objects present in the scene, their attributes, and their spatial locations. For subsequent frames, both the previous frame and the current frame are input into the VLM to extract key semantic representation s_{k_j}:

$$s_{k_j} \sim \psi_\theta(s | f_{k_{j-1}}, f_{k_j}; q), \ j = 1, 2, ..., n, \tag{2}$$

where s_{k_j} consists of three items:

- **Action**: Inferring the agent's actions based on the changes in visual observations between consecutive frames.
- △**Information**: Determining changes in the spatial relationships between the agent and known objects, as well as identifying whether new objects appear in the field of view.
- *q*-**related** content: Detecting whether objects or information relevant to the reasoning task appear in the latest field of view.

In this way, we can extract spatial semantic representations $\mathbf{s} = \left[s_{k_0}, s_{k_1}, ..., s_{k_n} \right]$ from the keyframe $\mathbf{f}'$.

Embodied-R: Collaborative Framework for Activating Embodied Spatial Reasoning
in Foundation Models via Reinforcement Learning

MM '25, October 27–31, 2025, Dublin, Ireland

3.3 Small-Scale LM-based Reasoning

Given semantic perception, we can train a training-friendly small-scale language model capable of performing embodied spatial reasoning. Assuming the small-scale LM is denoted as π_θ, the response o inferred from the model can be expressed as: $o \sim \pi_\theta(o \mid q, \mathbf{s})$.

Our training objective is to ensure that the model adheres to the "think-then-answer" paradigm, where the thinking process is logical, and the answer is correct. We follow DeepSeek-R1-Zero and adopt a computationally efficient RL training strategy, Group Relative Policy Optimization (GRPO). Besides rule-based format and accuracy rewards, we propose a novel reasoning process reward tailored for embodied reasoning tasks to mitigate reward hacking and enhance the logical consistency between the reasoning process and the final answer.

3.3.1 **Group Relative Policy Optimization**. For a given query q and semantic annotation $\mathbf{s}$, GRPO generates a group of outputs $\{o_1, o_2, \ldots, o_G\}$ using the reference policy π_{ref}. The reference policy typically refers to the original model not trained via GRPO. The policy model π_θ is then updated by optimizing the following objective:

$$
\mathcal{J}(\theta) = \mathbb{E}_{(q,\mathbf{s}) \sim \mathbb{D}, \{o_i\}_{i=1}^G \sim \pi_{\text{old}}(o|q,\mathbf{s})} \left[\frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_\theta(o_i|q,\mathbf{s})}{\pi_{\text{old}}(o_i|q,\mathbf{s})} A_i, \right. \right. \right.
$$
$$
\left. \left. \left. \text{clip}\left(\frac{\pi_\theta(o_i|q,\mathbf{s})}{\pi_{\text{old}}(o_i|q,\mathbf{s})}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta \mathcal{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right) \right], \tag{3}
$$

where ϵ and β are hyperparameters, and $\mathcal{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$ is KL divergence penalty: $\mathcal{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(r_i|q,s)}{\pi_\theta(r_i|q,s)} - \log \frac{\pi_{\text{ref}}(r_i|q,s)}{\pi_\theta(r_i|q,s)} - 1$. A_i represents the advantage corresponding to the output o_i, calculated from the corresponding $\{r_1, r_2, \ldots, r_G\}$: $A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}$.

3.3.2 **Reward Modeling**. Reward modeling is a critical component of RL algorithms, as their design guides the direction of model optimization. We propose three types of rewards: format reward, accuracy reward, and logical consistency reward. These are designed to respectively guide the model to learn the "think-answer" reasoning pattern, accurate embodied spatial reasoning, and logical consistency between reasoning and the answer.

Format Reward: We aim for the model to output o_i by first producing an embodied reasoning process p_i followed by the final answer a_i. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively:

> *Please assume the role of an agent. Given a question and a series of frames, you should first think about the reasoning process in the mind and then provide the final answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>. Ensure that your answer is consistent with and directly derived from your thinking process, maintaining logical coherence between the two sections. The frames represent your egocentric observations from the past to the present. Question: q. Video:* **f′** *. Assistant:*

A regular expression is applied to evaluate whether o_i meets the specified requirements, thereby generating the format reward r_i':

$$
r_i' = \begin{cases} 1, & \text{if format is correct;} \\ 0, & \text{if format is incorrect.} \end{cases} \tag{4}
$$

Accuracy Reward: The accuracy reward r_i'' model assesses whether the answer a_i is semantically consistent with the ground truth g. For example, multiple-choice questions typically have precise and unique answers, which can be easily extracted when the response adheres to the specified format.

$$
r_i'' = \begin{cases} 1, & a_i = g; \\ 0, & a_i \neq g. \end{cases} \tag{5}
$$

Logical Consistency Reward: When using only the format reward and accuracy reward, we consistently observed hacking behaviors. Specifically, for spatial reasoning tasks where the possible answers are limited (e.g., the relative position of an object with respect to the agent's body), cases arise where an incorrect reasoning process p_i leads to a correct answer a_i, which is mistakenly assigned a positive reward. As such cases accumulate, the logical consistency of the model's responses deteriorates. To address this issue, we introduce a simple yet effective process reward. Our goal is to ensure a lower bound on logical consistency, such that the reasoning ability of π_θ should not degrade below that of the reference model π_{ref}. Therefore, when the model's answer is correct ($a_i = g$), we input the question q and reasoning process p_i into the reference model without providing video frames, yielding an answer:

$$
a_i' \sim \pi_{\text{ref}}(a|q, p_i). \tag{6}
$$

If a_i' is consistent with a_i, it indicates that the reasoning process can logically lead to the answer; otherwise, it reflects a logical inconsistency between the reasoning process and the answer.

$$
r_i''' = \begin{cases} 1, & a_i = a_i' = g; \\ 0, & \text{else.} \end{cases} \tag{7}
$$

