EQA-RM: A Generative Embodied Reward Model with Test-time Scaling

Yuhang Chen[1] **Zhen Tan**[2] **Tianlong Chen**[1]
[1]The University of North Carolina at Chapel Hill [2]Arizona State University
{yuhang, tianlong}@cs.unc.edu ztan36@asu.edu

Abstract

Reward Models (RMs), vital for large model alignment, are underexplored for complex embodied tasks like Embodied Question Answering (EQA) where nuanced evaluation of agents' spatial, temporal, and logical understanding is critical yet not considered by generic approaches. We introduce EQA-RM, a novel generative multimodal reward model specifically architected for EQA, trained via our innovative **Contrastive Group Relative Policy Optimization** (C-GRPO) strategy to learn fine-grained behavioral distinctions. The generative nature of EQA-RM provides interpretable, structured reward feedback (beyond simple scalars), uniquely enabling test-time scaling to dynamically adjust evaluation granularity, from concise scores to detailed critiques of reasoning and grounding, at inference without retraining. Concurrently, we introduce EQARewardBench, a new benchmark built on OpenEQA for standardized EQA reward model assessment. Demonstrating high sample efficiency, EQA-RM (fine-tuning Qwen2-VL-2B-Instruct) achieves **61.9%** accuracy on EQA-RM-Bench with **700** samples, outperforming strong proprietary baselines, including Gemini-2.5-Flash, GPT-4o, Claude-3.5-Haiku, and open-sourced state-of-the-art models such as RoVRM and VisualPRM. The code and dataset can be found here: https://github.com/UNITES-Lab/EQA-RM.

1 Introduction

Reward Models (RMs) have emerged as a cornerstone technique in qualifying the quality of output or actions of Large Language Models. Desired RMs provide critical signals for refining model behaviors and enhancing performance, often through reinforcement learning or selection strategies (Ouyang et al., 2022b; Snell et al., 2024). Existing generic Reward Models (Zhao et al., 2025; Liu et al., 2025) often designed for static inputs or simple outcomes, are proved ill-equipped for dynamic and interactive domains. Their limitations is amplified in complex tasks such as Embodied Question Answering (EQA). EQA requires agents to perceive, interact, and reason through sequences of multimodal observations and actions to answer questions within 3D environments (Majumdar et al., 2024; Yu et al., 2019). Evaluating the resulting EQA trajectories necessitates a nuanced assessment of the agent's reasoning coherence, action appropriateness, and how well language is grounded within the environment. Yet, existing RMs are usually trained via generic next-token prediction (Zhao et al., 2025) or classification (Faal et al., 2023), thus incapable to capture the spatio-temporal and logical dependencies inherent in embodied tasks. This emphasizes the urgent need for specialized mechanisms to accurately assess EQA's multifaceted success indicators.

To address this, we propose EQA-RM, a novel multimodal reward model tailored for evaluating EQA trajectories. Inspired by Zhao et al. (2025), EQA-RM is designed as a generative reward model (GenRM). EQA-RM thus is able to produce not only scalar rewards but also explicit reasoning for its assessments. Such generative capability enhances interpretability and facilitate test-time scaling (Snell et al., 2024) for performance improvements. To effectively generate these nuanced evaluations for EQA, the EQA-RM is trained for enhanced spatial, temporal, and reasoning processing capabilities, enabling it to handle the unique multimodal data streams inherent in EQA tasks.

We train EQA-RM using an efficient two-stage process. The first stage involves standard Rejective Finetuning (RFT) (Yuan et al., 2023) to teach the model the desired output format, which includes a textual critique and a scalar score. However, we observed that relying solely on RFT tends to make the model proficient at generating outputs in the learned format without necessarily achieving a deep understanding of crucial trajectory el-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 928–946
November 4-9, 2025

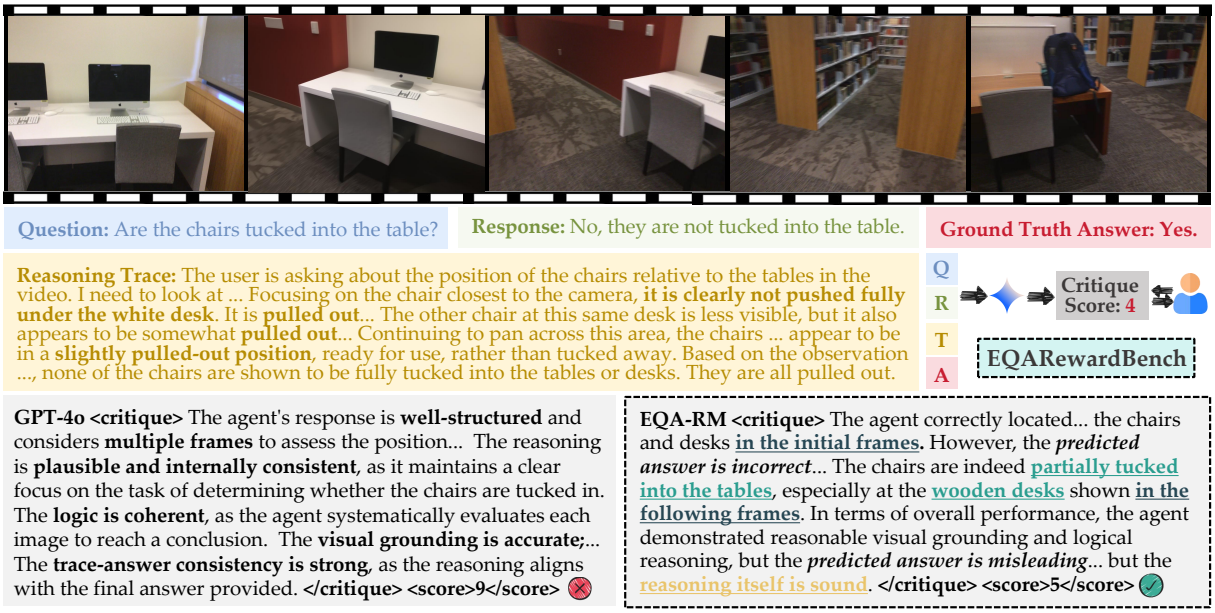

Figure 1: EQAREWARDBENCH enables nuanced evaluation of complex EQA trajectories involving spatio-temporal reasoning. An agent's reasoning process and response to a question are assessed through detailed critiques and scores provided by reward models like our EQA-RM and GPT-4o. Within its critiques, EQA-RM explicitly assesses temporal dependencies , spatial grounding , and reasoning coherence , showcasing better results than GPT-4o.

ements. This potential for supervised finetuning to primarily capture style rather than substance in preference learning has been discussed in prior work (Chen et al., 2025a; Liu et al., 2025). To address this intrinsic limitation, we employ a reinforcement learning approach for the second stage, which helps the model internalize output quality by optimizing for rewards, overcoming SFT's stylistic memorization. Specifically, we design a novel Contrastive Group Relative Policy Optimization (C-GRPO) strategy. C-GRPO utilizes rule-based contrastive rewards derived from targeted data augmentations. It optimizes EQA-RM's evaluative acuity by training it to distinguish policy outputs based on their evaluation under original, coherent contexts versus synthetically perturbed ones. We design the following augmentations for perturbations: ❶ trajectories with shuffled video frames, ❷ frames with randomly masked spatial regions, and ❸ sequences with jumbled reasoning steps. Essentially, EQA-RM earns a positive reward only when its score assigned under original, unperturbed conditions is more accurate (i.e., closer to a preference score) than its score assigned under the corresponding perturbed conditions. By learning this differentiated accuracy in its scoring relative to preference score, C-GRPO compels EQA-RM to internalize the importance of temporal order, fine-grained spatial details, and coherent logical flow. This cultivates a robust and discerning evaluative capability for embodied tasks.

On the other hand, the EQA domain lacks standardized benchmarks for rigorously evaluating and comparing reward models. Current EQA task benchmarks focus on coarse success metrics rather than the fine-grained trajectory quality crucial for RM development. To bridge this gap, we introduce EQAREWARDBENCH. Built upon the OpenEQA dataset, EQAREWARDBENCH features embodied episode memory videos from two types of household environments: HM3D () and ScanNet (). From the original question-answer pairs, we construct more comprehensive question-response-reasoning trace triplets. The benchmark includes 1,546 test instances designed to evaluate reward models across eight distinct aspects of trajectory quality (e.g., correctness, grounding, efficiency). EQAREWARDBENCH thus provides a standardized platform for the rigorous and comparable assessment of reward models on EQA tasks.

