

PhyCritic: Multimodal Critic Models for Physical AI

Supplementary Material

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 (ours)	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 [48], 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 [1] 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 [2] itself as the judge for pairwise knockout; and 3) using the physical RL-finetuned Cosmos-Reason1-7B [1] as the judge.

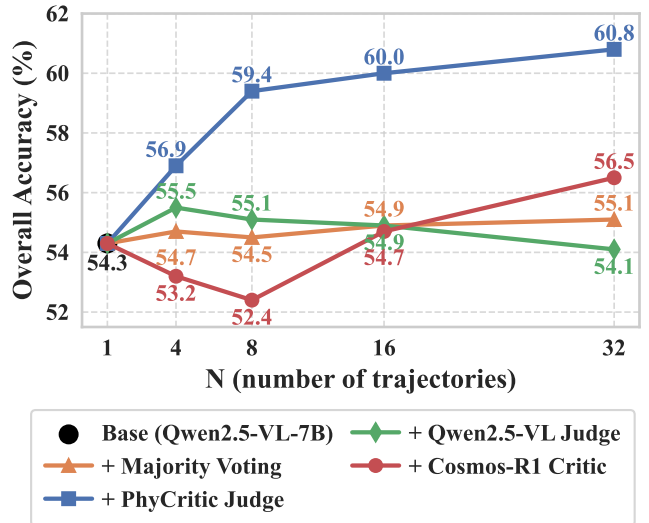


Figure 4. Comparison of Best-of- N ensemble mechanisms on CosmosReason1-Bench. Using PhyCritic-7B as the judge consistently improves the base Qwen2.5-VL-7B-Instruct model.

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.

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.

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.

Comparison with a general-domain critic. We compare PhyCritic with UnifiedReward-Think [50], a recent multimodal reward model fine-tuned 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.	Cosmos-B.	VL-Reward.	MM-Reward.
Qwen2.5-VL-7B	51.6	54.3	53.2	64.0
UnifiedReward-Think	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

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.

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.

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

PhyCritic-7B

<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}

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 [1]. 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 respond to 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. ### 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 [16]. 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.