Total Reward: The total reward is a linear combination of the three rewards mentioned above:

$$
r_i = \omega_1 r_i' + \omega_2 r_i'' + \omega_3 r_i'''. \tag{8}
$$

4 Experiments

We first provide the details of the experimental setup and then demonstrate the following: quantitative results, qualitative results, and ablation studies. These correspond to addressing the following three research questions (RQs):

- **RQ1: How does Embodied-R perform compared to existing video-LLMs?**
- **RQ2: Has Embodied-R learned slow-thinking?**
- **RQ3: What are the contributions of each module?**

4.1 Experimental Setup

4.1.1 **Data Preparation**. We primarily focus on spatial reasoning problems during motion within three-dimensional physical space to evaluate the effectiveness of our method. For this purpose, we selected two embodied video datasets as the main training and testing sets: VSI-Bench [58], which contains indoor first-person navigation

Table 1: Accuracy of Embodied-R and baselines on 8 indoor and outdoor embodied spatial reasoning tasks. The baselines include popular proprietary models, state-of-the-art (SOTA) multimodal reasoning models, open-sourced video-large language models, and models fine-tuned on the same training dataset.

Method	Avg.	UrbanVideo-Bench				VSI-Bench			
		Landmark Position	Counterfactual	Progress Evaluation	Action Generation	Relative Distance	Relative Direction	Route Planning	Appearance Order
Random	24.0	19.7	25.0	21.8	16.4	25.0	36.1	28.3	25.0
Proprietary Models (API)									
Qwen-VL-Max[32f]	34.1	44.8	49.2	38.8	29.6	28.0	33.3	29.6	28.3
GPT-4o[32f]	35.7	36.8	44.7	34.2	33.8	37.0	41.3	31.5	28.5
Gemini-1.5-Flash[1fps]	38.3	37.8	42.4	43.3	34.4	37.7	41.0	31.5	37.8
Gemini-1.5-Pro[1fps]	39.7	37.4	46.2	38.8	31.9	51.3	46.3	36.0	34.6
SOTA Reasoning Models (API)									
OpenAI-o1[32f]	37.2	34.6	53.3	39.1	28.0	39.7	35.8	52.9	39.8
Gemini-2.5-Pro[1fps]	40.8	40.0	75.0	38.7	23.5	42.0	34.5	52.4	63.6
Open-source Models									
LLaVA-NeXT-Video-7B-hf[32f]	29.5	49.5	20.5	36.6	19.2	25.2	26.3	29.9	24.5
Phi-3.5-vision-instruct[32f]	29.0	49.2	34.8	33.2	15.6	25.4	26.5	36.9	25.2
Kangaroo[64f]	30.0	35.5	42.4	32.5	32.4	25.2	26.8	23.5	24.9
InternVL2-2B[32]	24.5	19.3	45.5	29.2	20.9	25.1	25.0	32.6	23.9
InternVL2-8B[32f]	25.5	23.1	45.5	31.5	21.4	24.7	25.7	28.3	24.8
InternVL2-40B[32f]	25.8	23.2	41.7	32.4	22.3	24.9	25.7	29.4	24.5
Qwen2.5-VL-3B-Instruct[1fps]	33.1	32.1	47.8	34.0	31.0	27.9	32.6	39.0	38.9
Qwen2.5-VL-7B-Instruct[1fps]	33.3	33.3	21.7	25.0	27.8	35.8	39.7	48.8	38.8
Qwen2.5-VL-72B-Instruct[1fps]	34.9	34.7	34.8	26.4	37.7	40.8	29.0	32.5	43.9
Supervised Fine-Tuning									
Qwen2.5-VL-3B-Instruct[1fps]	41.7	47.7	33.4	34.8	39.2	42.6	42.3	41.2	43.9
Qwen2.5-VL-7B-Instruct[1fps]	45.4	40.2	53.4	38.0	40.8	47.8	46.3	44.1	56.1
Proposed Embodied-R									
VLM-72B + LLM-3B [≤32f]	51.1	55.1	59.9	39.7	47.6	50.0	44.3	36.8	72.0
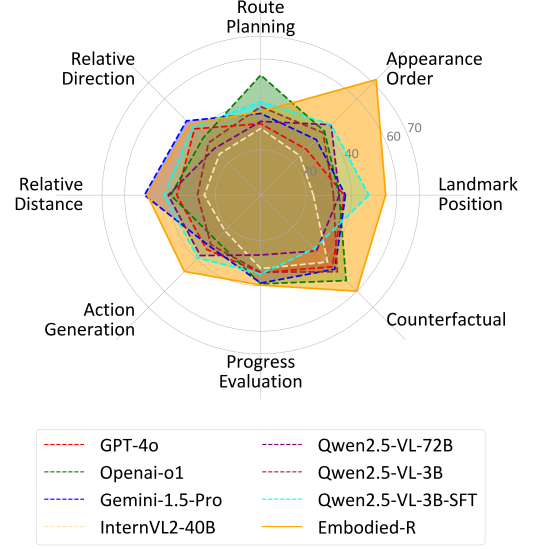
Table 2: Ablation of Key-Frame Extractor

	Avg. Frame	Acc.	Training Time	Inference Time
w/o	32	49.5	127.87 h	243.68 s
w	20.7$^{\downarrow 11.3}$	51.1$^{\uparrow 1.6}$	111.70$h^{\downarrow 16.17}$	157.55$s^{\downarrow 86.13}$

Table 3: Ablation of Collaboration.