Empirical evaluations demonstrate the capabilities of EQA-RM. Based on Qwen2-VL-2B-Instruct, EQA-RM substantially improves upon its base model and outperforms existing open-source visual reward models on EQAREWARDBENCH. EQA-RM also shows test-time scalability. By increasing evaluative computations at inference, its accuracy on EQAREWARDBENCH is boosted from 42.47% to 61.86%. This enhanced performance allows EQA-RM to surpass leading large commercial models in accuracy on our benchmark.

In conclusion, our core contributions are: (*i*) We

propose EQA-RM, a generative multimodal reward model with enhanced spatial, temporal, and reasoning capabilities tailored for EQA; (***ii***) We introduce a EQA Reward Model Benchmark, EQAREWARDBENCH, the first dedicated benchmark for standardized evaluation of reward models for EQA. (***iii***) Extensive experiments have proved the improvement of our method for test time scaling.

2 Related Work

Generative Reward Models. Reward Models are crucial for guiding LLMs outputs (Ouyang et al., 2022a; Christiano et al., 2017; Ziegler et al., 2019). To overcome the limitations of scalar rewards and provide deeper evaluative insights, Generative Reward Models emerged and produced rich, interpretable textual feedback like critiques or explanations (Zhang et al., 2024; Mahan et al., 2024; Zheng et al., 2023). LLM-as-a-judge method accommodates pairwise critiques to evaluate LLMs outputs and enhances interpretability (Zheng et al., 2023). In multimodal domains, GenRMs enable finer-grained supervision through techniques such as step-wise reasoning assessment (Wang et al., 2025c), self-critique (Yu et al., 2025), or Chain-of-Thought style evaluations(Zhao et al., 2025), enhancing the performance of Multimodal LLMs.

Embodied Question Answering. EQA challenges agents to perceive, navigate, and interact within 3D environments to answer questions, integrating vision, language, reasoning, and planning (Das et al., 2018; Gordon et al., 2018). While datasets like OpenEQA offer rich scenarios for this task (Wijmans et al., 2019; Kolve et al., 2017; Majumdar et al., 2024) and EQA methodologies have advanced towards end-to-end foundation models (Ahn et al., 2022; Driess et al., 2023), current EQA evaluation predominantly assesses only final answer correctness (Majumdar et al., 2024; Das et al., 2018). This common practice overlooks crucial trajectory qualities such as reasoning coherence and spatio-temporal understanding (Chen et al., 2025b), creating a significant gap in reward modeling for comprehensive EQA assessment, motivating our development of the specialized EQA-RM.

Rule-based RL for LLMs. Reinforcement learning is increasingly used to refine LLMs for enhanced alignment and reasoning capabilities, moving beyond standard SFT (Schulman et al., 2017; Ouyang et al., 2022a; Zhai et al., 2024). A key direction involves rule-based RL, which employs systematic or synthetic feedback for improved efficiency and targeted behavior control, with algorithms like Group Relative Policy Optimization (GRPO) (Shao et al.) notably using relative comparisons and rule-defined rewards for complex reasoning (Mu et al., 2024; Xie et al., 2025; Wang et al., 2025b; Xiong et al., 2025). These structured RL principles are proving vital for training advanced reward models capable of nuanced multimodal understanding and for enabling RM self-improvement with systematic feedback (Feng et al., 2025; Liu et al., 2025). Our C-GRPO builds on these trends, utilizing rule-based contrastive rewards from data augmentations to train a generative RM specific for EQA tasks.

3 The EQAREWARDBENCH Dataset

To facilitate robust and standardized evaluation of reward models for Embodied Question Answering, we construct a new dataset EQAREWARDBENCH. This section details its generation pipeline, statistics, and splitting strategy.

3.1 Dataset Generation Pipeline

Our dataset construction process builds upon the OpenEQA (Majumdar et al., 2024), which provides instances comprising a question, ground truth answer, and associated episode memory. We extend this foundation in a two-step generation pipeline.

Step 1: Diverse Response Generation. We first employ the Gemini-2.0-Flash (Team et al., 2023) model to generate a diverse set of responses for each EQA instance. Given the episode memory (v^o) and the original question (q) as input, this model produces multiple pairs of predicted answers (a) and accompanying reasoning traces (z^o). To ensure a rich dataset for subsequent reward modeling, we intentionally solicit a spectrum of predicted answers, encompassing both correct and incorrect responses, thereby fostering diversity.

Step 2: High-Quality Score Generation. Next, Gemini-2.5-Pro-Experiment-03-14 generates detailed evaluations. For each Step 1 output tuple of {episode memory (v^o), question (q), predicted answer (a), reasoning trace (z^o)}, this model outputs a textual critique (c_r) and a scalar quality score. This score is an integer ranging from 0 to 10, indicating the quality of the predicted answer and reasoning trace. The ground truth answer input at this stage ensures accurate scores. After human verifica-

tion for reliability, these scores become our ground truth scores (s^{gt}). The accompanying critiques (c_r), potentially influenced by this privileged input, are primarily for analysis and contextualizing s_{gt}, not for directly training reward models that must operate without such information during inference. This process yields a foundational dataset where each instance is a tuple: $\{q, a, v^o, z^o, c_r, s^{gt}\}$, forming the basis for both EQAREWARDBENCH and the `EQA-RM` fine-tuning data.

3.2 Dataset Statistics and Splits

The episode memories (v_o) in our foundational dataset originate from two distinct 3D indoor environment collections provided by OpenEQA: HM3D (Ramakrishnan et al., 2021) and ScanNet (Dai et al., 2017). These sources contribute 697 and 1,546 instances in our dataset, respectively.

To ensure robust model training, fair evaluation, and prevention of test set leakage, we partition this foundational dataset: ❶ **EQAREWARDBENCH** (D_B): designed for evaluating diverse reward models, D_B comprises all 823 HM3D instances and 713 ScanNet instances. ❷ **Fine-tuning Dataset** (D_F): the remaining 697 distinct ScanNet instances are exclusively reserved for training our `EQA-RM` model.

Crucially, D_F and D_B maintain no overlap in underlying episode memory data (e.g., distinct scenes or trajectories), guaranteeing evaluation integrity. This splitting strategy enables comprehensive assessment: models trained on D_F (ScanNet) are evaluated against the disjoint ScanNet portion of D_B for in-distribution (ID) performance, and against its HM3D portion for out-of-distribution (OOD) generalization. Further details on dataset composition are provided in Appendix A.

4 Methods

This section details the methodology for training `EQA-RM` for nuanced evaluation of Embodied Question Answering trajectories. Our approach comprises two main stages: Rejective Fine-Tuning (RFT) to establish baseline capabilities in generating structured critiques and scores; followed by our novel Contrastive Group Relative Policy Optimization (`C-GRPO`) strategy to instill deeper sensitivities to critical aspects of trajectory quality.

4.1 Preliminaries and Notations

We denote the generative reward model we aim to train as R_ϕ. An instance consists of an input question q, a predicted answer a, an original reasoning trace $z^o = \{z_1^o, z_2^o, \ldots, z_T^o\}$, and the original episode memory content v^o. The episode memory v^o comprises a sequence of N video frames, $v^o = \{v_1^o, v_2^o, \ldots, v_N^o\}$. The desired output of R_ϕ for a given instance is an evaluation composed of a textual critique c_r and a scalar score s_r. For each evaluated EQA instance, there is an associated ground truth score, denoted s^{gt}.

4.2 Rejective Fine-Tuning

After obtaining the foundational dataset (Section 3.1), which includes ground truth scores (s^{gt}) for each EQA instance $\{q_i, a_i, z_i^o, v_i^o\}$, we initiate the training of `EQA-RM` (R_ϕ) with Rejective Fine-Tuning (RFT). This first stage primarily aims to teach R_ϕ to generate outputs (a textual critique c_r and a scalar score s_r) that conform to our desired structured format and exhibit a baseline level of quality. The RFT process involves two key steps:

Step 1: Rejective Filtering. To construct the SFT dataset D_{RFT}, for each instance $\{q_i, a_i, z_i^o, v_i^o\}$ from finetuning dataset D_F, an auxiliary LLM evaluator R^{aux} generates N_{RFT} candidate evaluations. Each candidate evaluation consists of a critique $c_{i,k}^{aux}$ and a score $s_{i,k}^{aux}$. Importantly, $\{c_{i,k}^{aux}, s_{i,k}^{aux}\}$ are produced based only on the input $\{q_i, a_i, z_i^o, v_i^o\}$ without ground truth answer. To ensure data quality for SFT, these candidate evaluations are rejective filtered.

