

PhyCritic: Multimodal Critic Models for Physical AI

Tianyi Xiong^{1*}, Shihao Wang, Guilin Liu, Yi Dong, Ming Li¹, Heng Huang¹, Jan Kautz, Zhiding Yu[†]

Abstract

With the rapid development of large multimodal models, reliable judge and critic models have become essential for open-ended evaluation and preference alignment, providing pairwise preferences, numerical scores, and explanatory justifications for assessing model-generated responses. However, existing critics are primarily trained in general visual domains such as captioning or image question answering, leaving physical AI tasks involving perception, causal reasoning, and planning largely underexplored. We introduce PhyCritic, a multimodal critic model optimized for physical AI through a two-stage RLVR pipeline: a physical skill warmup stage that enhances physically oriented perception and reasoning, followed by *self-referential critic finetuning*, where the critic generates its own prediction as an internal reference before judging candidate responses, improving judgment stability and physical correctness. Across both physical and general-purpose multimodal judge benchmarks, PhyCritic achieves strong performance gains over open-source baselines and, when applied as a policy model, further improves perception and reasoning in physically grounded tasks.

Links: [Project Page](#)

1. Introduction

Physical AI has emerged as a new frontier that involves visual perception, physical commonsense, spatial reasoning, and action-centric decision making within a single computational framework. Unlike traditional visual recognition tasks, physical AI requires a model to interpret complex multi-view observations, understand object affordances, reason over causal dynamics, and assess how hypothetical actions unfold in real environments. This paradigm spans 3D perception and spatial grounding [Zhang et al. \(2025\)](#), robot-centric interaction understanding [Bu et al. \(2025\)](#); [Sermanet et al. \(2024\)](#); [Walke et al. \(2023\)](#); [Wang et al. \(2023\)](#), and safety-critical domains such as autonomous driving [Marcu et al. \(2024\)](#); [Nie et al. \(2024\)](#); [Sima et al. \(2024\)](#); [Xie et al. \(2025\)](#). As these systems grow in scale and autonomy, the community increasingly relies on multimodal evaluation to measure whether a model’s reasoning is physically correct, visually grounded, and aligned with human expectations.

However, despite rapid progress in multimodal large language models (MLLMs), the development of reliable multimodal *critic models*—models that evaluate other models’ outputs—lags far behind. Existing reward or judge models focus predominantly on general domains such as captioning, STEM reasoning, and image question answering [Lee et al. \(2024\)](#); [Wang et al. \(2025, 2026\)](#); [Xiong et al. \(2025\)](#); [Zhang et al. \(2025\)](#). Yet, evaluation in physical AI is fundamentally different: the critic must assess whether reasoning is causally valid, whether a visual explanation adheres to actual physical configurations, and whether the final answer respects temporal, spatial, and dynamical constraints.

Recent works have started to extend multimodal judges and RL-based critic training [Wang et al. \(2025,\); Zhang et al. \(2025\)](#) to physical-related scenarios, while early efforts such as DriveCritic [Song et al. \(2025\)](#) underscore the importance of domain-specific judgment capabilities. But existing critics remain limited in three essential ways. (1) They lack *physics awareness*, often failing to distinguish visually coherent but physically impossible reasoning. (2) Their training data focuses on broad multimodal evaluation rather than physically grounded scenarios involving manipulation, affordance reasoning, or embodied 3D interactions. (3) They do not ground

* Work done during an internship at NVIDIA. † Corresponding author: zhidingy@nvidia.com.

Additional affiliations: ¹ University of Maryland, College Park.

© 2026 NVIDIA. All rights reserved.

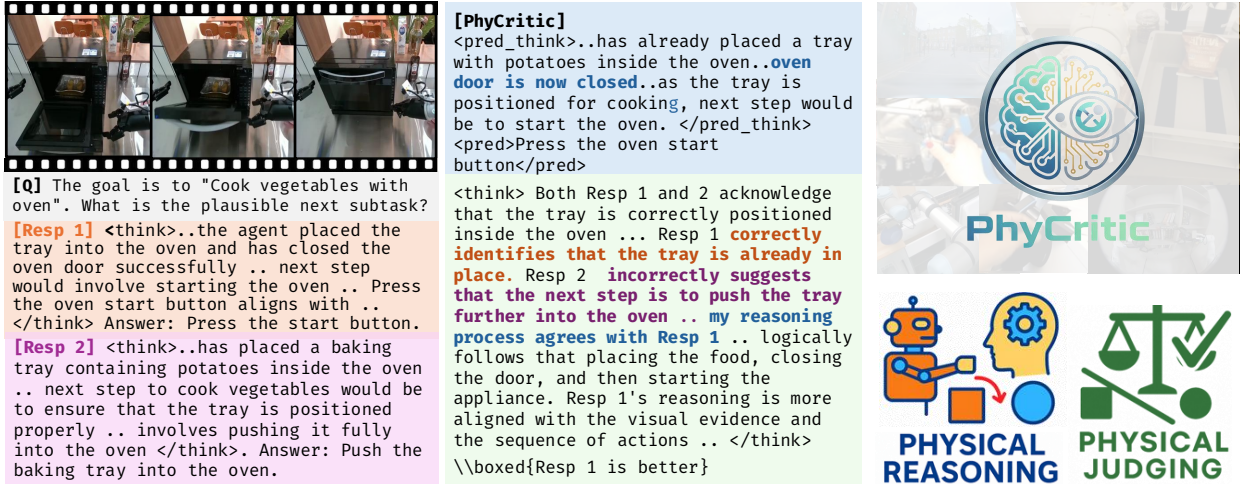


Figure 1: PhyCritic first produces its own physics-aware reasoning and prediction, then explicitly applies it as reference in judging a pair of model responses. In this example, PhyCritic first infers in its own prediction that “the oven is closed”. Based on this insight, the model then correctly identifies Response 1 as following the proper causal sequence while Response 2 proposes an unnecessary action. This self-referential process leads to more stable, physically correct judgments.

their decisions in their own physical understanding of the question, which potentially leads to inconsistent or superficial verdicts. In contrast, recent advances in reinforcement-finetuned multimodal *policy* models show that RLVR-style verifiable rewards Huang et al. (2025); Liu et al. (2025); Shen et al. (2025); Yu et al. (2025) and physically grounded reasoning datasets Azzolini et al. (2025) can significantly improve multimodal reasoning and temporal consistency. Yet, these insights have not been systematically transferred to physical-related critic models, especially in settings where judgments must reflect physical truth rather than linguistic form.

Our goal. We propose to bridge this gap by developing a new class of multimodal critics specifically designed for physical AI. Our model, called **PhyCritic**, aims to evaluate multimodal responses involving physical perception, causal reasoning, and action or plan assessment, and to do so in a manner that is grounded, stable, and physically correct.

As shown in Fig. 1, PhyCritic introduces the principle that a strong physical critic should behave like an expert human judge: before evaluating other models’ responses, it should first solve the problem itself. This intuition motivates *self-referential critic finetuning*, a two-stage reinforcement training framework:

- *Stage 1:* we apply standard GRPO on a small set of physical-related question–answer pairs to strengthen the model’s core physical perception and reasoning abilities, serving as a warm-up phase.
- *Stage 2:* Building upon the Stage-1 warm-up, the critic is trained to (i) generate its own internal reasoning and prediction for the question, and then (ii) evaluate candidate responses with explicit reference to this self-prediction. Using GRPO with both critic and self-prediction rewards encourages stable critic behavior and coherent physics-aware reasoning.

A new benchmark for physical critics. To rigorously evaluate judgment performance in physical contexts, we introduce **PhyCritic-Bench**, a novel benchmark explicitly targeting multimodal critic models for physical AI. Built from diverse embodied datasets such as RoboVQA, BridgeData V2, HoloAssist, and AgiBot World, PhyCritic-Bench includes high-quality physical reasoning questions derived from Cosmos-Reason1 and paired candidate responses scored via verifiable ground truth. This enables fine-grained evaluation of reasoning correctness, visual grounding, and causal validity.