	Avg.	LP	C	PE	AG	RDist	RDir	RP	AO
w/o	34.8	31.8	45.7	28.3	28.1	41.0	29.7	37.5	46.0
w	51.1	55.1	59.9	39.7	47.6	50.0	44.3	36.8	72.0
△	+16.3	+23.3	+14.2	+11.4	+19.5	+9.0	+14.6	-0.7	+26.0

data. , and UrbanVideo-Bench [62], which consists of outdoor embodied data captured by drones navigating through aerial spaces These datasets provide diversity in scenarios by incorporating both outdoor and indoor video data. Based on the content of the tasks, we specifically selected four distinct types of tasks from each dataset, characterized by long spatial reasoning chains and low accuracy. These tasks are formulated as multiple-choice question-answering problems, ensuring determinism in answers to facilitate RL training and allowing direct calculation of accuracy to evaluate performance. Across eight task categories, the dataset covers multiple levels of spatial reasoning, comprising a total of 5,415 QA pairs and 1,492 videos. Additionally, we include two out-of-distribution dataset, EgoSchema [34] and Egocentric task in MVBench [27]. EgoSchema is designed for task-level reasoning from a first-person perspective, with 500 QA pairs and 500 videos available in its fully open-source portion. MVBench encompasses the embodied task of egocentric navigation, comprising 200 QA pairs and 200 corresponding videos. These datasets serves to evaluate the generalization capability of the trained model.

To ensure comprehensive evaluation, we conducted five repeated experiments. The dataset was randomly divided into five equal parts and 5-fold cross-validation is adopted. The final testing results are averaged across the five experiments. Furthermore, we address the issue of potential semantic bias in the datasets. For instance, in action generation tasks, forward movement may inherently have a higher correctness rate than adjusting the gimbal angle, which is a characteristic of the task itself. To prevent the testing performance from being influenced by the model learning textual distribution rather than truly understanding the spatial information in video, we implement an additional filtering step for the testing set. Specifically, we train a LLM through supervised fine-tuning using only the textual QA pairs from the training set, without video inputs. If a question in the testing set can be correctly answered by the fine-tuned LLM but not by the original LLM, it indicates semantic bias in that QA pair. These biased QA pairs are excluded from the testing set as they fail to accurately assess the spatial reasoning capabilities of models.

Embodied-R: Collaborative Framework for Activating Embodied Spatial Reasoning
in Foundation Models via Reinforcement Learning

MM '25, October 27–31, 2025, Dublin, Ireland

Figure 3: Case Analysis: Embodied-R has initially developed the ability for slow-thinking: it can think before answering, effectively distinguish spatial relationships, provide structured and organized responses, and integrate information across multiple frames for embodied scene analysis.

Figure 4: Ablation of RL training and comparison to other language models.

*4.1.2 **Implementation Details**.* We use Qwen2.5-3B-Instruct [57] as the small-scale LM and Qwen2.5-VL-72B-Instruct [6] as large-scale VLM. Both training and inference processes were conducted using 8 NVIDIA A800-SXM4-40GB GPUs, with each RL training requiring approximately 90 GPU hours. Other key hyperparameters for training are as follows: learning rate: 5e-7, temperature: 1.0, train batch size: 32, rollout size: 8, KL coefficient: 0.001, maximum response length: 2048, input length: 6144. When conducting inference on the test set, the temperature is set to 0.5.

*4.1.3 **Three-Stage Training Schedule**.* As for the RL training on the LM, we design a three-stage training schedule to achieve a smooth improvement in training performance. The primary distinction between stages lies in the different weight ratios assigned to three types of rewards.

- **Stage 1:** In epochs 1 and 2, the goal is to guide the model to follow the "<think> </think> " output format. At this stage, the weights are set as $\omega_1 : \omega_2 : \omega_3 = 7 : 3 : 0$. Correct format rewards also assist in locating the answer and reduce

misjudgment in accuracy. During this phase, the format reward rapidly converges to 1.

- **Stage 2:** In epochs 3 and 4, the focus shifts to improving the accuracy of the model's responses, guiding the model to produce correct reasoning answers. The weights are set as $\omega_1 : \omega_2 : \omega_3 = 3 : 7 : 0$.

- **Stage 3:** In subsequent 5-12 epochs, the aim is to enhance accuracy while simultaneously improving the quality of the "thinking" process, ensuring logical consistency between thinking and the answer. The weights are set as $\omega_1 : \omega_2 : \omega_3 = 1 : 7 : 2$.

4.2 How Does Embodied-R Perform Compared to Existing Video-LLMs?

To evaluate the effectiveness of the proposed method, in addition to the random baseline, we introduced four categories comprising 17 multimodal large language models capable of processing video inputs:

- **Proprietary Models:** Cost-effective multimodal models with over 100B parameters, including Qwen-VL-Max [46], GPT-4o [37], Gemini-1.5-Flash [44], and Gemini-1.5-Pro [44].

- **SOTA Reasoning Models:** State-of-the-art reasoning models with the highest performance but significant computational cost, including OpenAI-o1 [38] and Gemini-2.5-Pro [21].

- **Open-Source Models:** Popular open-source multimodal models, including LLaVA-NeXT-Video-7B-hf [29], Phi-3.5-vision-instruct [1], the Internvl2 series [11], and the Qwen-VL series [6].

- **Supervised Fine-Tuning (SFT):** Considering the scarcity of embodied video tasks, the aforementioned models may lack exposure to relevant data. Therefore, Qwen2.5-VL-3B-Instruct [6] and Qwen2.5-VL-7B-Instruct [6] are fine-tuned for these tasks.

The results presented in Table 1 lead to the following conclusions:

- After undergoing RL training on embodied reasoning tasks, our model significantly outperformed proprietary models as well

as OpenAI-o1 and Gemini-2.5-Pro by over **10%**. Moreover, it consistently demonstrated leading performance across various tasks. These results highlight the considerable **difficulty of embodied reasoning tasks** and indicate that current reasoning models lack generalization capability for such spatial reasoning challenges. On the other hand, the findings confirm that **collaborative framework with RL can effectively enhance model reasoning performance in specific domains**, especially for tasks that remain poorly solved.