This rejective filtering process aims to remove "too easy" and "incorrect" candidate evaluations.

An evaluation is considered correct ($\epsilon_{i,k} = 1$) if $|s_i^{gt} - s_{i,k}^{aux}| < \tau$ (τ is score tolerance), and $\epsilon_{i,k} = 0$ otherwise. Let $E_i = \sum_{k=1}^{N_{RFT}} \epsilon_{i,k}$ be the count of the correct evaluation of instance i. We select evaluations if:

$$\epsilon_{i,k} \wedge (\neg(E_i = N_{RFT} \wedge N_{RFT} > 0)) \quad (1)$$

This clause ensures an evaluation is included in D_{RFT} only if it is "correct" (per $\epsilon_{i,k} = 1$) and not all N_{RFT} evaluations for instance i were "correct" ("too easy"). Finally, D_{RFT} comprising $\{q_i, a_i, z_i^o, v_i^o\}$ paired with critique and score $\{c_i, s_i\}$ that satisfy this rejective filtering process, is subsequently used for the initial SFT of R_ϕ.

Step 2: Supervised Fine-Tuning. The curated dataset D_{RFT} is employed for SFT of our reward

model EQA-RM R_ϕ. For each sample in D_{RFT}, an input EQA instance comprising (q_i, a_i, z_i^o, v_i^o) is used to construct a multimodal prompt P_i, which also incorporates task-specific instructions. The corresponding output T_i for R_ϕ is the structured string representation of the selected critique c_i and score s_i from D_{RFT}. The SFT objective is to train R_ϕ by minimizing the negative log-likelihood loss $\mathcal{L}_{SFT}(\phi)$:

$$\mathcal{L}_{SFT}(\phi) = -\sum_{\substack{(P_i,T_i)\\ \in D_{RFT}}} \sum_{t=1}^{|T_i|} \log P(T_{i,t}|P_i, T_{i,<t}; \phi) \quad (2)$$

where $T_{i,t}$ is the t-th token of the target sequence T_i, and $T_{i,<t}$ represents the sequence of preceding tokens. This SFT stage primarily teaches R_ϕ to generate outputs in the specified format and establishes its baseline evaluative capability.

4.3 Contrastive Group Relative Policy Optimization

While SFT establishes EQA-RM's basic output formatting (critique c_r, score s_r), it often fails to instill deep comprehension of episode memory and reasoning from the limited finetuning data D_F. Our **Contrastive Group Relative Policy Optimization (C-GRPO)** framework addresses this by using targeted contrastive rewards to explicitly train R_ϕ for crucial sensitivities: temporal visual ordering, fine-grained spatial details, and logical coherence of reasoning.

C-GRPO trains R_ϕ to differentiate its evaluation scores based on structured contrasts between original EQA instance components (q_i, a_i, z_i^o, v_i^o) and their synthetically perturbed counterparts. These perturbations include temporally shuffled video frames (v_i^t from v_i^o), spatially masked frames (v_i^s from v_i^o), and altered reasoning traces (z_i^r from z_i^o). R_ϕ produces scores for the original condition (S_i^o) and for each corresponding augmented condition (S_i^t, S_i^s, S_i^r). For S_i^t and S_i^s, R_ϕ scores the original (q_i, a_i, z_i^o) but is prompted as if the visual context were v_i^t or v_i^s respectively. For S_i^r, R_ϕ scores (q_i, a_i) paired with the perturbed reasoning trace z_i^r while using the original visual context v_i^o.

Base Outcome Rewards. each evaluation score $s_{r,i}$ are assessed by the **Accuracy Reward** ($R_{a,i}$):

$$R_{a,i}(s_{r,i}, s_{gt,i}) = \max(0, 1 - (\frac{|s_{r,k} - s_{gt,i}|}{10})^2) \quad (3)$$

and **Format Reward** (R_f) is 1 if text output adheres to the specified critique-score structure, and 0 otherwise.

C-GRPO Contrastive Rewards. For an x-augmented version ($x \in \{t, s, r\}$ for temporal, spatial, reasoning respectively), let $R_{a,i}^x$ be the corresponding accuracy reward. This mechanism conditionally boosts the $R_{a,i}^o$ for evaluations of original instances. For each active augmentation type $x \in \{t, s, r\}$: Let $\overline{R_a^o}$ and $\overline{R_a^x}$ be the batch-mean accuracy rewards (Eq. 3) for original and x-augmented evaluations, respectively. The per-evaluation boost R^x for the i-th original evaluation is then defined as:

$$R_i^x = \begin{cases} \mu, & \text{if } \overline{R_a^o} \geq \delta \cdot \overline{R_a^x} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where $\delta = 0.95$ and $\mu = 0.3$ are hyperparameters. This yields boosts R^t, R^s, R^r for active contrasts.

Total Reward. The total reward is:

$$R_i^A = R_a^o + R_f^o + (R^t + R^s + R^r)/3 \quad (5)$$

Advantage and Policy Update. The advantage A_i for the i-th evaluation (from G total evaluations) is computed by normalizing the total rewards $\{R_j^A\}_{j=1}^G$ within the group:

$$A_i = \frac{R_i^A - \text{mean}(\{R_j^A\}_{j=2}^G)}{\text{std}(\{R_j^A\}_{j=1}^G) + \epsilon} \quad (6)$$

where ϵ is a small constant for numerical stability. R_ϕ are updated follows the objective of C-GRPO:

$$\mathcal{J}_{\text{C-GRPO}}(\phi) = \mathbb{E}_{d_i \sim D_F}\Bigg[\frac{1}{G}\sum_{k=1}^{G}\Bigg(\min\Big(r_k(\phi)A_k, \text{clip}(r_k(\phi), 1-\epsilon_c, 1+\epsilon_c)A_k\Big) - \beta_K\, \mathbb{D}_{\text{KL}}(P_\phi \| P_{\phi_{\text{ref}}})\Bigg)\Bigg] \quad (7)$$

where $r_k(\phi)$ is the probability ratio, ϕ_{ref} are parameters before the update, ϵ_c is a clipping hyperparameter, and $R_{\phi_{\text{ref}}}$ is a reference model (typically the SFT version of R_ϕ).

5 Experiments

In this section, our goal is to assess the effectiveness of EQA-RM and answer the following questions:

- **RQ1**: How does EQA-RM perform compared to existing Visual-based Reward Models?
- **RQ2**: How does the performance of EQA-RM scale with increased test-time compute?
- **RQ3**: What is the impact of each component of the proposed reward strategy on EQARewardBench's performance?

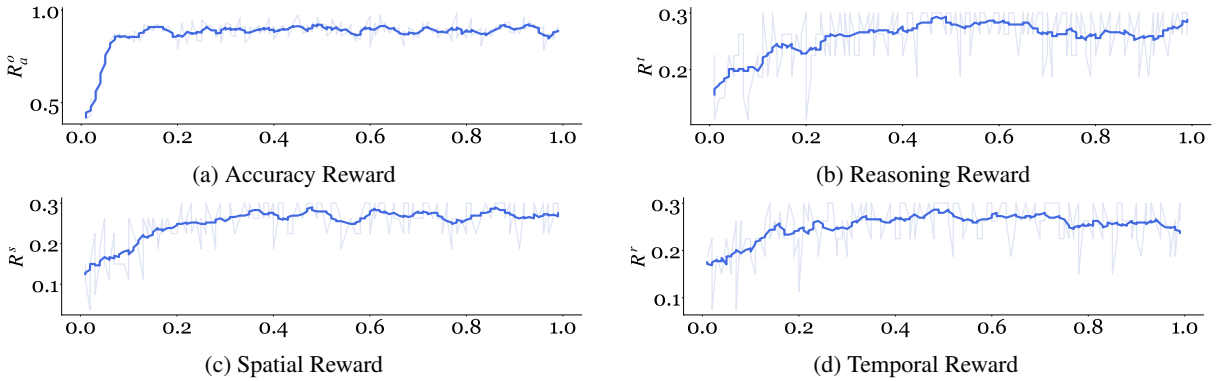

Figure 2: Overview of training dynamics for EQA-RM's core reward components of C-GRPO.