Contributions. Our main contributions in this work are summarized as follows:

- We introduce a self-referential critic learning framework that explicitly grounds the evaluation process in the model’s own physical perception and reasoning, implemented with a two-stage RLVR + GRPO pipeline.
- We develop **PhyCritic**, a multimodal critic specialized for assessing perception, causal reasoning, and planning in physical AI scenarios.
- We construct a high-quality physical critic dataset spanning diverse embodied domains with paired candidate responses and verifiable preference labels.
- We introduce **PhyCritic-Bench**, a benchmark for evaluating multimodal critic models in physical contexts.
- We demonstrate strong empirical gains across physical reasoning benchmarks (Cosmos-Reason1, CV-Bench, EgoPlan-Bench2) and general reward benchmarks (VL-RewardBench, Multimodal RewardBench), outperforming all open-source 7B/8B baselines.

Together, these results demonstrate that critic models benefit significantly from self-referential physical grounding, and that physical AI requires a new generation of physics-aware multimodal judge models.

2. Related Works

VLMs for physical AI. Recent progress in vision–language models (VLMs) has expanded their scope from passive perception to physical AI tasks requiring perception, action, and spatial reasoning. For spatial and 3D understanding, SPAR-7M and SPAR-Bench [Zhang et al. \(2025\)](#) establish large-scale 3D-aware QA benchmarks to enhance multi-view and video-based grounding, while models such as 3D-LLM [Hong et al. \(2023\)](#) and Point-LLM [Guo et al. \(2023\)](#) integrate point-cloud features for open-vocabulary geometric reasoning. In autonomous driving, benchmarks including Lingo-QA [Marcu et al. \(2024\)](#), DriveLM [Sima et al. \(2024\)](#), Reason2Drive [Nie et al. \(2024\)](#), and DriveBench [Xie et al. \(2025\)](#) assess multimodal reasoning, planning, and robustness under safety-critical conditions. Vision–Language–Action (VLA) frameworks such as RT-2 [Zitkovich et al. \(2023\)](#) and OpenVLA [Kim et al. \(2024\)](#) unify perception, language, and control within a single policy space, while Open-X-Embodiment [O’Neill et al. \(2024\)](#) facilitates large-scale cross-robot generalization through shared representation learning. Cosmos-Reason-1 [Azzolini et al. \(2025\)](#) further introduces physically grounded chain-of-thought reasoning via supervised and reinforcement fine-tuning, bridging perception and planning for embodied agents. Despite progress in perception and action, existing models often overlook the physical correctness and causal validity of their reasoning. PhyCritic introduces a physics-aware critic for diagnostic evaluation in physical AI.

Multimodal reward and critic models. Vision-language models (VLMs) have recently been extended beyond perception and reasoning to act as evaluators, assessing the quality of model-generated responses. Early studies showed that proprietary models such as GPT-4V [Zhang et al. \(2023\)](#) align closely with human judgments across diverse multimodal evaluation tasks. Subsequent work fine-tuned open-source VLMs into reward or critic models, typically following two paradigms: (1) BT-style reward modeling with explicit scalar heads to predict preference scores [Sun et al. \(2023\)](#); [Zang et al. \(2025\)](#); [Zhang et al. \(2025\)](#), and (2) generative judges that produce textual reasoning and final verdicts in an autoregressive manner [Lee et al. \(2024\)](#); [Wang et al. \(2025,\)](#); [Xiong et al. \(2025\)](#); [Zhang et al. \(2025\)](#), or combine both approaches [Zhang et al. \(2025\)](#). Later extensions incorporated critic training with reinforcement learning [Wang et al. \(2025,\)](#); [Zhang et al. \(2025\)](#) to enhance judgment accuracy. However, most efforts remain focused on general visual domains such as captioning, STEM reasoning, or visual question answering, leaving physical AI scenarios underexplored. Among earlier attempts, WorldModelBench [Li et al. \(2025\)](#) evaluates video-generation models on instruction following and physics adherence, underscoring the need for physics-aware multimodal evaluation. More recently, DriveCritic [Song et al. \(2025\)](#) employed LMM-based critics to assess trajectory pairs in autonomous driving, yet its scope remains and does not apply to diverse, grounded multimodal text responses. Our work broadens the critic capabilities

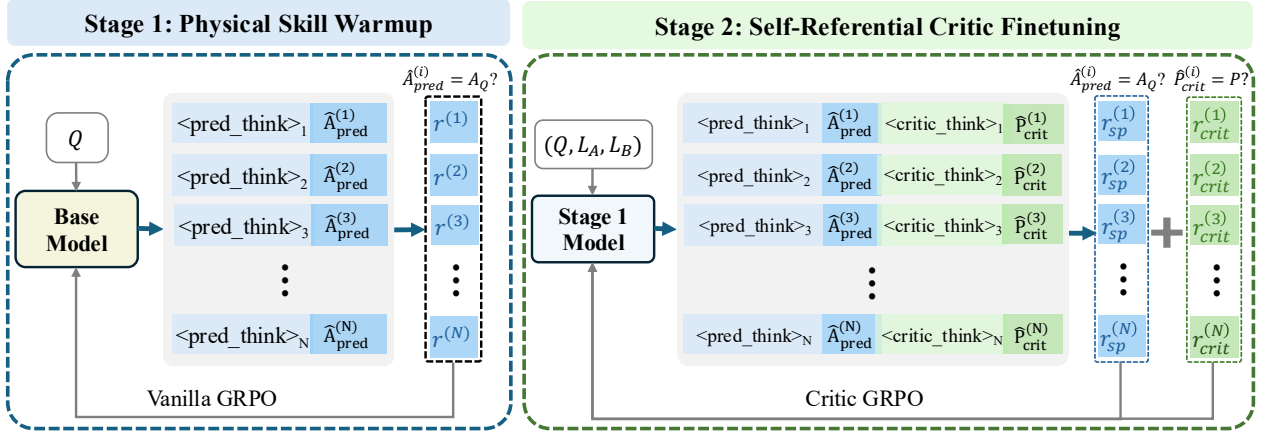


Figure 2: **PhyCritic training pipeline.** We begin with GRPO training on physical-related QA pairs to enhance the VLM’s physical reasoning ability (*left*), followed by self-referential critic finetuning to further develop its critique capacity (*right*).

of open-source VLMs toward general physical AI domains, emphasizing evaluation settings that involve physical perception, reasoning, and planning.

Multimodal reinforcement finetuning. Reinforcement finetuning has proven effective in enhancing the capabilities of vision-language models (VLMs). Early explorations applied reinforcement learning with human feedback (RLHF Ouyang et al. (2022)) and AI feedback (RLAIF Lee et al. (2023)) paradigms, typically using PPO Schulman et al. (2017) or DPO Rafailov et al. (2023), to mitigate hallucinations Sun et al. (2023); Yu et al. (2024, 2025) and improve video understanding Zhang et al. (2025). Following the success of DeepSeek-R1 Guo et al. (2025), subsequent studies have increasingly adopted Reinforcement Learning with Verifiable Rewards (RLVR) approaches to strengthen visual reasoning in multimodal mathematics and science Huang et al. (2025); Liu et al. (2025,); Meng et al. (2025); Wang et al. (2025); Yang et al. (2025), perception-related tasks Shen et al. (2025); Yu et al. (2025); Zhou et al. (2025), video reasoning Feng et al. (2025), and domain-specific applications including medical imaging Lai et al. (2025), GUI agents Luo et al. (2025), and robotic reasoning Azzolini et al. (2025). These efforts primarily target tasks with verifiable ground-truth supervision, aiming to improve models as policy for problem solving. More recently, RLVR-based methods have been extended to multimodal critic finetuning—training models to evaluate rather than generate responses—thereby enhancing judgment accuracy on image-grounded tasks Li et al.; Wang et al. (2025,); Zhang et al. (2025). PhyCritic extends multimodal RLVR to develop physical critics that evaluate responses involving perception, action, and planning in physical contexts, while enhancing the model’s ability for physical reasoning.

3. PhyCritic: Physical Critic for VLMs

Inspired by human critical thinking Lai (2011), we identify two key principles for consolidating physics-related critic capacities in VLMs. (1) These capacities should be developed from model’s intrinsic ability for physics-related perception and reasoning. (2) Building on the success of multimodal critics in both general image understanding Lee et al. (2024); Wang et al. (2025); Xiong et al. (2025); Zhang et al. (2025) and visual generation domains Liu et al. (2025); Wang et al. (2025), critic performance should be consolidated through explicit critic training with strong preference signals. Therefore, PhyCritic adopts a two-stage reinforcement fine-tuning pipeline. Starting from a base model, we first fine-tune it on a small set of physical-related question-answer pairs to enhance its physical perception and reasoning. We then introduce *self-referential critic fine-tuning*, a reinforcement learning stage where the critic is prompted to generate its own prediction for the user’s query before evaluating pairs of model responses. An overview is provided in Fig. 2.