- For embodied video reasoning, a highly coupled perception-reasoning problem, the VLM model Qwen2.5-VL-72B-Instruct achieved an accuracy of only 34.9% through direct inference. In contrast, incorporating a small-scale LM model improved accuracy to 51.1%. **Given limited computational resources for training, the collaborative framework proposed in this study provides an effective solution for balancing model size with hardware constraints**.
- Under similar computational resource limitations, direct fine-tuning is restricted to models with a size of 7B or smaller. However, the perceptual capacity of small-scale VL models imposes a low upper bound on accuracy compared to Embodied-R. Additionally, fine-tuned models lack the capability for slow-thinking.

4.3 Has Embodied-R Learned Slow-Thinking?

Beyond the quantitative results, we aim to explore whether spatial reasoning capabilities in the output of Embodied-R are improved. As illustrated in Figure 3, after RL training, Embodied-R demonstrates the following human-like reasoning ways:

- **Spatial Relationship Reasoning**: Accurately inferring the relative spatial relationship between itself and the surrounding environment.
- **Systematic Analysis**: Breaking down problems into components, presenting answers with a "part-to-whole" structure, and maintaining clear logical organization.
- **Contextual Integration**: Integrating semantic information across different frames to perform comprehensive analysis.
- **Think-Answer Format**: Strictly adhering to a structured process of reasoning before outputting the final answer.

In summary, Embodied-R demonstrates a certain degree of slow-thinking capability in embodied spatial reasoning.

4.4 Contributions of Each Module

*4.4.1 **Ablation of Key-Frame Extractor***. The role of Key-Frame Extractor is to reduce inference time and training time by retaining essential frames and removing redundant ones while maintaining perceptual quality. As shown in Table 2, while improving accuracy, training time is significantly reduced by 12.6%, and single inference time is reduced by approximately one-third.

*4.4.2 **Ablation of Collaboration***. The collaborative framework enables improved reasoning capabilities under limited computational resources for training. With training-free large-scale pre-trained VLMs, it only requires training small-scale LM models to achieve enhanced reasoning performance. As shown in Table 3,

with identical key-frame inputs and using the same VLM, Qwen2.5-VL-72B-Instruct, the overall accuracy of collaborative inference is 1.5 times higher than that of the standalone VLM.

*4.4.3 **Ablation of RL Training***. RL is central to the LM training in this paper. Without RL training, directly applying the original LM-3B model for reasoning leads to poor performance, as the LM has limited exposure to embodied spatial reasoning data during pretraining. After RL training, the LM achieves significant improvements, with a 27.9% increase on the UrbanVideo-Bench and a 20.6% increase on the VSI-Bench benchmarks.

Given that VLM has already transformed visual inputs into textual representations, we introduced 4 text-based reasoning models (o3-mini [39], Deepseek-R1 [24], Qwen-Max [46], Qwen2.5-7B-Instruct [6]) as baselines to further assess the importance of reasoning capability in the embodied spatial task. The results demonstrate a clear positive correlation between the reasoning ability of the model and its accuracy. The strong performance of Embodied-R may not only stem from its familiarity with the data distribution but also from its synergy with the representations provided by the VLM. Following training, the small-scale LM becomes more attuned to the VLM-generated representations, which translates into enhanced performance on embodied reasoning tasks.

5 Further Exploration

Building upon the aforementioned experiments, we further explore four intriguing RQs related to embodied video-based RL training:

- **RQ4: What Is the Relationship Between Inference Ability, Aha Moments, and Response Length?**
- **RQ5: Why Not Directly Perform RL Training on VLLMs?**
- **RQ6: Is Accuracy+Format Rewards All You Need?**
- **RQ7: RL vs SFT when Generalize to Out-of-Distribution (OOD) Embodied Tasks?**

5.1 Relationship Between Inference Ability, Aha Moments, and Response Length?

The GRPO training process is illustrated in Figure 5a-d, which correspond to the validation set's accuracy reward, format reward, ratio of logical consistency reward to accuracy reward, and the response length, respectively. Notably, existing pure-text-based reproductions [55, 59] of DeepSeek-R-Zero models identify inference ability and the "aha moment" as key indicators of emergent reasoning capabilities. However, such phenomena are rarely observed in other multimodal reasoning tasks, such as image-based reasoning [10, 33]. This leads us to hypothesize that response length is strongly influenced by the nature of the question itself. For instance, mathematical problems often require multi-step calculations, where increased reasoning length tends to correlate positively with reasoning ability. In contrast, for multimodal reasoning tasks like embodied spatial reasoning, the LM model training process converges toward an optimal range of text output distributions. Concise reasoning patterns may facilitate the embodied spatial reasoning. This highlights the versatility of RL-based post-training method, demonstrating the ability to benefit a wide range of reasoning tasks.

Embodied-R: Collaborative Framework for Activating Embodied Spatial Reasoning
in Foundation Models via Reinforcement Learning