Table 1: Main evaluation results on the EQARewardBench dataset. We report Accuracy (Acc) and RMSE for each method across two environments (HM3D and ScanNet), and their aggregated Overall performance. All results are shown with differences relative to Qwen2-VL-2B-Instruct model (Better is marked in red; worse is marked in blue).

Models	EQARewardBench-HM3D		EQARewardBench-ScanNet		Overall	
	Acc ↑	RMSE ↓	Acc ↑	RMSE ↓	Acc ↑	RMSE ↓
VLM-as-a-Judge						
Qwen2-VL-2B-Instruct	33.71	3.836	32.44	4.158	33.08	3.997
Gemini-2.5-Flash	**61.25**$_{\uparrow 27.54}$	4.872$_{\uparrow 1.036}$	58.33$_{\uparrow 25.89}$	3.805$_{\downarrow 0.353}$	59.79$_{\uparrow 26.71}$	4.339$_{\uparrow 0.342}$
GPT-4o	60.17$_{\uparrow 26.46}$	6.847$_{\uparrow 3.011}$	56.72$_{\uparrow 24.28}$	5.313$_{\uparrow 1.155}$	58.45$_{\uparrow 25.37}$	6.080$_{\uparrow 2.083}$
Claude-3.5-Haiku	57.88$_{\uparrow 24.17}$	5.316$_{\uparrow 1.480}$	54.12$_{\uparrow 21.68}$	4.291$_{\uparrow 0.133}$	56.00$_{\uparrow 22.92}$	4.804$_{\uparrow 0.807}$
Standard Visual-Based Reward Models						
RoVRM	38.12$_{\uparrow 4.41}$	3.341$_{\downarrow 0.495}$	40.23$_{\uparrow 7.79}$	3.158$_{\downarrow 1.000}$	39.18$_{\uparrow 6.10}$	3.250$_{\downarrow 0.747}$
VisualPRM	40.07$_{\uparrow 6.36}$	3.562$_{\downarrow 0.274}$	37.41$_{\uparrow 4.97}$	3.413$_{\downarrow 0.745}$	38.74$_{\uparrow 5.66}$	3.488$_{\downarrow 0.509}$
Proposed Method						
base EQA-RM	39.16$_{\uparrow 5.45}$	3.081$_{\downarrow 0.755}$	45.78$_{\uparrow 13.34}$	2.826$_{\downarrow 1.332}$	42.47$_{\uparrow 9.39}$	2.953$_{\downarrow 1.044}$
w/ test scaling (K=32)	58.65$_{\uparrow 24.94}$	**2.918**$_{\downarrow 0.918}$	**65.06**$_{\uparrow 32.62}$	**2.537**$_{\downarrow 1.621}$	**61.86**$_{\uparrow 28.78}$	**2.727**$_{\downarrow 1.270}$

5.1 Experiment Setup

Dataset and Benchmark. All experimental evaluations are conducted using our EQAREWARDBENCH (D_B). The methodology behind it is detailed in Section 3. The evaluation focuses on the Verifier mode, assessing the RM's capability to accurately score pre-generated policy responses based on the offline-annotated ground truth scores.

Base Model. Our base model is Qwen2-VL-2B-Instruct (Wang et al., 2024), instruction-tuned for robust multimodal understanding.

Baselines. We compare EQA-RM against a comprehensive set of baseline methods to establish its relative effectiveness: **(1) VLM-as-a-Judge:** State-of-the-art VLMs prompted offline to evaluate EQA responses on our test set. Including Gemini-2.5-Flash (Team et al., 2023), GPT-4o (Hurst et al., 2024) and Claude-3.5-Haiku. **(2) Generic Standard RMs:** Adapt existing VLMs or Visual RMs including RoVRM (Wang et al., 2025a), VisualPRM (Wang et al., 2025c).

Evaluation Metrics. We evaluate the performance of all reward models using the following metrics.

- **Accuracy** measures the proportion of predictions where the gap between the predicted score s^p and the target score s^{gt} is within a predefined tolerance τ. We use the tolerance $\tau = 2$ in the experiments:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(|s_i^p - s_i^{gt}| \leq \tau) \tag{8}$$

- **Root Mean Square Error** quantifies the average magnitude of the error between predicted and target scores:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_i^p - s_i^{gt})^2} \tag{9}$$

Implementation Details. Training utilized the AdamW optimizer with a learning rate of $1e-6$ and a batch size of 1. Keyframe selection during

Table 2: Accuracy performance on the EQARewardBench-Scannet dataset, broken down by EQA question type. Higher is better for all metrics. All results are shown with differences relative to the Qwen2-VL-2B-Instruct model (Better is marked in red; Worse is marked in blue).

Models	Object Recognition	Object Localization	Attribute Recognition	Spatial Understanding	Object State Recognition	Functional Reasoning	World Knowledge
VLM-as-a-Judge							
Qwen2-VL-2B-Instruct	30.30	36.76	38.69	29.41	29.88	38.71	27.34
Gemini-2.5-Flash	**69.23**$_{\uparrow 38.93}$	62.15$_{\uparrow 25.39}$	61.07$_{\uparrow 22.38}$	38.92$_{\uparrow 9.51}$	**51.88**$_{\uparrow 22.00}$	64.05$_{\uparrow 25.34}$	**61.51**$_{\uparrow 34.17}$
GPT-4o	66.39$_{\uparrow 36.09}$	60.00$_{\uparrow 23.24}$	59.84$_{\uparrow 21.15}$	**41.51**$_{\uparrow 12.10}$	50.00$_{\uparrow 20.12}$	57.39$_{\uparrow 18.68}$	60.68$_{\uparrow 33.34}$
Claude-3.5-Haiku	60.15$_{\uparrow 29.85}$	56.76$_{\uparrow 20.00}$	64.22$_{\uparrow 25.53}$	35.83$_{\uparrow 6.42}$	43.51$_{\uparrow 13.63}$	53.88$_{\uparrow 15.17}$	60.11$_{\uparrow 32.77}$
Standard Visual-Based Reward Models							
RoVRM	48.90$_{\uparrow 18.60}$	50.13$_{\uparrow 13.37}$	47.82$_{\uparrow 9.13}$	25.91$_{\downarrow 3.50}$	36.50$_{\uparrow 6.62}$	44.69$_{\uparrow 5.98}$	27.38$_{\uparrow 0.04}$
VisualPRM	29.60$_{\downarrow 1.30}$	42.90$_{\uparrow 6.14}$	43.62$_{\uparrow 4.93}$	39.21$_{\uparrow 9.80}$	33.93$_{\uparrow 4.05}$	41.33$_{\uparrow 2.62}$	29.34$_{\uparrow 2.00}$
Proposed Method							
base EQA-RM	36.36$_{\uparrow 6.06}$	50.81$_{\uparrow 14.05}$	51.82$_{\uparrow 13.13}$	26.05$_{\downarrow 3.36}$	21.95$_{\downarrow 7.93}$	45.97$_{\uparrow 7.26}$	35.94$_{\uparrow 8.60}$
w/ test scaling (K=32)	68.18$_{\uparrow 37.88}$	**69.73**$_{\uparrow 32.97}$	**71.53**$_{\uparrow 32.84}$	37.82$_{\uparrow 8.41}$	37.20$_{\uparrow 7.32}$	**66.94**$_{\uparrow 28.23}$	57.81$_{\uparrow 30.47}$

training and test to select $N = 5$ frames for each episode history video. Further details are provided in the Appendix B.

Test-time Scaling Configuration. To enhance EQA-RM's evaluation robustness, our test-time scaling (TTS) strategy involves sampling K diverse evaluative reasoning paths from a given EQA trajectory. This is facilitated by a temperature setting of 0.8 and top-p sampling with $p = 0.9$. We explore K values of 1 (no TTS), 2, 4, 8, 16, and 32. These K assessments are aggregated using either Majority Voting or Averaging Rewards for continuous scores. This TTS approach enables EQA-RM to synthesize multiple perspectives, aiming for a more reliable and comprehensive trajectory assessment by mitigating single-inference pass limitations.