3.1. Task Formulation

To implement this mechanism, our training data are organized as tuples (Q, L_A, L_B, A_Q, P) , where:

- Q denotes a multimodal prompt consisting of a user question together with visual inputs;
- L_A and L_B are two candidate responses to be evaluated;
- A_Q denotes the ground-truth answer to Q ;
- $P \in A, B$ is a binary preference label indicating which response demonstrates higher quality.

3.2. Two-Stage Training Pipeline

As existing VLMs are primarily trained on general image–text pairs with limited exposure to physically grounded data, we devise a two-stage training pipeline to enhance physical critic capacity.

Stage 1: Physical Skill Warmup with RLVR. Before introducing the complex critic task, we first ensure that the model possesses solid foundational perception and reasoning capabilities within the physical domain. In this stage, we use a verifiable dataset containing only (Q, A_Q) pairs and adopt a standard reinforcement learning setup with accuracy reward

$$r = \mathbb{I}(\hat{A}_{pred}(Q) = A_Q)$$

The objective is to align the model to produce accurate and reliable predictions \hat{A}_{pred} , establishing a strong foundation for critic training in the next stage.

Stage 2: Self-Referential Critic Finetuning. After acquiring foundational physical capacities, we utilize the complete dataset (Q, L_A, L_B, A_Q, P) to align the model’s judgment over a pair of responses, $\hat{P}_{pred}(Q, L_A, L_B)$, with the ground-truth preference label P . As will be illustrated in Section 3.3, this stage encourages the model to anchor its external critic judgment to its internal self-prediction \hat{A}_{pred} , thereby promoting the development of more grounded and interpretable physical critic capacities.

3.3. Self-Referential Critic Finetuning

Our core idea is that a critic’s judgment should be grounded in its own understanding of the problem. Therefore, during training, the model π_θ is required to perform two tasks concurrently:

1. **Self-Prediction:** The model first generates its own internal prediction \hat{A}_{pred} in response to the prompt Q .
2. **Preference Judgement:** Subsequently, acting as a critic, the model produces its preference prediction \hat{P}_{pred} for the context (Q, L_A, L_B) , while being explicitly instructed to ground its evaluation process on the previously generated self-prediction \hat{A}_{pred} .

Reward Design. The total reward r_{total} is composed of an accuracy reward r_{acc} and a format reward r_{format} .

$$r_{total} = r_{acc} + r_{form} * \alpha_{form}$$

where α_{form} is a weighting coefficient for format reward.

Accuracy Reward. In self-referential critic fine-tuning, the accuracy reward r_{acc} consists of two components:

- **Self-Prediction Reward (r_{sp}):** Evaluates the correctness of the model’s internal knowledge by comparing its self-prediction \hat{A}_{pred} with the ground-truth answer A_Q :

$$r_{sp} = \mathbb{I}(\hat{A}_{pred} = A_Q)$$

which encourages the VLM to first become a reliable problem solver.

- **Critic Reward (r_{crit}):** Evaluates the model’s judgment capability as a critic by checking whether its

You are an expert in evaluating responses from multimodal models. Your task is to compare two model responses to the same question grounded on the provided image or video. You should assess both the **quality of the reasoning process** and the **factual correctness of the final answer** in each response.

Your evaluation should focus on two main aspects:

1. Reasoning Process Quality:

Evaluate the model’s reasoning based on the following aspects:

– *Truthfulness*: All factual statements must be correct and verifiable. No fabricated details or contradictions with the visual input; – *Visual Groundedness*: The reasoning must accurately reference and interpret visual elements from the provided image or video; – *Logical Validity*: The reasoning should follow a logically consistent, step-by-step progression; – *Efficiency and Clarity*: Reasoning should be clear, purposeful, and focused. Thoughtful self-correction is acceptable, but avoid unnecessary repetition or irrelevant information.

2. Final Answer Accuracy:

Evaluate whether the final answer is factually correct given the question and the visual input.

Here are the inputs for evaluation:

[Question]:{question}; [Response 1]:{resp1}; [Response 2]:{resp2}

To begin your evaluation, first generate your own reasoning process towards solving the question, enclosing it within `<pred_think></pred_think>`. Then, provide your own answer to the question in `<pred></pred>`. After that, provide a detailed justification comparing the two responses, using your own response as a reference point. Evaluate each response according to the criteria above, considering both reasoning quality and final answer correctness. Your explanation must clearly reference specific details from both responses, your own response, and the visual evidence. Finally, make a clear and strict decision on which response is better overall. Output your decision in one of the following formats:

– Response 1 is better; – Response 2 is better

You **FIRST** write your reasoning process to the question in `<pred_think></pred_think>`, **THEN** generate your own answer in `<pred></pred>`, **THEN** think about the comparison and judgment process as an internal monologue (enclosed within `<think></think>`), and **FINALLY** provide the final answer in `\boxed{}`.

Table 1: **Critic prompt for self-referential critic fine-tuning.** After presenting detailed evaluation criteria, it explicitly instructs the judge model to first generate its own reasoning and prediction for the given question, then use its self-prediction as a reference during the critique process. **Format prompt** regions are highlighted in green.

predicted preference \hat{P}_{crit} matches the golden preference P :

$$r_{\text{crit}} = \mathbb{I}(\hat{P}_{\text{crit}}(Q, L_A, L_B) = P)$$

The overall accuracy reward r_{acc} is computed as a weighted sum of the two components:

$$r_{\text{acc}} = \alpha_{\text{sp}} r_{\text{sp}} + \alpha_{\text{crit}} r_{\text{crit}}$$

where α_{sp} and α_{crit} are weighting coefficients controlling their relative importance.

Format Reward (r_{form}). This reward evaluates whether the critic’s output follows the predefined self-referential structure:

$$r_{\text{form}} = \begin{cases} 1.0 & \text{if } \langle \text{pred_think} \rangle, \langle \text{pred} \rangle, \\ & \langle \text{think} \rangle \text{ and } \backslash \text{boxed} \{ \} \text{ exist,} \\ 0.5 & \text{if only } \langle \text{think} \rangle \text{ and } \backslash \text{boxed} \{ \} \text{ exist,} \\ 0 & \text{otherwise.} \end{cases}$$

Optimization Algorithm. Inspired by the strong performance DeepSeek-R1 Guo et al. (2025), we employ Group Relative Policy Optimization (GRPO) as the optimization algorithm. GRPO is a policy-gradient based reinforcement-learning method originally introduced in the context of mathematical reasoning Shao et al. (2024). It omits the need for a learned value-network in PPO, and instead computes advantages by comparing

multiple sampled trajectories as a group.

$$\mathcal{L}_{\text{GRPO}} \mathbb{E}_{o \sim \pi_o} [\min(\rho_o A_o, \text{clip}(\rho_o, 1 - \epsilon, 1 + \epsilon) A_o)] - \beta \mathcal{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$$

where $\rho_o = \frac{\pi_\theta(o)}{\pi_{\text{ref}}(o)}$ is the probability ratio, $A_o = \frac{r_o - \bar{r}}{\text{std}(r)}$ is the group-relative advantage, computed by normalizing each outcome reward r_o for the sampled trace o .

Critic Prompt. As shown in Tab. 1, we design a rigorous critic prompt template that explicitly instructs the model to follow a clear set of predefined evaluation criteria, and to reference its critique process with respect to its own prediction. Following DeepSeek-R1 Guo et al. (2025), we append a format prompt at the end to standardize the model’s reasoning behavior and output structure.

3.4. Critic Training Dataset

To enhance critic capacities, we construct a dataset spanning diverse physical scenarios and video domains, paired with high-quality questions probing perception, reasoning, and causality. Candidate responses of varying quality enable the critic to learn nuanced distinctions in reasoning accuracy and visual grounding.

We select videos from four major robotics and embodied datasets—RoboVQA Sermanet et al. (2024), Bridge-Data V2 Walke et al. (2023), HoloAssist Wang et al. (2023), and AgiBot World Bu et al. (2025)—covering both egocentric and third-person views, a wide range of manipulation behaviors (e.g., grasping, stacking, folding, assembling), and diverse physical contexts from simulated kitchens to real-world workspaces. For questions, we build upon the Cosmos-Reason1 RL dataset Azzolini et al. (2025), which provides 800 high-quality question–answer pairs requiring rich physical perception, planning, and reasoning capacities.