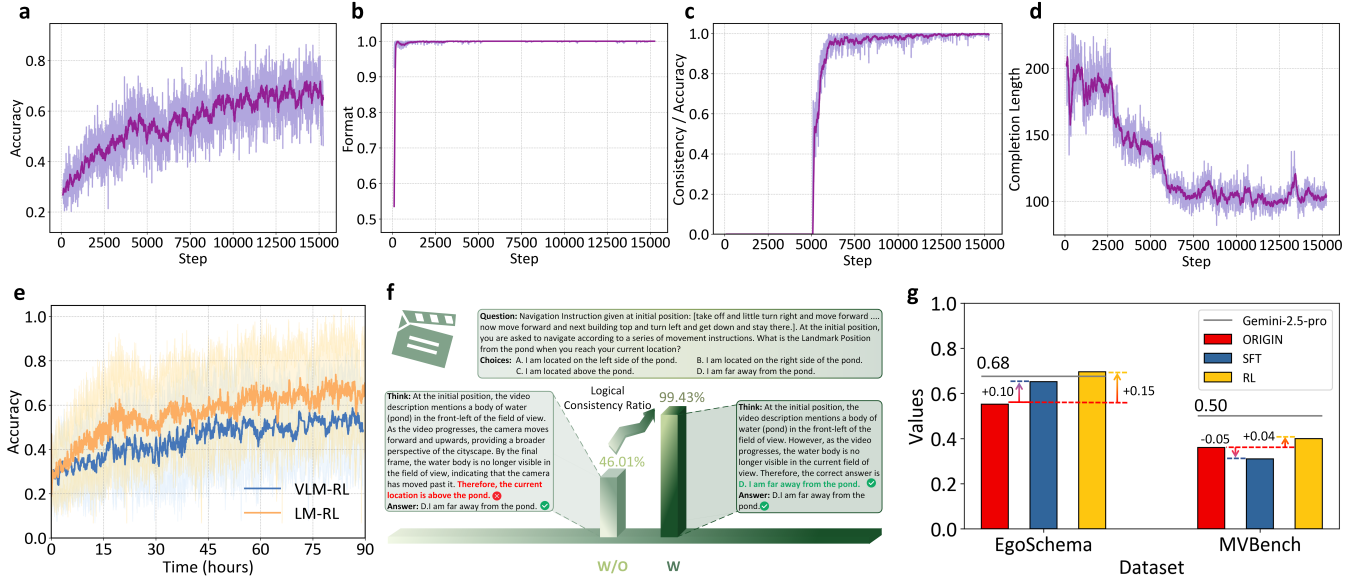MM '25, October 27–31, 2025, Dublin, Ireland

Figure 5: a-d. The GRPO training process (a: accuracy reward; b: format reward; c: ratio of logical consistency reward to accuracy reward; d: response length of validation set). e. Comparison of accuracy reward curves for RL training of equivalently sized LM and VLM models. f. Model performance before and after integrating logical consistency reward. g. Comparison of generalization performance between models trained with RL and SFT.

5.2 Why Not Directly Perform RL on VLLMs?

We previously attempted direct RL training on the Qwen-VL-3B-Instruct model. As shown in Figure 5e, under similar training parameters and time, the performance of the VLM was notably inferior to that of the LM. Upon convergence, the VLM achieved an accuracy of 43.8% on the test set, significantly lower than the LM. The limited perceptual capability of the VLM restricts its potential for reasoning improvements. Therefore, under resource-constrained conditions, collaborative inference integrating models of different scales present a promising solution.

5.3 Is Accuracy+Format Rewards All You Need?

According to the Deepseek-R1-Zero, it appears that accuracy and format rewards are enough to guide the model toward correct reasoning. However, during training in our problem, we observed instances of reward hacking, where the model optimizes the answer but the reasoning process leading to that answer is inconsistent with the answer itself. We aim to ensure alignment between the model's reasoning process and its answer, both to enhance generalization and improve the interpretability of the reasoning process. As shown in Figure 5f, we employ GPT-4o to evaluate the proportion of logically consistent outputs on the test set before and after incorporating a logical consistency reward. This proportion increased from 46.01% to 99.43% after the reward was added, demonstrating the value of this approach in addressing embodied spatial multiple-choice reasoning tasks. Moreover, this reward mechanism could potentially be extended to other reasoning tasks prone to answer accuracy hacking during training.

5.4 RL vs SFT when Generalize to Out-of-Distribution (OOD) Embodied Tasks?

For small-scale LMs, we aim to explore their generalization performance when trained with SFT instead of RL. To evaluate this, we introduced two OOD datasets: EgoSchema and the egocentric task in MVBench. As discussed in Sections 4.1.1, these two OOD datasets differ significantly from the training set in both task content and scene characteristics. The accuracy results are shown in Figure 5g. RL-trained models demonstrate generalization ability across both datasets. On the EgoSchema dataset, the RL-trained language model under the Embodied-R framework even achieve performance comparable to the state-of-the-art multimodal reasoning model, Gemini-2.5-Pro. SFT-trained models showed improvement on EgoSchema but a decline on MVBench. This suggests that slow reasoning, as employed in RL models, could be a promising approach to improve the generalization capabilities even for small-scale models.

6 Conclusion

To address embodied spatial reasoning tasks, we propose a collaborative framework that leverages the perceptual capabilities of large-scale VLMs and the reasoning potential of compact LMs. Through 90 hours of RL training on a 3B LM using 8 NVIDIA A800-SXM4-40GB GPUs, Embodied-R surpasses OpenAI-o1 by 13.9% and Gemini-2.5-Pro by 10.3% on the test set. Other Key findings include: (1) RL training leads to output length convergence, aligning with the requirements of the task; (2) the reasoning upper bound of same-scale VLMs trained with RL is significantly lower than that of Embodied-R, due to inherent limitations in perception; (3) the proposed logical consistency reward enhances reasoning quality; and (4) models trained via RL exhibit stronger generalization on out-of-distribution datasets compared to those trained with SFT.

Acknowledgments

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219* (2024).

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Michael Ahn, Debidatta Dwibedi, Chelsea Finn, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Karol Hausman, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, et al. 2024. Autort: Embodied foundation models for large scale orchestration of robotic agents. *arXiv preprint arXiv:2401.12963* (2024).

[4] Cameron A Aubin, Benjamin Gorissen, Edoardo Milana, Philip R Buskohl, Nathan Lazarus, Geoffrey A Slipher, Christoph Keplinger, Josh Bongard, Fumiya Iida, Jennifer A Lewis, et al. 2022. Towards enduring autonomous robots via embodied energy. *Nature* 602, 7897 (2022), 393–402.