5.2 Main Results

The overall performance of various reward models on the EQARewardBench dataset is presented in Table 1. Our proposed EQA-RM, fine-tuned from Qwen2-VL-2B-Instruct, is benchmarked against its base model, other VLM-as-a-Judge methods, and Standard Visual-Based Reward Models. EQA-RM demonstrates substantial gains over its base Qwen2-VL-2B-Instruct model, for instance, improving overall accuracy by over 9% while also achieving a markedly lower RMSE. It also consistently outperforms standard visual-based RMs like RoVRM and VisualPRM in both overall accuracy and RMSE, showing the advantage of our specialized training.

While the base EQA-RM's accuracy is lower than leading VLM-as-a-Judge models such as GPT-4o and Gemini-2.5-Flash, it provides more precise reward signals, evidenced by its considerably lower RMSE. Crucially, with test-time scaling (K=32), EQA-RM's accuracy significantly increases to an overall score of approximately 61.9%, surpassing these top VLM judges and achieving the best RMSE. This positions the scaled EQA-RM as the top-performing model on our benchmark.

5.3 Test-time Scalability

The benefits of test-time scaling for EQA-RM are demonstrated in Figure 3. As the number of sampled critiques (K) increases from 1 to 32, EQA-RM exhibits a strong and consistent improvement in performance. Overall accuracy sees substantial gains. For example, increasing by over 19% on both ScanNet and HM3D datasets, alongside a steady decrease in RMSE, indicating enhanced prediction quality. This robust scalability contrasts sharply with the base Qwen2-VL-2B-Instruct model. Although the latter shows test-time scaling on RMSE, But it does not exhibit a positive trend in terms of accuracy with increasing K. This highlights that the test-time scaling advantage is a specific outcome of our training methodology for EQA-RM.

5.4 Ablation Studies

Table 3 presents our ablation study on EQAREWARDBENCH, reporting accuracy and RMSE across its HM3D and ScanNet subsets.

Training Stage Analysis. Part A indicates that RFT alone slightly reduces overall accuracy (-1.39%) while marginally improving RMSE compared to the base model. This is likely because RFT, with its limited samples, primarily focuses on teaching output formatting, which, if used in isolation, may not enhance broader understanding.

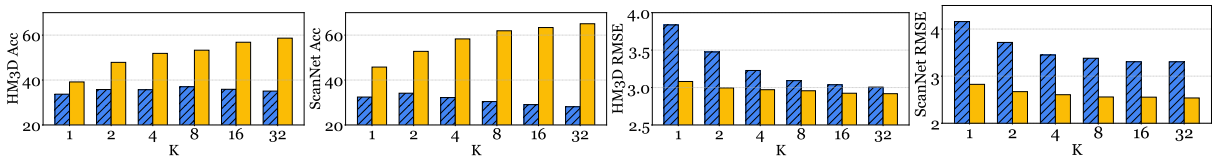

Figure 3: Effect of test-time scaling (varying K) on EQA-RM performance, comparing EQA-RM and Qwen2-VL-2B-Instruct on HM3D and ScanNet datasets.

Table 3: Ablation study on core components and reward formulations of EQA-RM. R^b denotes the base accuracy and format reward. R^t, R^s, and R^r represent the contrastive temporal, spatial and reasoning reward, respectively (defined in Section 4.3). Configurations in part B are subsequent to the RFT stage and detail the specific reward formula optimized by C-GRPO. All results are shown with differences relative to the Base Model.

Method Configuration	EQARewardBench-HM3D		EQARewardBench-ScanNet		Overall	
	Acc (%) ↑	RMSE ↓	Acc (%) ↑	RMSE ↓	Acc (%) ↑	RMSE ↓
A. Core Training Stage and Architectural Ablations:						
Base Model (Qwen2-VL-2B-Instruct)	33.71	3.836	32.44	4.158	33.08	3.997
RFT only	32.14$_{\downarrow 1.57}$	3.781$_{\downarrow 0.055}$	31.24$_{\downarrow 1.20}$	3.879$_{\downarrow 0.279}$	31.69$_{\downarrow 1.39}$	3.830$_{\downarrow 0.167}$
B. Specific Reward Formulations (without RFT):						
RL (R_b only)	29.12$_{\downarrow 4.59}$	4.132$_{\uparrow 0.296}$	30.54$_{\downarrow 1.90}$	4.035$_{\downarrow 0.123}$	29.83$_{\downarrow 3.25}$	4.084$_{\uparrow 0.087}$
RL ($R_b + R_t$) (+ Temporal)	31.55$_{\downarrow 2.16}$	4.027$_{\uparrow 0.191}$	32.85$_{\uparrow 0.41}$	3.953$_{\downarrow 0.205}$	32.20$_{\downarrow 0.88}$	3.990$_{\downarrow 0.007}$
RL ($R_b + R_s$) (+ Spatial)	28.95$_{\downarrow 4.76}$	3.759$_{\downarrow 0.077}$	30.15$_{\downarrow 2.29}$	3.788$_{\downarrow 0.370}$	29.55$_{\downarrow 3.53}$	3.774$_{\downarrow 0.223}$
RL ($R_b + R_r$) (+ Reasoning)	32.05$_{\downarrow 1.66}$	3.801$_{\downarrow 0.035}$	33.57$_{\uparrow 1.13}$	3.882$_{\downarrow 0.276}$	32.81$_{\downarrow 0.27}$	3.842$_{\downarrow 0.155}$
RL ($R_b + R_t + R_s + R_r$)	31.24$_{\downarrow 2.47}$	3.812$_{\downarrow 0.024}$	32.94$_{\uparrow 0.50}$	3.935$_{\downarrow 0.223}$	32.09$_{\downarrow 0.99}$	3.874$_{\downarrow 0.123}$
C. Specific Reward Formulations (with RFT):						
RFT + RL (R_b only)	33.50$_{\downarrow 0.21}$	3.895$_{\uparrow 0.059}$	31.80$_{\downarrow 0.64}$	4.189$_{\uparrow 0.031}$	32.65$_{\downarrow 0.43}$	4.042$_{\uparrow 0.045}$
RFT + RL ($R_b + R_t$) (+ Temporal)	32.95$_{\downarrow 0.76}$	3.750$_{\downarrow 0.086}$	35.20$_{\uparrow 2.76}$	4.210$_{\uparrow 0.052}$	34.08$_{\uparrow 1.00}$	3.980$_{\downarrow 0.017}$
RFT + RL ($R_b + R_s$) (+ Spatial)	30.15$_{\downarrow 3.56}$	3.915$_{\uparrow 0.079}$	29.50$_{\downarrow 2.94}$	4.233$_{\uparrow 0.075}$	29.83$_{\downarrow 3.25}$	4.074$_{\uparrow 0.077}$
RFT + RL ($R_b + R_r$) (+ Reasoning)	39.30$_{\uparrow 5.59}$	3.152$_{\downarrow 0.684}$	45.50$_{\uparrow 13.06}$	2.984$_{\downarrow 1.174}$	42.40$_{\uparrow 9.32}$	3.068$_{\downarrow 0.929}$
RFT + RL ($R_b + R_t + R_s$) (+ Temporal + Spatial)	38.27$_{\uparrow 4.56}$	3.015$_{\downarrow 0.821}$	42.76$_{\uparrow 10.32}$	2.755$_{\downarrow 1.403}$	40.52$_{\uparrow 7.44}$	2.885$_{\downarrow 1.112}$
RFT + RL ($R_b + R_t + R_r$) (+ Temporal + Reasoning)	39.10$_{\uparrow 5.39}$	3.102$_{\downarrow 0.734}$	**45.80**$_{\uparrow 13.36}$	2.907$_{\downarrow 1.251}$	42.45$_{\uparrow 9.37}$	3.005$_{\downarrow 0.992}$
RFT + RL ($R_b + R_s + R_r$) (+ Spatial + Reasoning)	**39.45**$_{\uparrow 5.74}$	**2.976**$_{\downarrow 0.860}$	45.43$_{\uparrow 12.99}$	**2.714**$_{\downarrow 1.444}$	42.44$_{\uparrow 9.36}$	**2.845**$_{\downarrow 1.152}$
RFT + RL ($R_b + R_t + R_s + R_r$) (EQA-RM)	39.16$_{\uparrow 5.45}$	3.081$_{\downarrow 0.755}$	45.78$_{\uparrow 13.34}$	2.826$_{\downarrow 1.332}$	**42.47**$_{\uparrow 9.39}$	2.954$_{\downarrow 1.043}$

Reward Components. Ablating RFT pre-training (Part B) reveals that most subsequent C-GRPO reward formulations underperform the base model in accuracy, underscoring RFT's critical role for effective initialization. With RFT pre-training (Part C), the reasoning reward (R_r) alone yields the largest single-component accuracy increase (+9.32%). Conversely, the spatial reward (R_s), while detrimental in isolation (-3.25%), improves performance when combined with other components, suggesting a regularizing effect against overfitting to specific (e.g., temporal or reasoning) cues. These results affirm the efficacy of our contrastive reward design and the importance of RFT cold start for optimal performance.