Model outputs are collected from seven multimodal models spanning proprietary (GPT-4o Hurst et al. (2024), Gemini-2.5-Flash Comanici et al. (2025)), open-source (Qwen2.5-VL-72B Bai et al. (2025), InternVL3-38B Zhu et al. (2025)), and RL-finetuned systems with enhanced long-horizon reasoning (Cosmos-Reason1-7B Azzolini et al. (2025), Video-R1 Feng et al. (2025), MiMo-VL-7B Team et al. (2025)). All models are prompted in a Chain-of-Thought (CoT) manner to produce explicit reasoning traces with their final responses.

For ground-truth critic preference, we apply a simple accuracy-based labeling: GPT-4o verifies each response against the ground-truth answer and assigns a binary score (1 = preferred, 0 = rejected). We then construct response pairs, each consisting of one chosen and one rejected response. After balancing response lengths and distribution, we obtain a final critic dataset of 3,258 samples.

4. PhyCritic-Bench

Existing multimodal reward benchmarks primarily cover prompts from general visual domains such as captioning Ye et al. (2025), STEM reasoning Chen et al. (2024); Xiong et al. (2025); Zhang et al. (2025) and general image question answering Li et al. (2025); Yasunaga et al. (2025), does not cover physical AI domains that involves judgments related to physical perception, actions and planning. We thus introduce PhyCritic-Bench, a fine-grained and challenging benchmark targeted in physical AI domains for multimodal judge models.

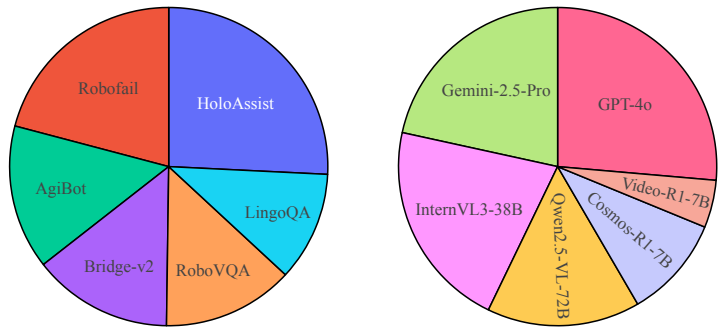


Figure 3: Distribution of prompt sources (left) and model responses (right) in PhyCritic-Bench.

Model	PhyCritic-Bench							VL-RewardBench					MM-RB.
	overall	AgiB.	Holo.	R-VQA	Bridge.	R-Fail	Lingo.	overall	macro	General	Hallu.	Reason	overall
Gemini-2.5-Pro	78.2	87.9	82.8	86.7	68.8	76.6	60.0	74.9	72.9	59.1	85.2	74.4	85.4
Gemini-2.5-Flash	67.1	63.6	70.7	73.3	81.2	63.8	44.0	71.9	72.0	58.0	77.0	81.0	82.5
GPT-4o	64.7	57.6	70.2	46.7	75.0	72.3	56.0	65.8	62.4	49.6	67.6	70.5	71.5
<i>Open-Source 7B/8B VLMs and Physical-Task-Tuned Variants</i>													
Eagle-2.5-8B	56.0	75.8	50.0	60.0	50.0	48.9	60.0	50.2	49.7	41.4	48.6	59.3	64.4
Qwen2.5-VL-7B	51.6	51.5	51.7	46.7	50.0	57.4	48.0	53.2	50.9	40.9	54.3	57.4	64.0
Robobrain2.0-7B	54.7	51.5	65.5	43.3	46.9	55.3	56.0	42.4	44.2	39.2	55.8	37.5	50.5
Cosmos-R1-7B	51.1	36.4	53.4	43.3	68.8	51.1	52.0	44.8	44.8	33.1	41.3	59.9	54.8
PhyCritic-7B (ours)	68.0	78.8	65.5	86.7	65.6	57.4	60.0	57.3	54.9	45.3	58.6	60.9	65.9

Table 2: Performance comparison on PhyCritic-Bench and two general multimodal reward benchmarks. Best and second-best results of *open-source* VLMs are marked in **bold** and underline, respectively. *MM-RB. is an abbreviated form of Multimodal RewardBench Yasunaga et al. (2025).

PhyCritic-Bench primarily comprises 225 evaluation samples covers evaluations in two physical-AI scenarios: 1) *Robotics tasks* – videos from RoboVQA Sermanet et al. (2024), BridgeData V2 Walke et al. (2023), HoloAssist Wang et al. (2023), AgiBot Bu et al. (2025), and RoboFail Liu et al. (2023), with questions adapted from CosmosReason1-Bench Azzolini et al. (2025); 2) *Autonomous driving* – videos and action- or prediction-related questions from LingoQA Marcu et al. (2024). Model responses are collected from seven VLMs spanning proprietary, open-source, and RL-finetuned families. An overview of the prompt and response model statistics is shown in Fig. 3.

Following prior works, we formulate each evaluation instance in PhyCritic-Bench as a pairwise preference tuple (x, l_a, l_b, p) , where q is the multimodal prompt containing the visual input and user query, l_a and l_b are candidate responses, and $p \in a, b$ is the ground-truth preference label. For response-pair and preference-label construction, we follow the JudgeBench Tan et al. (2024) pipeline. Each prompt is assigned to one model from a pool, which generates $N=8$ CoT responses via temperature sampling. GPT-4o verifies each response against the ground-truth answer, identifying correct and incorrect predictions. The final pair comprises one correct and one incorrect response, with correctness serving as the preference label. During evaluation, each VLM acts as a pairwise judge, and performance is measured by its preference consistency with the ground truth:

$$Acc = \mathbb{I}(\text{VLM}(q, l_a, l_b) = p)$$

5. Experiments

Implementation Details. We adopt Qwen2.5-VL-7B-Instruct Bai et al. (2025) as our base model and the veRL framework for RL finetuning. In the warmup stage, we train the model using vanilla GRPO on the Cosmos-Reason1-RL dataset Azzolini et al. (2025) for 80 steps. In the critic finetuning stage, we perform self-referential critic finetuning for 300 steps. For both stages, we use a batch size of 128, a learning rate of 1×10^{-6} , and set the KL coefficient to 0.01. For critic finetuning, we set $\alpha_{sp} = 0.2$, $\alpha_{crit} = 0.7$ and $\alpha_{form} = 0.1$.

Evaluation Benchmark. We evaluate our approach on both critic performance and its capacity for handling Physical-AI tasks. For critic performance, we use our curated PhyCritic-Bench for physical-related critique, together with two general reward benchmarks—VL-RewardBench Li et al. (2025) and Multimodal-RewardBench Yasunaga et al. (2025)—to evaluate critic capacity in general visual understanding and knowledge-intensive reasoning. For physical capacity, we evaluate our methods on three physical-related multimodal benchmarks. CosmosReason1-Bench probes physical commonsense and causal/affordance reasoning across agility, holistic perception, robot-VQA, tool-use/bridging, and failure-robustness sub-tasks Azzolini et al. (2025). CV-Bench assesses whether visual representations support downstream control through 2D/3D spatial cog-

Model	CosmosReason1-Bench						CV-Bench			EgoPlanBench2				
	overall	AgiB.	Holo.	R-VQA	Bridge.	RoboFail	overall	2D	3D	overall	Daily	Hobb.	Recreat.	Work
Gemini-2.5-Pro	64.7	45.0	84.0	83.6	45.0	74.0	85.4	79.3	91.5	42.8	44.2	43.1	46.4	39.6
Gemini-2.5-Flash	57.9	42.7	68.2	78.2	35.5	62.8	74.0	74.4	73.6	37.1	38.7	35.6	43.7	33.4
GPT-4o	56.3	38.0	64.0	78.2	33.0	66.0	79.1	73.2	85.0	41.8	47.4	40.0	44.8	35.6
<i>Open-Source 7B/8B VLMs and Physical-Task-Tuned Variants</i>														
Eagle-2.5-8B	54.3	39.0	<u>67.0</u>	67.4	29.0	68.0	77.9	77.5	78.3	43.0	49.4	<u>41.7</u>	<u>47.0</u>	<u>35.1</u>
Qwen2.5-VL-7B	54.3	40.4 [†]	50.0 [†]	82.5 [†]	<u>38.0[†]</u>	57.6 [†]	78.9	75.1	<u>82.7</u>	32.6	36.7	27.7	33.9	32.5
Robobrain-2.0-7B	50.4	37.0	41.0	<u>85.4</u>	29.0	56.0	85.8[†]	–	–	33.2 [†]	39.4 [†]	32.2 [†]	33.9 [†]	29.7 [†]
Cosmos-R1-7B	<u>63.0</u>	<u>49.4[†]</u>	63.0 [†]	83.8 [†]	58.8[†]	60.0 [†]	75.2	74.9	75.7	29.8	33.5	31.2	30.6	24.5
PhyCritic-7B (ours)	63.9	53.0	68.0	95.5	37.0	<u>63.0</u>	<u>79.7</u>	<u>75.5</u>	83.9	<u>42.3</u>	<u>45.6</u>	<u>38.3</u>	<u>47.5</u>	39.4

Table 3: Performance comparison on Physical-related multimodal benchmarks. †: Results reported in original benchmark papers. Best and second-best results of *open-source* VLMs are marked in **bold** and underline, respectively.

dition and cross-view consistency [Tong et al. \(2024\)](#). EgoPlanBench2 targets egocentric daily-task planning (Daily/Hobbies/Recreation/Work), measuring the full perception–decomposition–planning pipeline [Qiu et al. \(2024\)](#).