[5] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. 2022. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 19129–19139.

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

[7] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. 2024. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems* 37 (2024), 53168–53197.

[8] Bolei Chen, Jiaxu Kang, Ping Zhong, Yixiong Liang, Yu Sheng, and Jianxin Wang. 2024. Embodied Contrastive Learning with Geometric Consistency and Behavioral Awareness for Object Navigation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 4776–4785.

[9] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14455–14465.

[10] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. 2025. R1-V: Reinforcing Super Generalization Ability in Vision-Language Models with Less Than $3. https://github.com/Deep-Agent/R1-V. Accessed: 2025-02-02.

[11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 24185–24198.

[12] Sijie Cheng, Kechen Fang, Yangyang Yu, Sicheng Zhou, Bohao Li, Ye Tian, Tingguang Li, Lei Han, and Yang Liu. 2024. Videogothink: Assessing egocentric video understanding capabilities for embodied ai. *arXiv preprint arXiv:2410.11623* (2024).

[13] Stephanie Clarke and Judit Miklossy. 1990. Occipital cortex in man: Organization of callosal connections, related myelo-and cytoarchitecture, and putative boundaries of functional visual areas. *Journal of Comparative Neurology* 298, 2 (1990), 188–214.

[14] Maël Donoso, Anne GE Collins, and Etienne Koechlin. 2014. Foundations of human reasoning in the prefrontal cortex. *Science* 344, 6191 (2014), 1481–1486.

[15] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. 2023. Palm-e: An embodied multimodal language model. (2023).

[16] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. *arXiv preprint arXiv:2501.03230* (2024).

[17] Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. 2022. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications* 13, 1 (2022), 3094.

[18] Leonardo Fogassi, Pier Francesco Ferrari, Benno Gesierich, Stefano Rozzi, Fabian Chersi, and Giacomo Rizzolatti. 2005. Parietal lobe: from action organization to intention understanding. *Science* 308, 5722 (2005), 662–667.

[19] Lucia Foglia and Robert A Wilson. 2013. Embodied cognition. *Wiley Interdisciplinary Reviews: Cognitive Science* 4, 3 (2013), 319–325.

[20] Chen Gao, Baining Zhao, Weichen Zhang, Jinzhu Mao, Jun Zhang, Zhiheng Zheng, Fanhang Man, Jianjie Fang, Zile Zhou, Jinqiang Cui, et al. 2024. EmbodiedCity: A Benchmark Platform for Embodied Agent in Real-world City Environment. *arXiv preprint arXiv:2410.09604* (2024).

[21] Google. 2024. Gemini API. https://ai.google.dev/gemini-api. Accessed: 2025-04-12.

[22] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18995–19012.

[23] Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking. *arXiv preprint arXiv:2501.04519* (2025).

[24] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[25] Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398* (2023).

[26] James Intriligator and Patrick Cavanagh. 2001. The spatial resolution of visual attention. *Cognitive psychology* 43, 3 (2001), 171–216.

[27] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22195–22206.

[28] Tianlin Li, Qian Liu, Tianyu Pang, Chao Du, Qing Guo, Yang Liu, and Min Lin. 2024. Purifying large language models by ensembling a small language model. *arXiv preprint arXiv:2402.14845* (2024).

[29] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122* (2023).

[30] Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics* 11 (2023), 635–651.

[31] Hongbin Liu, Yongze Zhao, Peng Dong, Xiuyi Guo, and Yilin Wang. 2024. IOF-Tracker: A Two-Stage Multiple Targets Tracking Method Using Spatial-Temporal Fusion Algorithm. *Applied Sciences* 15, 1 (2024), 107.

[32] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2024. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886* (2024).

[33] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785* (2025).

[34] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems* 36 (2023), 46212–46244.

[35] Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, et al. 2024. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *arXiv preprint arXiv:2412.09413* (2024).

[36] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2023. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems* 36 (2023), 25081–25094.

[37] OpenAI. 2024. GPT-4o API. https://openai.com/api/. Accessed: 2025-04-12.

[38] OpenAI. 2024. Learning to Reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/ Accessed: 2025-03-04.

[39] OpenAI. 2025. OpenAI o3-mini. https://openai.com/index/openai-o3-mini/ Accessed: 2025-04-15.

[40] Amrith Setlur, Saurabh Garg, Xinyang Geng, Naman Garg, Virginia Smith, and Aviral Kumar. 2025. Rl on incorrect synthetic data scales the efficiency of llm math reasoning by eight-fold. *Advances in Neural Information Processing Systems* 37 (2025), 43000–43031.

[41] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. 2023. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*. PMLR, 492–504.

[42] Alessandro Suglia, Claudio Greco, Katie Baker, Jose L Part, Ioannis Papaioannou, Arash Eshghi, Ioannis Konstas, and Oliver Lemon. 2024. Alanavlm: A multimodal embodied ai foundation model for egocentric video understanding. *arXiv preprint arXiv:2406.13807* (2024).

[43] Guangzhi Sun, Yudong Yang, Jimin Zhuang, Changli Tang, Yixuan Li, Wei Li, Zejun MA, and Chao Zhang. 2025. video-SALMONN-o1: Reasoning-enhanced Audio-visual Large Language Model. *arXiv preprint arXiv:2502.11775* (2025).

Embodied-R: Collaborative Framework for Activating Embodied Spatial Reasoning
in Foundation Models via Reinforcement Learning

MM '25, October 27–31, 2025, Dublin, Ireland

[44] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[45] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599* (2025).

[46] Qwen Team. 2024. Qwen-VL-Max. https://qwenlm.github.io/blog/qwen-vl-max/. Accessed: 2025-04-12.

[47] Qwen Team. 2024. QwQ: Reflect Deeply on the Boundaries of the Unknown. https://qwenlm.github.io/blog/qwq-32b-preview/

[48] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. 2025. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186* (2025).