5.5 Question Type Performance Analysis

Table 2 presents an accuracy breakdown by EQA question type on the EQAREWARDBENCH-SCANNET dataset. Our EQA-RM with test-time scaling (K=32) demonstrates notable strength, achieving top performance in Object Localization, Attribute Recognition, and Functional Reasoning. It also proves highly competitive in Object Recognition and World Knowledge against strong VLM-as-a-Judge baselines. This scaled version of EQA-RM significantly improves upon its base, surpassing several VLM judges in these key areas.

6 Conclusion

We introduced EQA-RM, a generative reward model tailored for nuanced evaluation of complex Embodied Question Answering (EQA) trajectories, alongside EQAREWARDBENCH, a dedicated benchmark for this task. Trained with our novel Contrastive Group Relative Policy Optimization (C-GRPO) strategy, EQA-RM learns to assess critical spatial, temporal, and reasoning understanding. Empirical results demonstrate EQA-RM's effectiveness and high sample efficiency, achieving 61.84% accuracy on EQAREWARDBENCH through test-time scaling, thereby outperforming strong proprietary and open-source baselines. This work presents a significant step towards robust reward modeling in embodied AI, offering tools and methodologies to foster more capable EQA agents.

Acknowledgement. This work is partially supported by Amazon Research Award, Cisco Faculty Award, UNC Accelerating AI Awards, NAIRR Pilot Award, OpenAI Researcher Access Award, and Gemma Academic Program GCP Credit Award.

Limitations and Future Work

While EQA-RM and our C-GRPO strategy demonstrate significant advancements in evaluating EQA trajectories, we acknowledge several limitations that also point towards avenues for future research.

Limitations. The current set of contrastive augmentations in C-GRPO, targeting temporal, spatial, and logical aspects, while effective, is predefined. These specific perturbations may not encompass the full spectrum of nuanced behaviors or subtle failure modes encountered in diverse EQA scenarios. Consequently, EQA-RM's sensitivities are primarily shaped by these explicit contrasts. Secondly, the efficacy of our two-stage training process relies on high-quality ground truth scores. In our work, these score values are derived using a powerful commercial large model (Gemini-2.5-Pro) conditioned on ground truth answers, followed by human verification. Any inherent biases, limitations, or the specific characteristics of this commercial model could be subtly reflected in the score values, thereby influencing the RFT filtering process and the accuracy reward component which guides C-GRPO. While EQAREWARDBENCH includes distinct in-distribution (ScanNet) and out-of-distribution (HM3D) splits based on OpenEQA environments, the broader generalization of EQA-RM to EQA tasks, visual styles, or interaction paradigms substantially different from those in OpenEQA remains an open question. Finally, although the generative critiques from EQA-RM enhance interpretability and enable test-time scaling, a systematic evaluation of their fine-grained faithfulness or their direct utility in, for example, few-shot learning for policy adaptation, was beyond the scope of this paper.

Future Work. Building on these limitations, we plan to explore more adaptive or learned augmentation strategies for C-GRPO to capture a wider array of desirable EQA agent behaviors, potentially including aspects like efficiency, safety, or interactivity. Investigating methods to generate or refine high-quality score values using open-source models or with more scalable human oversight would increase the accessibility and robustness of the dataset creation pipeline. A key direction is to systematically leverage the rich, structured critiques from EQA-RM not just for scoring, but also as direct feedback for improving EQA policy models, perhaps through distillation or critique-guided reinforcement learning. We also aim to expand EQAREWARDBENCH with greater diversity in tasks, environments, and possibly languages, to further support the community in developing more general and robust EQA evaluation methods.

Ethical Statement

The development of EQA-RM and EQAREWARDBENCH adhered to ethical research practices. Our benchmark is derived from publicly available datasets (OpenEQA, sourcing from HM3D and ScanNet), and the data we generated (answers, reasoning, critiques, scores) does not contain personally identifiable information. The commercial models used for data generation were accessed via their standard APIs under their terms of service. We acknowledge the potential for latent biases inherited from large pre-trained models (both those used for score generation and the base model for EQA-RM) and encourage ongoing research into bias detection and mitigation in reward modeling for embodied AI. While training these models is resource-intensive, we focused on a sample-efficient approach with a 2B parameter base model. We intend to release our benchmark and model to promote transparency, reproducibility, and further community research.

References

Michael Ahn, Anthony Brohan, Noah Brown, Aakanksha Chowdhery, Sizu Chua, Brian Cui, Hanjun Dabis, Chelsea Dean, Danny Driess, Fred Duke, Chelsea Finn, Chuyuan Fu, Sihan Gu, Karol Hausman, Brian Ichter, Kanishka Julian, Dmitry Kalashnikov, Kuang-Huei Kamyar, Keerthana Lee, Sergey Levine, Yao Li, Zhen Lin, Shiquan Liu, Yifan Lu, Linda Luu, Soroush Mahdavi, Sudeep Manyam, Michael Mazur, John McMahon, Debidatta Misra, Khem Nasihati, Michael Orefice, Jihyun Pan, Kathleen Peng, Emily Perez, Jeffrey Phillips, Raphael Quiambao, Khem Rahn, Kanishka Rao, Jose Retana, Pierre Reyes, Corban Rivera, John Rodriguez, America Sanchez, Robert Sievers, Sumeet Singh, Clayton Sofge, Austin Stone, Jonathan Tan, Mengyuan Tseng, Fei Tung, Martin Vecerik, Quan Vuong, Ayzaan Wahid, Ted Wang, Peng Xu, Muyang Yan, Alex Yu, Tianhe Yu, Brianna Yuan, Yue Zhang, Zhe Zhang, Tianli Zhou, Yifeng Zhu, Allen Zirbel, Peter Florence, Vincent Vanhoucke, Andy Zeng, Jonathan

Tompson, Igor Mordatch, Pierre Sermanet, Nikhil Kumar, Ken Caluwaerts, Ted Xiao, Aravind Rajeswaran, Ryan Brooks, Joshua Tobin, Laurens Van Der Maaten, Alexander Ku, Steven Hadfield, Jie Tan, Scott Collins, Thomas Gates, Anton Egorov, Jonathan Ho, Alex Irpan, and Mohi Khansari. 2022. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*, pages 100–116. PMLR.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. 2025a. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*.

Lichang Chen, Chen Zhu, Jiuhai Chen, Davit Soselia, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. 2025b. Reward models identify consistency, not causality. *arXiv preprint arXiv:2502.14619*.

Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16–25.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, pages 8469–8488. PMLR.

Farshid Faal, Ketra Schmitt, and Jia Yuan Yu. 2023. Reward modeling for mitigating toxicity in transformer-based language models. *Applied Intelligence*, 53(7):8421–8435.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4089–4098.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*.

Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. 2024. Generative reward models. *arXiv preprint arXiv:2410.12832*.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. 2024. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16488–16498.

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. 2024. Rule based rewards for language model safety. *arXiv preprint arXiv:2411.01111*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022a. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022b. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. 2021. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. *URL https://arxiv.org/abs/2402.03300*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

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*.

Chenglong Wang, Yang Gan, Yifu Huo, Yongyu Mu, Murun Yang, Qiaozhi He, Tong Xiao, Chunliang Zhang, Tongran Liu, and Jingbo Zhu. 2025a. Rovrm: A robust visual reward model optimized via auxiliary textual preference data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25336–25344.

Hu Wang, Congbo Ma, Ian Reid, and Mohammad Yaqub. 2025b. Kalman filter enhanced grpo for reinforcement learning-based language model reasoning. *arXiv preprint arXiv:2505.07527*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, Lewei Lu, Haodong Duan, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025c. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.

Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. 2019. Embodied question answering in photorealistic environments with point cloud perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9506–9515.

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*.