Baselines. We compare PhyCritic against two groups of baselines. (1) *General-purpose VLMs*: leading open-source models trained for visual instruction following, including Qwen2.5-VL-7B [Bai et al. \(2025\)](#) and Eagle2.5-8B [Chen et al. \(2025\)](#). (2) *Physical reasoning-oriented VLMs* — two representative lines built upon the same Qwen2.5-VL backbone as ours: Cosmos-Reason1 [Azzolini et al. \(2025\)](#), optimized for physical reasoning through large-scale supervised finetuning on millions of physical common-sense and embodied reasoning samples distilled from DeepSeek-R1 traces, followed by reinforcement finetuning on object interaction and temporal-consistency tasks; and RoboBrain2 [Team et al. \(2025\)](#), a perception-to-action model trained via multi-stage supervised and reinforcement finetuning that emphasizes task decomposition, subgoal generation, and robustness in embodied environments.

5.1. Main Results

SOTA open-source 7B/8B critic on physical judgment. On PhyCritic-Bench (Tab. 2), our model attains the best overall accuracy among open-source 7B/8B models (68.0), outperforming leading general-purpose VLMs—Eagle-2.5-8B (56.0; +12.0), Qwen2.5-VL-7B (51.6; +16.4)—as well as the RL-finetuned physical models RoboBrain2.0-7B (54.7; +13.3) and Cosmos-R1-7B (51.1; +16.9). Across sub-suites, PhyCritic-7B delivers best or tied-best performance on AgiBot (78.8), HoloAssist (65.5; tied), and RoboVQA (86.7), while ranking second on Bridge-v2 (65.6). Notably, its strong judgment capability also generalizes to unseen subdomains, achieving top performance on RoboFail (57.4; tied with the base Qwen2.5-VL) and LingoQA (60.0; tied best). While physical-related RL by itself offers limited benefit to accuracy, our two-stage pipeline yields a *strict* and *stable* judge capable of handling embodied tasks with long-horizon, affordance-centric reasoning.

From physical critic to general-domain judging. As in Tab. 2, despite being tuned only in physical contexts, PhyCritic generalizes well to general-domain judging scenarios. On VL-RewardBench and Multimodal-RewardBench, it outperforms the Qwen2.5-VL base model by +4.1 and +1.9 overall, with consistent gains across general, hallucination-related, and reasoning-related judgments. These results suggest that our physical critic training—rooted in physical perception, planning, and action evaluation—transfers effectively to broader multimodal judging across general visual domains.

Improved physical reasoning as a policy. As shown in Tab. 3, PhyCritic not only judges well but also strengthens *policy-like* physical task handling. On CosmosReason1-Bench, it achieves the best open-source accuracy (63.9), surpassing Cosmos-R1-7B (63.0; +0.9), even though the latter is trained on millions of

in-domain distilled reasoning traces followed by RL finetuning. This strong performance further generalizes to physical perception and reasoning tasks beyond the training distribution. On CV-Bench, PhyCritic-7B achieves the second-best average (79.7) and the best 3D score (83.9), demonstrating improved spatial grounding and cross-view consistency. On EgoPlanBench2, PhyCritic-7B ranks second overall (42.3) with top-two performance across all domains, highlighting robust egocentric planning capabilities.

Data efficiency. Notably, our two-stage RL pipeline requires only 80+300 RL steps, with a total of 4,058 training samples. In contrast to approaches that depend on millions of supervised traces, PhyCritic leverages *verifiable* physical QA and *self-referential* critic signals, enabling a more *data-efficient* path toward strong critic and policy performance.

5.2. Ablation Studies

We further conduct two ablation studies to evaluate the impact of the two-stage RL pipeline and the proposed self-referential critic finetuning on model performance.

Two-stage RL is necessary and complementary.

We compare our method against various RL strategies. In all experiments, s_1 uses 80 RL steps and s_2 uses 300. As shown in Tab. 4, our two-stage pipeline achieves the best results across physical critic judgment, physical reasoning, and general-domain evaluation. Stage 1 physical RL improves physical reasoning (+7.5 on CosmosReason1-Bench) but brings limited judgment gains (+2.0 on PhyCritic-Bench), while Stage 2

mainly strengthens critic capability (+14.4 on PhyCritic-Bench) and further enhances reasoning performance (+2.1 on CosmosReason1-Bench). Notably, the full pipeline surpasses physical RL alone in both stages on physical reasoning tasks, showing that critic training further pushes the model’s problem-solving capacities. Overall, these results indicate that (i) Stage 1 builds foundational physical skills through policy RL warmup, and (ii) Stage 2 self-referential critic finetuning—leveraging diverse reasoning traces and high-quality preference signals—further consolidates judgment consistency, improves generalization, and reduces overfitting.

Self-referential critic finetuning drives the gains.

As shown in Tab. 5, removing the explicit self-referential process during finetuning (*no self-refer.*) reduces PhyCritic-Bench performance from 68.0 to 64.4 (−3.6). Keeping the self-reference prompt but removing the self-prediction reward R_{sp} leads to a smaller yet noticeable drop to 65.8 (−2.2). Similar trends appear on CosmosR1-Bench (63.9 → 62.6 /

63.5) and VL-Reward (57.3 → 56.6 / 56.5). By explicitly instructing the judge model to ground its evaluations in its own reasoning and problem-solving behavior, and by applying rewards that reinforce accurate self-reasoning, the model simultaneously learns to produce *generalized*, *consistent judgment* and *faithful reasoning*.

5.3. Analysis

Better self-predictions ⇒ better judgments.

To assess how self-prediction accuracy affects judgment quality, we conduct chi-square tests between the correctness of PhyCritic’s own answer and its downstream judgment in the self-referential critic process. The Stage 1 model with physical RL warmup already shows a strong positive association ($\chi^2 = 51.07$, $p = 8.93 \times 10^{-13}$), and self-referential critic finetuning further strengthens

Method	PhyCritic-B.		CosmosR1-B.		VL-Reward.
	overall	macro	overall	macro	overall
Qwen2.5-VL-7B	51.6	50.9	54.3	53.7	53.2
- physical RL (s_1)	53.6	54.6	61.8	61.2	52.0
- physical RL (s_1+s_2)	52.7	53.6	63.1	62.6	53.0
- critic RL (s_1+s_2)	62.2	62.4	57.1	56.5	54.0
- mixed RL (s_1+s_2)	66.7	68.1	60.2	59.5	55.5
- Two-stage RL (ours)	68.0	69.0	63.9	63.3	57.3

Table 4: Ablation on RL strategy. “s” denotes training stage.

Method	PhyCritic-B.		CosmosR1-B.		VL-Reward.
	overall	macro	overall	macro	overall
PhyCritic-7B	68.0	69.0	63.9	63.3	57.3
- no self-refer.	64.4	63.8	62.6	61.9	56.6
- no r_{sp}	65.8	64.8	63.5	62.9	56.5

Table 5: Ablations on self-referential critic finetuning.

this dependence in the final model ($\chi^2 = 161.76$, $p = 4.66 \times 10^{-37}$). These results support the premise that a strong physics-aware critic should “solve before judging,” using its own grounded prediction to avoid spurious correlations and unsupported verdicts.