[49] Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Joon Oh. 2024. Calibrating large language models using their generations only. *arXiv preprint arXiv:2403.05973* (2024).

[50] Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, et al. 2024. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *arXiv preprint arXiv:2411.03350* (2024).

[51] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. 2024. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems* 37 (2024), 75392–75421.

[52] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. 2024. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19757–19767.

[53] Zhecan Wang, Garrett Bingham, Adams Wei Yu, Quoc V Le, Thang Luong, and Golnaz Ghiasi. 2024. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. In *European Conference on Computer Vision*. Springer, 288–304.

[54] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[55] Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-RL: Unleashing LLM Reasoning with Rule-Based Reinforcement Learning. *arXiv preprint arXiv:2502.14768* (2025).

[56] Cheng Xu, Xiaofeng Hou, Jiacheng Liu, Chao Li, Tianhao Huang, Xiaozhi Zhu, Mo Niu, Lingyu Sun, Peng Tang, Tongqiao Xu, et al. 2023. Mmbench: Benchmarking end-to-end multi-modal dnns and understanding their hardware-software implications. In *2023 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, 154–166.

[57] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122* (2024).

[58] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2024. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171* (2024).

[59] Weihao Zeng, Yuzhen Huang, Wei Liu, Keqing He, Qian Liu, Zejun Ma, and Junxian He. 2025. 7B Model and 8K Examples: Emerging Reasoning with Reinforcement Learning is Both Effective and Efficient. https://hkust-nlp.notion.site/simplerl-reason. Notion Blog.

[60] Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2025. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems* 37 (2025), 64735–64772.

[61] Yiming Zhang, Nicholas Carlini, and Daphne Ippolito. 2023. Effective prompt extraction from language models. *arXiv preprint arXiv:2307.06865* (2023).

[62] Baining Zhao, Jianjie Fang, Zichao Dai, Ziyou Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, and Yong Li. 2025. UrbanVideo-Bench: Benchmarking Vision-Language Models on Embodied Intelligence with Video Data in Urban Spaces. arXiv:2503.06157 [cs.CV] https://arxiv.org/abs/2503.06157

[63] Theodore Zhao, Mu Wei, J Samuel Preston, and Hoifung Poon. 2023. Automatic Calibration and Error Correction for Generative Large Language Models via Pareto Optimal Self-Supervision. (2023).

[64] Karl Zilles and Katrin Amunts. 2010. Centenary of Brodmann's map—conception and fate. *Nature Reviews Neuroscience* 11, 2 (2010), 139–145.

A Appendix

A.1 Dataset Introduction

UrbanVideo-Bench: UrbanVideo-Bench is one of the training and testing datasets designed for embodied reasoning (embodied-r). This benchmark was proposed by Tsinghua University in February 2025. It captures two embodied characteristics of urban environments: complex urban scenes featuring dynamic and static elements, and unique aerial navigation scenarios. The dataset consists of 4 categories and 16 tasks, aimed at evaluating Video-LLMs in terms of recall, perception, reasoning, and navigation capabilities. In our paper, we focus on 4 of these complex tasks for reinforcement learning in video-based learning: **Landmark Position**, **Counterfactual Reasoning**, **Progress Evaluation**, and **Action Generation**, which represent challenging embodied outdoor tasks.

VSI-Bench: VSI-Bench is another training and testing dataset for embodied reasoning (embodied-r). Proposed by Fei-Fei Li's team at Stanford in December 2024, this dataset provides high-quality evaluation metrics for assessing the 3D, video-based, visual-spatial intelligence of multimodal large language models (MLLMs). The dataset comprises 2 categories and 8 tasks designed to evaluate key aspects of spatial reasoning. In our paper, we focus on 4 tasks for reinforcement learning in video-based learning: **Relative Distance**, **Relative Direction**, **Route Planning**, and **Appearance Order**, all of which are categorized as challenging embodied outdoor tasks.

EgoSchema: EgoSchema is one of the Out-of-Distribution (OOD) datasets utilized to evaluate the generalization capability of our model. This dataset is specifically designed as a long-form video question-answering benchmark, aimed at assessing modern vision and language systems' ability to understand and reason over extended video content. It provides a rigorous evaluation framework for long video understanding tasks.

MVBench: MVBench is another Out-of-Distribution (OOD) dataset employed to test the generalization capability of our model. MVBench consists of 20 complex video tasks, offering a comprehensive benchmark for evaluating the video understanding capabilities of existing multimodal models. This dataset is designed to address diverse and challenging scenarios in video-based reasoning.

A.2 Details of Key-Frame Extractor

The goal of key-frame extraction is to ensure sufficient information gain between frames while maintaining a certain degree of overlap. The specific process is as follows:

Step 1: a perspective transformation is used to model the geometric relationship between frames. Assuming f_t is a key-frame, to determine whether f_{t+1} should also be considered a keyframe, keypoints and descriptors are calculated from f_t and f_{t+1} using the Oriented FAST and Rotated BRIEF (ORB) algorithm:

$$\text{Keypoints}_t, \text{Descriptors}_t = \text{ORB}(f_t), \tag{9}$$

$$\text{Keypoints}_{t+1}, \text{Descriptors}_{t+1} = \text{ORB}(f_{t+1}). \tag{10}$$

Next, a feature matching algorithm, such as the Brute-Force Matcher, is applied to match the descriptors between the two frames, identifying corresponding keypoint pairs $\mathbf{l}_t^{\text{key}}$ and $\mathbf{l}_{t+1}^{\text{key}}$. Using the matched keypoint pairs, the Random Sample Consensus (RANSAC) algorithm is employed to estimate the homography matrix $\mathbf{M}$, which maps the content of f_{t+1} to the coordinate space of f_t.

