VILBENCH: A Suite for Vision-Language Process Reward Modeling

Haoqin Tu[1] **Weitao Feng**[1*] **Hardy Chen**[1] **Hui Liu**[2]
Xianfeng Tang[2] **Cihang Xie**[1]
[1]UC Santa Cruz [2]Amazon Research

Abstract

Process-supervised reward models serve as a fine-grained function that provides detailed step-wise feedback to model responses, facilitating effective selection of reasoning trajectories for complex tasks. Despite its advantages, evaluation on PRMs remains less explored, especially in the multimodal domain. To address this gap, this paper first benchmarks current vision large language models (VLLMs) as two types of reward models: output reward models (ORMs) and process reward models (PRMs) on multiple vision-language benchmarks, which reveal that neither ORM nor PRM consistently outperforms across all tasks, and superior VLLMs do not necessarily yield better rewarding performance. To further advance evaluation, we introduce VILBENCH, a vision-language benchmark designed to require intensive process reward signals. Notably, OpenAI's GPT-4o with Chain-of-Thought (CoT) achieves only 27.3% accuracy, challenging current VLLMs. Lastly, we preliminarily showcase a promising pathway towards bridging the gap between general VLLMs and reward models—by collecting 73.6K vision-language process reward data using an enhanced tree-search algorithm, our 3B model is able to achieve an average improvement of 3.3% over standard CoT and up to 2.5% compared to its untrained counterpart on VILBENCH by selecting OpenAI o1's generations. We will release our code, model, and data at https://ucsc-vlaa.github.io/ViLBench.

1 Introduction

Reward models (RMs) play a crucial role in aligning model outputs with human preferences, benefiting Large Language Models (LLMs) in both training and inference stages (Schulman et al., 2017; Bai et al., 2022; Ouyang et al., 2022; Rafailov et al., 2023; Chen et al., 2025). The most popular RMs include output reward models (ORMs) and process-supervised reward models (PRMs). While ORMs assess responses at the final output level (Zheng et al., 2023; Stiennon et al., 2020), PRMs provide detailed, step-wise feedback, making them particularly useful for complex reasoning tasks (Lightman et al., 2023; Wang et al., 2023; Zhang et al., 2025b). Despite their advantages in the language domain, the application of PRMs in multimodal contexts remains underexplored, with most vision-language RMs following the ORM paradigm (Xiong et al., 2024; Lee et al., 2024a; Zang et al., 2025).

To advance the study of vision-language process reward modeling, this paper presents a comprehensive suite of contributions encompassing (1) a benchmarking study of state-of-the-art VLLMs as reward models, (2) a newly curated dataset designed for fine-grained step-wise reward evaluation, and (3) an advanced vision-language PRM trained on large-scale vision-language step reward data. Our goal is to provide a deeper understanding of the effectiveness of current vision-language reward models and to pave the way for future improvements in multimodal step-wise evaluations.

As our first contribution, we evaluate seven VLLMs (six open-weight and one private) following MLLM-as-a-judge (Chen et al., 2024a; Ge et al., 2023) across five challenging vision-language tasks. This benchmarking effort systematically analyzes the models' rewarding capabilities in various domains, revealing several key insights. For example, we observe that neither ORM nor PRM consistently outperforms the other across all tasks, indicating that different reasoning structures benefit from different rewarding approaches (Zhang et al., 2025b). Additionally, we find that better VLLMs do not always translate to superior reward capabilities, suggesting that rewarding and generation abilities are not inherently correlated. Our results also highlight that in specific domains such as text-dominant tasks, PRMs

[*]Work done during an internship at UCSC.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 6776–6791
November 4-9, 2025