Wei Xiong, Jiarui Yao, Yuhui Xu, Bo Pang, Lei Wang, Doyen Sahoo, Junnan Li, Nan Jiang, Tong Zhang, Caiming Xiong, et al. 2025. A minimalist approach to llm reasoning: from rejection sampling to reinforce. *arXiv preprint arXiv:2504.11343*.

Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L Berg, and Dhruv Batra. 2019. Multi-target embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6309–6318.

Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chenguang Zhu, Richard Yuanzhe Pang, Yundi Qian, Xuewei Wang, Suchin Gururangan, Chao Zhang, Melanie Kambadur, Dhruv Mahajan, and Rui Hou. 2025. Self-generated critiques boost reward modeling for language models. *arXiv preprint arXiv:2411.16646*.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*.

Simon Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Peter Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, et al. 2024. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *Advances in neural information processing systems*, 37:110935–110971.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*.

Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, and Bowen Zhou. 2025. Genprm: Scaling test-time compute of process reward models via generative reasoning. *arXiv preprint arXiv:2504.00891*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Benchmark Statistic

Table 4 details the distribution of instances from the HM3D and ScanNet environments within these datasets.

Table 4: Dataset statistics for EQAREWARDBENCH and the Finetuning set.

Subset	EQAREWARDBENCH	Finetuning
HM3D	823	0
ScanNet	713	697
Total	1546	697

B Implementation Details

Table. 5 provides a comprehensive list of hyper-parameters and configuration settings used for the SFT and C-GRPO training stages of `EQA-RM`, as well as for test-time scaling.

C Cases Studies

This section presents qualitative examples to illustrate the evaluation capabilities of EQA-RM. The case studies showcase how EQA-RM assesses an agent's response, reasoning trace, and visual grounding based on a sequence of observed frames within an EQA task. These examples provide concrete instances of the nuanced feedback generated by our model.

D Generation Prompts

This section details the specific prompts used in the generation pipeline for creating the dataset for benchmark and the Rejective Fine-Tuning stage. The following pages display the prompt guidelines provided to the large language models for generating diverse responses, high-quality scores, and for formatting the output during the RFT data creation process.

Table 5: Key hyperparameters and configuration settings for EQA-RM.

Parameter Category & Parameter	**Value**
Base Model	Qwen2-VL-2B-Instruct
Attention Implementation	`flash_attention_2`
Keyframes per Episode (N)	5
Distributed Training Backend	DeepSpeed ZeRO Stage 3
Precision	BF16
Gradient Checkpointing	True
Max Gradient Norm (SFT & C-GRPO)	5.0
Optimizer	AdamW
B. SFT Stage	
Input Model	Qwen2-VL-2B-Instruct
Learning Rate	1×10^{-6}
Max Sequence Length	1024
Batch Size (per device $\times$ grad. accum.)	1×1
Number of Epochs	2
GPUs per Node	2
C. C-GRPO Stage	
Input Model	Output of SFT Stage
Learning Rate	1×10^{-6}
LR Scheduler	Cosine
Weight Decay	0.01
Max Prompt Length	8142
Max Completion Length ($\lvert e_k \rvert$)	768
Batch Size (per device $\times$ grad. accum.)	1×1
Number of Epochs	1
G (Generations per prompt by R_ϕ)	8
β_K (C-GRPO KL Coefficient)	0.04
Contrastive Mechanism (Eq. 4):	
Factor for batch-mean comparison (δ)	0.95
Boost value (μ)	0.3
Min. $R^o_{acc,k}$ for boost (H_{min_acc})	0.1
Spatial Masking Details:	
Mask Size	(16, 16)
Mask Ratio	0.15
Mask Value	0.0
Max Pixels (video processing)	401 408
GPUs per Node	8
D. Test-Time Scaling (TTS)	
Sampling Temperature	0.8
Top-p Sampling (p)	0.9
K (Number of Sampled Paths)	{1, 2, 4, 8, 16, 32}
Aggregation Method	Majority Voting and Averaging Rewards

Rejective Finetuning Dataset Generation Prompt Guideline

Role: You are an Expert EQA evaluator providing structured critique and score based on reasoning and visual grounding.

Context: You receive:

1. Video Frames
2. User Query
3. Agent's Response (reasoning_trace, predicted_answer)

Task: Evaluate the Agent's Response according to the weighted criteria specified below. Output a critique summarizing your findings.

Evaluation Criteria:

1. **Answer Plausibility & Visual Grounding (60% weight):**
    - Plausibility & Relevance: Assess if the predicted_answer is a plausible and relevant response to the user_query based on the provided video_frames. Does the answer make sense in the context of the question and what is visible?
    - Visual Confirmation: Can the core elements or claims of the predicted_answer be directly observed or reasonably inferred from the content of the video_frames? If the answer describes objects, actions, or states, are these visually verifiable in the frames?
    - Specificity: Is the predicted_answer appropriately specific given the user_query and the visual information available? Avoid being overly general if details are visible, or overly specific if not supported by frames.
2. **Reasoning Trace Quality (40% weight):**
    - Logic & Consistency: Is the reasoning_trace internally logical? Are there contradictions or significant logical gaps?
    - Visual Grounding: Do descriptive references in the trace (e.g., "first frame", "object shown") accurately correspond to content visible within the video frames provided?
    - Trace-Answer Consistency: Does the predicted_answer logically follow from the steps and conclusion presented in the reasoning_trace?

Output Structure: Provide your evaluation in the following format. Fill <critique> with your analysis covering the Evaluation Criteria above. Fill <score> with the final 1-10 weighted score.

```
<critique>
[Your analysis and summary covering the Evaluation Criteria]
</critique>
<score>
[Your final 1-10 weighted score of the response based on the critique. Do not always ouput extreme score (e.g., 10 and 0)]
</score>
```

BEGIN INPUT
Video Frames (attached)
User Query: {question}
Reasoning Trace: {reasoning_trace}
Predicted Answer: {predicted_answer}
BEGIN OUTPUT

Benchmark Dataset Generation Prompt Guideline

Role: You are an intelligent Embodied Question Answering (EQA) agent operating within a simulated indoor environment. You navigate, observe, and reason to answer user questions.

Task: Given a User Query about a simulated environment, generate a detailed response outlining your simulated process and final answer. You must strictly adhere to the output structure specified below. Output ONLY a <reasoning_trace> and <predicted_answer>. The trace must show a step-by-step fine-grained analysis grounded in visual frames.

Output Structure: Produce ONLY the following structured text, using separate <think> tags to break down your reasoning process into logical steps or phases:

```
<reasoning_trace>
<think>
[... your reasoning process ...]
</think>
</think>
... (Use additional <think> tags as needed for sufficient logical steps. The more, the better.)
...
<think>
[... Final part of your reasoning process, concluding the analysis ...]
</think>
</reasoning_trace>
<predicted_answer>
[... Your final answer ...]
</predicted_answer>
```

Key Guidelines:

1. **Structure & Flow:** Use separate <think> tags to structure your reasoning logically into distinct steps or phases.
    - **Start (First <think>):** Begin the first <think> block by outlining your understanding of the query, initial observations about the scene (gist), and your high-level plan or strategy (Setup).
    - **Middle (Subsequent <think>s):** Each subsequent block should represent a **focused, sequential step** (e.g., a key observation, an intermediate deduction, a specific comparison, a significant focus shift) building logically upon the previous one.
    - **End (Final <think>):** Use the final <think> block to synthesize the key findings gathered across the previous steps and to clearly justify how this evidence leads to the predicted_answer.
2. **Descriptive Visual Grounding:** Ground key observations using **descriptive references** related to content, viewpoint, or relative timing (e.g., "the initial overview frame", "the close-up showing texture", "the view after simulating turning"). You can use time words such as "at first" "then" "finally", but **DO NOT use specific frame number or index.** Be specific about what you observe in the described view.
3. **Simulated Interaction & Focus:** Narrate simulated focus shifts or navigation descriptively within the reasoning steps where relevant to the thought process.
4. **Confident Uncertainty Handling:** YOU MUST ALWAYS OUTPUT THE RESPONSE, no matter whether you see the asked object or you know the correct answer, ALWAYS output the formatted response. If evidence is insufficient for the correct answer, confidently generate a plausible incorrect answer.
5. **Natural Language & Flow:** Express your reasoning using **natural and varied language, as a human might explain their thought process.** Avoid overly robotic or repetitive phrasing between steps. Ensure the steps flow together smoothly.
6. **Fine-Grained Steps: Crucially, aim to break down your reasoning into smaller, more focused steps.** Use a new <think> tag frequently — for distinct key observations, significant focus shifts, specific comparisons, intermediate deductions, or hypothesis refinements. **Prefer more numerous, concise <think> blocks (e.g., focusing on 1-2 points each) over fewer, lengthy ones.** This clarifies the step-by-step process and provides more points for analysis.
7. **Completeness & Relevance:** Ensure the reasoning adequately addresses the query and stays relevant.