Limitations. While self-referential critic fine-tuning has proven more effective than conventional pairwise critic training, it additionally requires ground-truth answers for the multimodal prompts. Although actions and planning can often be verified through physical-world feedback, this requirement limits its applicability to fully open-ended scenarios. Future work may explore self-verification or meta-judging strategies to replace the explicit accuracy reward used for self-prediction. Furthermore, PhyCritic could be extended toward multi-round critic self-refinement, leveraging critic signals to iteratively improve self-generation.

6. Conclusion

We presented PhyCritic, a multimodal critic model tailored for physical AI scenarios involving perception, causal reasoning, and action or plan evaluation. By introducing a two-stage RLVR pipeline with *self-referential critic finetuning*, PhyCritic grounds its judgments in its own physically informed predictions, yielding more consistent, interpretable evaluations and stronger physical reasoning. We further introduced PhyCritic-Bench, a challenging benchmark for physical-domain multimodal judging. Experiments show that PhyCritic achieves the best performance among open-source 7B/8B models on physical judgment, generalizes well to broader multimodal reward tasks, and enhances physical reasoning. Together with our analysis, PhyCritic paves the way toward more reliable, physical-aware multimodal AI evaluation systems.

A. Additional Results and Analysis

In this section, we first analyze the robustness of (i) reward weight hyperparameters and (ii) critic prompt design; we then validate PhyCritic as a reward for (iii) test-time scaling and (iv) downstream policy training, and finally compare it with (v) a general-domain multimodal critic.

Ablations on α_{sp} . We fix the format reward weight at $\alpha_{form} = 0.1$ and vary the self-prediction reward weight α_{sp} while setting the critic reward weight to $\alpha_{crit} = 0.9 - \alpha_{sp}$. As shown in Tab. 6, PhyCritic exhibits relatively robust performance across different choices of α_{sp} . We adopt $\alpha_{sp} = 0.2$ in our final configuration, as it provides the best overall balance across judge and reasoning benchmarks.

α_{sp}	PhyCritic-B.		CosmosR1-B.		VL-Reward.
	overall	macro	overall	macro	overall
0.1	65.8	66.1	62.9	62.3	58.6
0.2	68.0	69.0	63.9	63.3	57.3
0.3	66.2	66.4	63.3	62.7	57.2
0.4	65.3	65.2	63.7	63.1	56.4

Table 6: Ablations on the self-prediction reward weight (α_{sp}).

Ablations on critic prompts. As shown in Tab. 1, our critic prompts include detailed evaluation criteria that guide the model’s rollout during the critic process—for example, truthfulness, visual groundedness, logical validity, and efficiency for reasoning assessments. To assess the importance of these hand-crafted criteria, we conduct an ablation study that removes all criteria and report the results in Tab. 7. Removing the criteria leads to clear drops in judgment accuracy, physical reasoning performance, and generalization to image-domain judgments. Since our GRPO objective relies only on the self-prediction reward r_{sp} and critic accuracy reward r_{crit} , the criteria in the critic prompts become crucial for shaping and regularizing PhyCritic’s critic reasoning behavior during reinforcement finetuning.

Critic Prompt	PhyCritic-B.		CosmosR1-B.		VL-Reward.
	overall	macro	overall	macro	overall
Full prompt	68.0	69.0	63.9	63.3	57.3
- no criteria	63.9	64.8	62.0	61.3	55.1

Table 7: Ablations on critic prompt criteria.

This finding contrasts with prior work such as Wang et al. (2025), which uses a simple pairwise critic prompt for general-domain critic training. Our results indicate that building reliable critic capabilities for physical-AI tasks is considerably more challenging and requires explicit, structured critic guidance.

PhyCritic for best-of- N sampling. Here, we evaluate the effectiveness of PhyCritic as a test-time scaling judge for best-of- N sampling. We first employ Qwen2.5-VL-7B-Instruct as the *policy model* to generate N candidate reasoning trajectories for each question in CosmosReason1-Bench Azzolini et al. (2025) under a thinking prompt with a temperature of $\tau = 0.6$. PhyCritic-7B is then applied through a *pairwise knockout* procedure: starting from the first two trajectories, PhyCritic judges the pair and advances the preferred response to compete with the next candidate. This process continues for $N - 1$ rounds until a final winner is selected. We compare this strategy against three baselines: 1) majority voting over the final predicted answers; 2) using Qwen2.5-VL-7B-Instruct Bai et al. (2025) itself as the judge for pairwise knockout; and 3) using the physical RL-finetuned Cosmos-Reason1-7B Azzolini et al. (2025) as the judge.

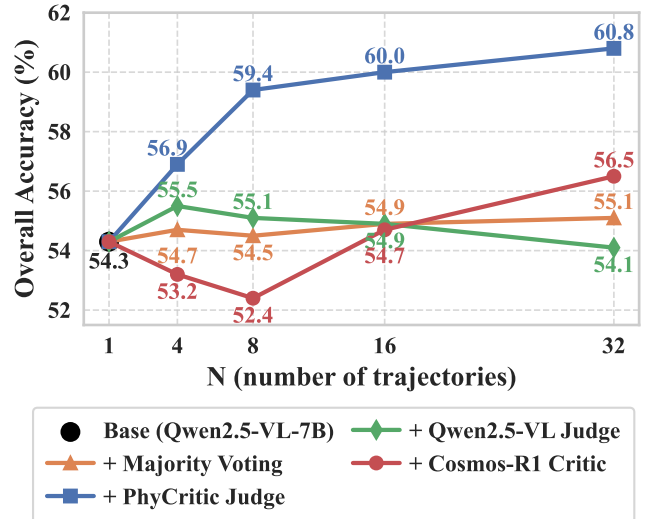


Figure 4: Comparison of Best-of- N ensemble mechanisms on CosmosReason1-Bench. Using PhyCritic-7B as the judge consistently improves the base Qwen2.5-VL-7B-Instruct model.

As shown in Fig. 4, applying PhyCritic to select the best reasoning trajectory yields the most consistent performance gains as N increases, achieving a +6.5-point improvement at $N = 32$ for the base model on CosmosReason1-Bench (60.8 vs 54.3). In contrast, using the physically finetuned Cosmos-Reason1-7B or the base Qwen2.5-VL models as the critic yield much smaller or even no improvements, reflecting their limited physical-domain judgment capacity. By reliably identifying the high-quality trajectory among multiple candidates, PhyCritic serves as an effective ensemble mechanism for enhancing test-time performance.

PhyCritic for downstream policy training. To validate PhyCritic as a reward signal for guiding downstream policy training, we apply PhyCritic within a self-improving DPO framework. Starting from Qwen2.5-VL-7B-Instruct, we generate 8 responses per prompt on Cosmos-RL data, use PhyCritic-7B to score all ordered response pairs and select best–worst responses as preference pairs for DPO training (learning rate $1e-6$, 5 epochs). The resulting DPO-trained policy is evaluated on CosmosReason1-Bench. As shown in Tab. 8, PhyCritic-guided DPO achieves substantial gains over the base Qwen2.5-VL model and consistently outperforms an answer-verifier baseline (random correct–incorrect pairs), indicating that PhyCritic provides an effective reward signal beyond answer correctness for improving downstream physical reasoning.

Model	overall	AgiBot	HoloAssist	RoboVQA	BridgeV2	RoboFail
Qwen2.5-VL-7B-Instruct	54.3	40.4	50.0	82.5	38.0	57.6
+ answer-verify DPO	57.5	39.0	60.0	86.4	39.0	60.0
+ PhyCritic-7B DPO	60.0	41.0	66.0	90.0	39.0	61.0

Table 8: PhyCritic for guiding DPO training on Qwen2.5-VL-7B-Instruct, evaluated on CosmosReason1-Bench.

Comparison with a general-domain critic. We compare PhyCritic with UnifiedReward-Think Wang et al. (2025), a recent multimodal reward model fine-tuned with RL on general understanding and generation preference data. As shown in Tab. 9, UnifiedReward-Think performs strongly on general-domain reward benchmarks, but shows only marginal gains on physical-domain judgment and reduced physical reasoning performance. In contrast, PhyCritic achieves robust gains on physical-domain judgment and reasoning, and its critic capacity transfers to general-domain judging, despite not being trained on general-domain critic data.