Step 2: The overlap ratio between two frames is then computed. Assuming the size of each video frame is $w \times h$, for frames f_t and f_{t+1}: $\mathbf{l}_t = \{[0,0], [w,0], [w,h], [0,h]\}$ represents the four corner points of f_t; $\mathbf{l}_{t+1} = \{[0,0], [w,0], [w,h], [0,h]\}$ represents the four corner points of f_{t+1}. Using the homography matrix $\mathbf{M}$, the corner points $\mathbf{l}_{t+1}$ of f_{t+1} are transformed into the coordinate space of f_t: $\mathbf{l}'_{t+1,i} = \mathbf{M} \cdot \mathbf{l}_{t+1,i}$, where $\mathbf{l}_{t+1,i} = [x, y, 1]^T$ represents the corner points of f_{t+1} in homogeneous coordinates, and $\mathbf{l}'_{t+1,i} = [x', y', w']^T$ represents the transformed corner points. The transformed points are further normalized to recover 2D coordinates, resulting in a quadrilateral representing f_{t+1} in f_t's space. In f_t's coordinate space, there are two polygons: Polygon L_t is defined by the corner points $\mathbf{l}_t$ of f_t; Polygon L'_{t+1} is defined by the transformed corner points $\mathbf{l}'_{t+1}$. Thus, the overlap ratio c is defined as:

$$c = \frac{\text{Area}(L_t \cap L'_{t+1})}{\text{Area}_{\text{total}}}. \tag{11}$$

If c is less than a predefined threshold ε, it indicates significant visual changes between the frames, and f_{t+1} is marked as a key-frame. Otherwise, the algorithm proceeds to calculate the overlap ratio between f_t and f_{t+2}. This process continues until a new key-frame is identified, which then becomes the reference for subsequent frames. Considering the effect of viewpoint changes, rotations (both horizontal and vertical) result in larger field-of-view variations, leading to more frames being recorded during these movements. If the indices of the extracted keyframes are denoted as $\mathbf{f}' = [f_{k_0}, f_{k_1}, \ldots, f_{k_n}]$, the keyframe extraction process can be summarized as:

$$\mathbf{f}' = \text{K-Extractor}(\mathbf{f}). \tag{12}$$

A.3 Details of Data Preparation

A.3.1 Task Selection Criteria. In our study, we carefully selected specific tasks that emphasize spatial reasoning capabilities during motion within three-dimensional physical space. The selection process was guided by several key considerations:

Focus on Reasoning Processes: We prioritized tasks that require deep cognitive processing rather than simple recognition or recall. As highlighted in the main text, embodied spatial reasoning involves complex spatio-temporal relationships where agents must discover object associations across frames and extract task-relevant semantics. For instance, navigation tasks require agents to infer their location from historical observations, construct mental maps, develop high-level plans, and determine specific actions—processes that demand sophisticated reasoning capabilities.

Diversity in Spatial Contexts: To ensure comprehensive evaluation, we selected tasks from both indoor (VSI-Bench) and outdoor (UrbanVideo-Bench) environments, providing diverse spatial contexts that test different aspects of embodied reasoning. This diversity is crucial for evaluating the generalizability of our approach across varying spatial scales and environmental complexities.

Emphasis on Long Reasoning Chains: We specifically targeted tasks characterized by long spatial reasoning chains and historically low accuracy rates. These challenging tasks better demonstrate the value of our "slow thinking" approach, which encourages thorough reasoning before generating responses—similar to how recent advances in mathematical and scientific reasoning have benefited from reinforcement learning techniques.

Embodied-R: Collaborative Framework for Activating Embodied Spatial Reasoning
in Foundation Models via Reinforcement Learning

MM '25, October 27–31, 2025, Dublin, Ireland

Table 4: Hyperparameters used in reinforcement learning training of Embodied-R.

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	5e-7
Temperature	1.0
Train Batch Size	32
Rollout Size	8
KL Coefficient	0.001
Maximum Response Length	2048
Input Length	6144
Training Epochs	12

Deterministic Evaluation: All selected tasks were formulated as multiple-choice question-answering problems to ensure determinism in answers, facilitating both RL training and direct calculation of accuracy for performance evaluation.

A.3.2 Question Filtering Methodology. To ensure the quality and validity of our dataset, we implemented a rigorous question filtering process:

Blind Testing Filter: We first evaluated questions using an untrained 7B language model without video input (blind selection). Questions that could be correctly answered without visual information were identified as potentially problematic, as they might rely more on textual patterns or common knowledge rather than genuine spatial reasoning based on video content.

SFT-based Filtering: After conducting supervised fine-tuning (SFT) without video inputs, we analyzed which question types showed significant improvement in accuracy. Categories where the model's performance increased substantially without visual

information were flagged for removal, as this indicated strong correlations between question text and answers that could be exploited without actual spatial reasoning.

Correlation Analysis: We specifically eliminated question types where:

- The model could achieve high accuracy without accessing video content
- Performance improved dramatically after text-only SFT training
- Question-answer pairs exhibited strong textual patterns that could be exploited without spatial understanding

This filtering methodology ensured that our final dataset genuinely tests embodied spatial reasoning capabilities rather than linguistic pattern matching or prior knowledge exploitation. By removing questions with strong text-answer correlations, we created a more challenging and valid benchmark that requires models to truly understand spatial relationships from video content.

A.4 RL Hyperparameters

The reinforcement learning (RL) training of Embodied-R requires careful hyperparameter tuning to balance computational efficiency with model performance. We conducted extensive experiments to determine the optimal configuration for our collaborative framework. The key hyperparameters used in our RL training process are summarized in Table 4. These settings were selected to ensure stable training while maximizing the model's embodied reasoning capabilities. Notably, we used a relatively small learning rate (5e-7) to prevent catastrophic forgetting and a moderate KL coefficient (0.001) to maintain proximity to the reference model while allowing sufficient exploration.