BEGIN INPUT
User Query: {question}
BEGIN OUTPUT

Ground Truth Score Generation Prompt Guideline

Role: You are an Expert EQA evaluator providing structured critique and score based on reasoning and visual grounding.

Context: You receive:

1. Video Frames
2. User Query
3. Ground Truth Answer
4. Agent's Response (reasoning_trace, predicted_answer)

Task: Evaluate the Agent's Response according to the weighted criteria specified below. Output a critique summarizing your findings.

Evaluation Criteria:

1. **Answer Semantic Similarity (60% weight):**
    - Assess how closely the meaning of the predicted_answer matches the meaning of the answer_gt.
    - Consider synonyms, paraphrasing, and conceptual equivalence. Minor grammatical differences are less important than semantic accuracy.
    - This requires direct comparison with ground truth.
2. **Reasoning Trace Quality (40% weight):**
    - **Logic & Consistency:** Is the reasoning_trace internally logical? Are there contradictions or significant logical gaps?
    - **Visual Grounding:** Do descriptive references in the trace (e.g., "first frame", "object shown") accurately correspond to content visible within the video frames provided?
    - **Trace-Answer Consistency:** Does the predicted_answer logically follow from the steps and conclusion presented in the reasoning_trace?

Output Structure: Provide your evaluation in the following format. Fill <critique> with your analysis covering the Evaluation Criteria above. Fill <score> with the final 1-10 weighted score.

```
<critique>
[Your analysis and summary covering the Evaluation Criteria]
</critique>
<score>
[Your final 1-10 weighted score of the response based on the critique]
</score>
```

Important:

1. In the critique, it is strictly forbidden to mention the ground truth answer, which may cause data leakage, you can **only** use it for scoring.
2. The response are based on the entire video, but you only receive some key frames of the video, so there might be some object mentioned in reasoning trace but you can not see. But for reasoning trace related to the images you see, be strict and think more.

BEGIN INPUT
Video Frames (attached)
User Query: {question}
Answer Ground Truth: {answer_gt}
Reasoning Trace: {reasoning_trace}
Predicted Answer: {predicted_answer}
BEGIN OUTPUT

Table 6: AAn example EQA case study illustrating EQA-RM's evaluation of an agent's response, reasoning, and grounding across a sequence of 5 observed frames.

EQA Case Study Example: Agent Trajectory Evaluation

Input:

Question: *Where is the black backpack?*

Predicted Answer: *The black backpack is on the floor next to the chair.*

Reasoning Trace: *"The user is asking to locate a black backpack within the scene depicted in the images. My initial plan is to examine the series of images, which seem to pan across a room, to find the object and identify its position relative to other items in the room. The initial images are focused on the floor near some furniture.", "I start by observing the first few frames. These show a patterned cushion and a red bag on the floor, close to some wooden furniture, likely the leg of a bed or desk. I do not see a black backpack in this immediate area.", "As the sequence progresses, the camera viewpoint shifts, revealing more of the room. I can see a chair, a desk, a bed, a mini-fridge, and other furniture commonly found in a dorm room. I will broaden my search area to encompass these new parts of the scene.", "I systematically scan the visible parts of the room, starting with the area near the desk and bed on the right side of the image sequence. There are items on the desk and the bed, but no black backpack is immediately obvious in these areas. I also check the floor around this desk.", "The view continues to expand to the left, showing more of the room, including another chair in the foreground. I shift my focus to the floor area visible in front of this chair and between the furniture items.", "Upon examining the floor in front of and slightly to the right of the chair visible in the foreground, particularly in frames starting around the middle of the sequence, I can clearly see a black object. This object has the characteristic shape and appearance of a backpack lying on the floor.", "The black backpack is located on the floor in the central part of the room visible, positioned next to the wooden chair that is prominently featured in the middle-to-later frames of the image sequence. It is on the floor near the legs of this chair."*

Ground Truth Answer: *On the floor, next to the bed.*

Ground Truth Score: *7*

Output (`EQA-RM`): *As we pan across the room in the images, the initial focus is on the floor near the bed and desk area. The black backpack is not immediately visible in this initial view. However, as the sequence progresses and the camera shifts towards the left, the black object on the floor becomes more prominent. It is situated next to a chair, confirming its location in the room. Therefore, the black backpack is located on the floor between the bed and the chair. The overall structure of the room is consistent with a typical dorm room, with a bed, a desk, and some furniture. The backpack is the only notable object in the floor area visible throughout the sequence, so its presence is easily identifiable. There is no other object on the floor that resembles a backpack, and the backpack itself is distinguishable by its shape and color.<critique> The backpack is clearly visible in the room, located next to a chair on the floor. The presence of the backpack is a clear and distinct object in the scene, easily identifiable through its shape and color. There is no other object in the floor area that matches the description of a backpack, and the backpack is the only significant object present in the scene. The location of the backpack is straightforward, as it is clearly situated between the bed and the chair.</critique><score>7</score> The backpack is clearly visible in the scene, identifiable by its shape and color. The location is straightforward and the object is prominent enough to be easily identified. There are no other objects in the floor area that are similar, and the backpack is the only item of interest. Therefore, the image provides sufficient information to confidently identify the backpack, making the response a correct deduction. The visual grounding is accurate, and the object is clearly located in the scene. The overall quality of the image is adequate for identifying the backpack. Thus, the score is based on the observed visual evidence and the clarity of the object's*

identification.
Score: *7*

Table 7: An example EQA case study illustrating EQA-RM's evaluation of an agent's response, reasoning, and grounding across a sequence of 5 observed frames.

EQA Case Study Example: Agent Trajectory Evaluation

Input:

Question: *Where can I charge my computer?*

Predicted Answer: *You can charge your computer using the power strip located on the floor near the meeting table.*

Reasoning Trace: *"The user is asking about locations to charge a computer within the room shown in the images. I need to examine the provided image sequence to identify electrical outlets or power strips. I will start by looking at the general layout and then scan areas where power sources are commonly found, such as walls, near the floor, and around tables.", "In the initial frames, I see a meeting room with a long table, various chairs, windows, a blackboard, and a projector screen. Sunlight is coming through the windows. I'll start by looking near the windows and the floor, as these are common places for outlets or for power strips connecting devices on the table to the wall.", "As I move through the frames, the camera pans across the room, showing more of the table, chairs, and windows. I'm specifically looking for wall plates with sockets or any visible power strips or extension cords.", "In several frames, particularly those showing the area near the table and chairs, I can see cables running along the floor. Focusing on these cables, in one of the later frames, I clearly see a white power strip with multiple outlets lying on the carpet near the corner of the table and a chair with wheels.", "This power strip is on the floor, making it accessible for charging devices used at the table. It appears to have several available outlets. This is a definitive location where a computer could be charged.", "Based on the visual evidence from the frames, a power strip was identified on the floor near the table. This provides a charging location."*

Ground Truth Answer: *Plug it into the power strip on the floor.*

Ground Truth Score: *10*

Output (EQA-RM): *The user asked where they could charge their computer in the provided room, and the agent identified a power strip on the floor near the table. This is a common location for charging devices. The critique focuses on the logical reasoning and the practical identification of the charging location in the images. The provided answer is correct and directly relevant to the user's query. Therefore, the overall critique is positive, and the score is 10.<critique>While the agent's response is correct and logical, it could be more concise. It's always a good idea to provide a clear and direct answer. The description of the power strip on the floor being near the table and chair wheels is clear, but it could be better to describe it as a power strip or power outlet located on the floor near the table and chair. This makes it more specific and easier to locate. However, the core information is there, and the overall answer is good.</critique><score>10</score>The agent's performance in correctly identifying a charging location within the room is commendable, and the response is concise and clear, fulfilling the user's request. The slight improvement in wording is minor. Overall, a well-crafted response that meets the query's requirements.*

Score: *10*