Model	PhyCritic-B.	CosmosReason1-B.	VL-RewardB.	Multimodal-RewardB.
Qwen2.5-VL-7B-Instruct	51.6	54.3	53.2	64.0
UnifiedReward-Think-Qwen-7B	52.4	51.8	73.8	66.1
PhyCritic-7B	68.0	63.9	57.3	65.9

Table 9: Comparison with UnifiedReward-Think. Both models are initialized from Qwen2.5-VL-7B-Instruct.

B. Qualitative Examples

In this section, we provide additional case studies to further compare the judging and physical reasoning behaviors of PhyCritic with those of baseline models.

- Table 10 provides an example of PhyCritic judging a pair of model responses for physical-related evaluation tasks via the self-referential critic process. By first generating its own reasoning to solve the problem and then explicitly referencing this reasoning during critique, PhyCritic can accurately distinguish differences in the quality of physical-related reasoning rather than relying on shallow stylistic cues, leading to more consistent judgments supported by grounded and clear justifications.
- Table 11 illustrates a case where PhyCritic is used as a *policy model* for physical reasoning. The self-referential critic training process further strengthens its physical perception and reasoning capabilities, enabling PhyCritic to generate visually grounded and internally consistent reasoning that leads to reliable final answers.
- Table 12 presents an example of PhyCritic judging a pair of model responses in a general image domain. Although PhyCritic is trained exclusively on physical-related critic data, its physically grounded critic capability transfers effectively to broader visual tasks, enabling reliable and visually grounded judgments even when instructed to use different critic formats and to evaluate general image-based model responses.

PhyCritic in Judging Physical-Related Domains



Question	What should you do next in this driving scenario? A. Continue driving as the truck is stopped and you have the right of way. B. Stop and wait for the truck to move first. C. Reverse to allow the truck to proceed. D. Overtake the truck by entering the opposite lane.
Response 1	<think> In this scenario, the vehicles are on a two-lane road with traffic in both directions. The UPS truck appears to be stationary or moving very slowly, creating a potential obstruction. Although the reason for the stop is unclear, it is blocking the left lane. Option A is unsafe since the truck may be stopped legitimately, and entering its lane could cause a collision. Option B safely checks that the road is clear before moving. Option C is dangerous due to possible traffic behind you. Option D is risky because of oncoming traffic. </think> <answer> B </answer>
Response 2	<think> In this driving scenario, the UPS truck appears to be preparing to turn or has stopped temporarily, possibly to yield to oncoming traffic. However, the road ahead is clear, and there seems to be no immediate obstacle preventing the driver from continuing to drive forward safely. Since the truck is currently stationary and there is no sign indicating that the vehicle must stop, it suggests that the driver has the right of way. </think> <answer> A </answer>

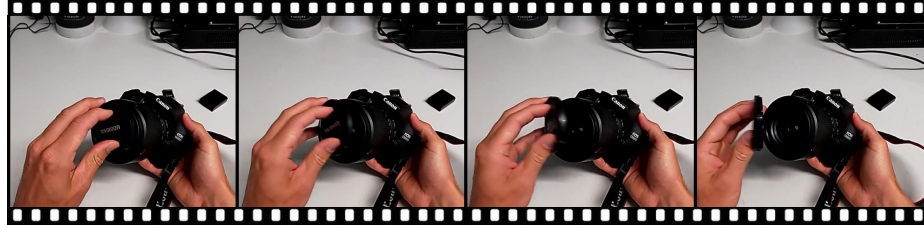
Ground Truth Response 2 > Response 1

Qwen2.5-VL-7B-Instruct	<p>### Response 1: - Reasoning Process Quality: - Truthfulness: The response correctly identifies that the UPS truck is stationary or moving slowly and may obstruct the lane. - Visual Groundedness: It references the image showing the UPS truck on the left side, partially blocking the left lane. - Logical Validity: The reasoning is sound—assessing the truck’s movement, the obstruction risk, and the safest action. - Efficiency and Clarity: The response is clear and concise, avoids repetition, and presents a coherent flow of reasoning. - Final Answer Accuracy: The final answer is accurate; stopping and waiting for the truck to move first is indeed a safe and responsible action.</p> <p>### Response 2: - Reasoning Process Quality: - Truthfulness: The response correctly observes that the truck appears stationary or preparing to turn and notes no immediate obstacle. - Visual Groundedness: It references the visual evidence where the truck is visible on the left side and no traffic ahead obstructs the lane. - Logical Validity: The reasoning is logical, given the truck’s stationary state and the unobstructed lane, and the conclusion about having the right of way is reasonable. - Efficiency and Clarity: The response is clear but slightly less structured than Response 1, as it does not explicitly outline all safety-checking steps. - Final Answer Accuracy: The final answer is accurate; continuing to drive as the truck is stopped and there is no immediate danger is safe and efficient.</p> <p>### Evaluation Summary: Both responses demonstrate factual accuracy, strong visual grounding, a coherent logical flow, and clear, efficient communication. However, Response 1 is more comprehensive and structured, clearly detailing the steps for assessing the situation and identifying the safest action, which makes its reasoning easier to follow. Decision: Response 1 is better</p>
------------------------	---

PhyCritic-7B	<p><pred_think> The image shows a busy street with a UPS truck in the left lane, seemingly stopped or moving slowly. There are other cars and pedestrians on the sidewalks. Given the situation, the truck is not obstructing the right lane where the driver’s car is positioned, allowing for a safe passage. The road ahead is clear, and there is no visible sign that necessitates stopping. Therefore, the most logical action is to continue driving, assuming the truck will eventually move out of the way. </pred_think> <pred> A </pred> <think> Both Model Responses 1 and 2 acknowledge that the truck is stationary and there is no immediate danger from the truck itself, allowing the driver to continue driving. However, they differ in their assessment of the safety and appropriateness of waiting versus continuing. Model Response 1 emphasizes the safety aspect by advising to stop and wait for the truck to move first, citing the possibility of the truck being stopped for a legitimate reason. This response is grounded in the visual evidence that the truck is stationary and not obstructing the path, but it slightly overemphasizes the need to wait due to the lack of visible signs or indications that stopping is necessary. Model Response 2 correctly identifies that the truck is stationary and does not present an obstacle, allowing the driver to continue driving without waiting. This approach aligns well with the visual evidence and the logic of the situation. The road ahead is clear, and the truck is not obstructing the right lane, making it safe to continue driving. My reasoning process aligns more closely with Model Response 2. </think> \boxed{Response 2 is better}</p>
--------------	---

Table 10: Example on PhyCritic-Bench. Qwen2.5-VL-7B focuses on **superficial structures** and produces **uninformative critiques**. In contrast, PhyCritic delivers more aligned judgments by first **identifying the truck’s status during its own reasoning** and then grounding this reasoning in the critic process, allowing it to accurately **capture the quality differences** between the responses.

PhyCritic in Physical Reasoning



Question The overall goal is “The student detaches the cover.”. The agent in the video is currently performing one subtask out of many to complete this goal. For the agent in the video, what is the most plausible next immediate subtask from the choices below?
 A: grab lens_cover
 B: grab dslr
 C: lift dslr
 D: place lens_cover

Answer D

Cosmos-Reason1-7B <think> Okay, let’s see. The user provided the video and the current action. The main goal is for the student to detach the cover. **The agent is currently holding the camera with both hands, preparing to remove the lens cap. So, the current step is that the left hand is adjusting the lens cover. The next logical step after adjusting would be to actually take it off.** Since the lens cap is already being handled, the next action should involve removing it. The options are A (grab lens_cover), B (grab dslr), C (lift dslr), D (place lens_cover). Option A makes sense because once you’ve adjusted or positioned the lens cap correctly, the next step is to grab it to remove it. Option B and C involve grabbing or lifting the DSLR, which might not be necessary if the goal is just to detach the cover. Option D suggests placing the lens cover somewhere else, but since the task is detachment, placing isn’t the immediate next step. So the most plausible is A.</think>
 <answer> A </answer>

PhyCritic-7B <think> The agent is holding a camera with a lens cover attached. The goal is to detach the cover. **The agent has their hand on the lens cover, which suggests they are removing it. The most logical next step would be to place the lens cover down after detaching it.** - Option A (grab lens_cover) seems redundant since the lens cover is already being held. - Option B (grab dslr) does not make sense because the agent should not be grabbing the camera again at this point. - Option C (lift dslr) is incorrect because lifting the DSLR is not necessary to detach the lens cover. - Option D (place lens_cover) aligns with the action needed to properly detach the lens cover. Therefore, the most plausible next immediate subtask is to place the lens cover. </think>
 \boxed{D}

Table 11: Example of PhyCritic on CosmosReason1-Bench [Azzolini et al. \(2025\)](#). For *policy-like* problem solving on physical reasoning tasks, PhyCritic performs more visual-grounded, logically consistent, and efficient reasoning to infer the most plausible next subtask, without introducing **stepwise contradictions**.

PhyCritic in Judging General Visual Domains (Out of Training Distribution)



Question	What color is the tissue box on the counter?
Response 1	The tissue box on the counter in the picture is green.
Response 2	The tissue box on the counter in the picture is white.
Ground Truth	Response 1 > Response 2
Qwen2.5-VL-7B-Instruct	<p>1. Accuracy: - Both responses directly address the color of the tissue box mentioned in the image query. However, only one of them can be correct based on what is visible in the image. - Based on the visual evidence in the image, there is no green present, confirming that Response 1 is not accurate. - The tissue box is indeed white in the image.</p> <p>2. Completeness: - Both responses provide full responses to the question by naming a single color. Complete responses are typically expected to include only relevant responses without additional unnecessary text. Both Response 1 and Response 2 do this well as they solely focus on the color.</p> <p>3. Clarity: - Both responses are straightforward and unambiguous, making it clear what color the tissue box is. - There is no extraneous information, ensuring clarity for the reader.</p> <p>4. Relevance: - Both responses directly relate to the question and the image by providing the color of a specific object (the tissue box) mentioned in the query. - Neither response includes unrelated details, thus being relevant.</p> <p>Overall Judgment: After evaluating the accuracy, completeness, clarity, and relevance, Response 2 is better.</p>
PhyCritic-7B	<p>1. Accuracy: - Response 1: The tissue box on the counter appears green. Response 2: The tissue box on the counter appears white.</p> <p>2. Completeness: - Both responses provide a straightforward response to the question about the color of the tissue box. However, they do not offer additional information that might not be relevant to the question, such as the location or condition of the tissue box.</p> <p>3. Clarity: - Both responses are clear and easy to understand. They directly response the question without ambiguity.</p> <p>4. Relevance: - Both responses directly relate to the question by addressing the color of the tissue box in the picture.</p> <p>Overall Judgment: - Response 1 is better because the color of the tissue box in the image is green, whereas Response 2 states it is white, which is incorrect. ###</p> <p>Detailed Response: - The tissue box on the counter is noticeably green, as seen in the image. It appears compact and situated near the sink. The surrounding surfaces are neutral in color, and the green hue of the tissue box contrasts with the lighter colors of the countertop and the sink.</p> <p>Final Response: Response 1 is better.</p>

Table 12: Example of PhyCritic on VL-RewardBench Lambert et al. (2024). Due to improved physical perception, PhyCritic accurately identifies the tissue box as green and avoids hallucinated colors, producing judgments that better align with human preference. Note that neither image-domain critic data nor the critic prompts used in VL-RewardBench were seen during PhyCritic’ training.

References

- [1] Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiabin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [3] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.
- [4] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024.
- [5] Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Tuomas Rintamaki, et al. Eagle 2.5: Boosting long-context post-training for frontier vision-language models. *arXiv preprint arXiv:2504.15271*, 2025.
- [6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- [7] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.
- [8] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [9] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023.
- [10] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.
- [11] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.
- [12] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [13] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.
- [14] Emily R Lai. Critical thinking: A literature review. *Pearson’s research reports*, 6(1):40–41, 2011.

- [15] Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025.
- [16] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.
- [17] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. 2023.
- [18] Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the association for computational linguistics ACL 2024*, pages 11286–11315, 2024.
- [19] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025.
- [20] Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, et al. V1-rewardbench: A challenging benchmark for vision-language generative reward models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24657–24668, 2025.
- [21] Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, Haoquan Zhang, Wang Bill Zhu, and Kaipeng Zhang. To think or not to think: A study of thinking in rule-based visual reinforcement fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- [22] Chenxi Liu, Tianyi Xiong, Yanshuo Chen, Ruibo Chen, Yihan Wu, Junfeng Guo, Tianyi Zhou, and Heng Huang. Modality-balancing preference optimization of large multimodal models by adversarial negative mining. *arXiv preprint arXiv:2506.08022*, 2025.
- [23] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.
- [24] Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*, 2025.
- [25] Zeyi Liu, Arpit Bahety, and Shuran Song. Reflect: Summarizing robot experiences for failure explanation and correction. *arXiv preprint arXiv:2306.15724*, 2023.
- [26] Run Luo, Lu Wang, Wanwei He, Longze Chen, Jiaming Li, and Xiaobo Xia. Gui-r1: A generalist r1-style vision-language action model for gui agents. *arXiv preprint arXiv:2504.10458*, 2025.
- [27] Ana-Maria Marcu, Long Chen, Jan Hünemann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision*, pages 252–269. Springer, 2024.
- [28] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

- [29] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*, pages 292–308. Springer, 2024.
- [30] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [31] Abby O’Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.
- [32] Lu Qiu, Yi Chen, Yuying Ge, Yixiao Ge, Ying Shan, and Xihui Liu. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. *arXiv preprint arXiv:2412.04447*, 2024.
- [33] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.
- [34] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [35] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 645–652. IEEE, 2024.
- [36] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [37] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.
- [38] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pages 256–274. Springer, 2024.
- [39] Jingyu Song, Zhenxin Li, Shiyi Lan, Xinglong Sun, Nadine Chang, Maying Shen, Joshua Chen, Katherine A Skinner, and Jose M Alvarez. Drivecritic: Towards context-aware, human-aligned evaluation for autonomous driving with vision-language models. *arXiv preprint arXiv:2510.13108*, 2025.
- [40] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.
- [41] Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*, 2024.

- [42] BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Xiansheng Chen, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, et al. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025.
- [43] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.
- [44] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37: 87310–87356, 2024.
- [45] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.
- [46] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20270–20281, 2023.
- [47] Xiyao Wang, Chunyuan Li, Jianwei Yang, Kai Zhang, Bo Liu, Tianyi Xiong, and Furong Huang. Llava-critic-r1: Your critic model is secretly a strong policy model. *arXiv preprint arXiv:2509.00676*, 2025.
- [48] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025.
- [49] Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*, 2025.
- [50] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.
- [51] Yibin Wang, Yuhang Zang, Feng Han, Yujie Zhou, Jiazi Bu, Cheng Jin, and Jiaqi Wang. Unified personalized reward model for vision generation. *arXiv preprint arXiv:2602.02380*, 2026.
- [52] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. *arXiv preprint arXiv:2501.04003*, 2025.
- [53] Tianyi Xiong, Yi Ge, Ming Li, Zuolong Zhang, Pranav Kulkarni, Kaishen Wang, Qi He, Zeying Zhu, Chenxi Liu, Ruiibo Chen, Tong Zheng, Yanshuo Chen, Xiyao Wang, Renrui Zhang, Wenhui Chen, and Heng Huang. Multi-crit: Benchmarking multimodal judges on pluralistic criteria-following, 2025. URL <https://arxiv.org/abs/2511.21662>.
- [54] Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13618–13628, 2025.
- [55] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

- [56] Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal rewardbench: Holistic evaluation of reward models for vision language models. *arXiv preprint arXiv:2502.14191*, 2025.
- [57] Qinghao Ye, Xianhan Zeng, Fu Li, Chunyuan Li, and Haoqi Fan. Painting with words: Elevating detailed image captioning with benchmark and alignment learning. *arXiv preprint arXiv:2503.07906*, 2025.
- [58] En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025.
- [59] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024.
- [60] Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, et al. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19985–19995, 2025.
- [61] Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, et al. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*, 2025.
- [62] Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yanpeng Zhou, Yujie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*, 2025.
- [63] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander G Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 694–717, 2025.
- [64] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023.
- [65] Yi-Fan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, et al. R1-reward: Training multimodal reward model through stable reinforcement learning. *arXiv preprint arXiv:2505.02835*, 2025.
- [66] Yi-Fan Zhang, Haihua Yang, Huanyu Zhang, Yang Shi, Zezhou Chen, Haochen Tian, Chaoyou Fu, Haotian Wang, Kai Wu, Bo Cui, et al. Basereward: A strong baseline for multimodal reward model. *arXiv preprint arXiv:2509.16127*, 2025.
- [67] Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, et al. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv preprint arXiv:2502.10391*, 2025.
- [68] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's "aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025.
- [69] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

- [70] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.