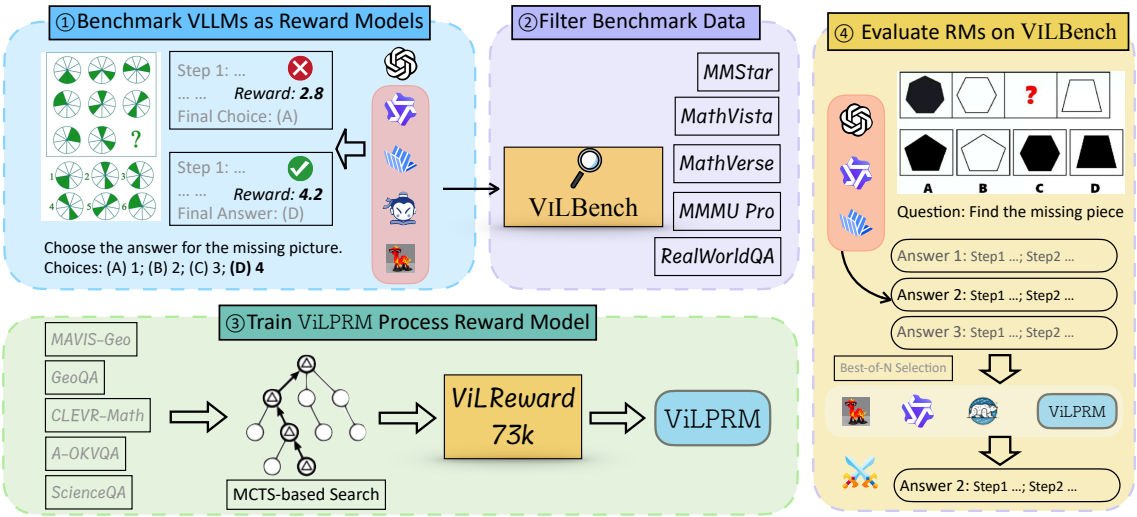

Figure 1: We present a suite of vision-language process reward modeling. We first benchmark current vision-language models as different reward models, and present ViLBench that requires intensive step-wise reward. Then we collect 73K+ preference reward data to train a vision-language process reward model ViLPRM that performs better than other baselines on ViLBench.

is able to provide a greater advantage, suggesting their strong potential in tasks requiring intricate, step-wise reasoning.

Next, we introduce ViLBench, a vision-language benchmark that demands step-wise reward feedback. Current vision-language benchmarks primarily focus on evaluating final outputs, which limits their ability to distinguish between improvements driven by ORMs and PRMs. To address this limitation, we curate a dataset of 600 examples that emphasize the necessity of step-wise feedback. Our filtering protocol assembles judges from six open-weight VLLMs to select examples that require fine-grained rewards beyond simple correctness assessments. Notably, advanced models like GPT-4o achieve only 27.3% accuracy on ViLBench, benefiting 3.0% more from PRM-driven step-wise rewards than from ORMs, underscoring our benchmark's difficulty and its emphasis on fine-grained reward assessment.

Lastly, as a preliminary but promising step towards bridging the gap between general VLLMs and vision-language PRMs, we employ an enhanced multimodal Monte Carlo Tree Search (MCTS) (Zhang et al., 2025a) to generate ViLReward-73K, a dataset of 73.6K stepwise vision-language reward samples drawn from five training datasets. With this dataset, we train a 3B vision-language PRM that significantly improves the evaluation accuracy of step-wise rewards. Specifically, this model substantially surpasses existing PRMs, achieving an average improvement of 3.3% over standard CoT approaches and up to 2.5% compared to its untrained counterpart on ViLBench. We also discuss potential challenges and future directions to conclude the paper.

Model Name	LLM	Model Size	Date
InternLM-X2.5 (2024a)	InternLM2	7B	07/2024
LLaVA-OneVisoin (2024a)	Qwen2	7B	08/2024
Qwen2-VL (2024c)	Qwen2	7B	08/2024
InternVL-2.5 (2024c)	Qwen2.5	8B	12/2024
Qwen2.5-VL (2025)	Qwen2.5	3B, 7B	02/2025
GPT-4o (2024)	Unknown	Unknown	05/2024

Table 1: VLLMs used as different RMs for ViLBench.

2 Part I: Benchmarking VLLMs as Reward Models

VLLMs are demonstrating increasing strength across a variety of tasks. One effective way to further enhance their performance is by evaluating their test-time scaling ability. To assess the step-wise critique capabilities of VLLMs, we benchmark seven different models (see Table 1 for model details) following the paradigm of LLM-as-a-judge (Chen et al., 2024a; Zheng et al., 2023) on five widely used vision-language tasks: MMStar (Chen et al., 2024b), MathVista (Lu et al., 2024), MathVerse (Zhang et al., 2024b), MMMU Pro (Yue et al., 2024), and RealWorldQA (Grok-1.5 Team, 2024). To further explore their inference-time scaling potential, we adopt the Best-of-N (BoN) setting, where VLLMs select the best response from a pool of N candidate responses (Wang et al., 2022; Lightman et al., 2023). In detail, we adopt GPT-4o as the base so-

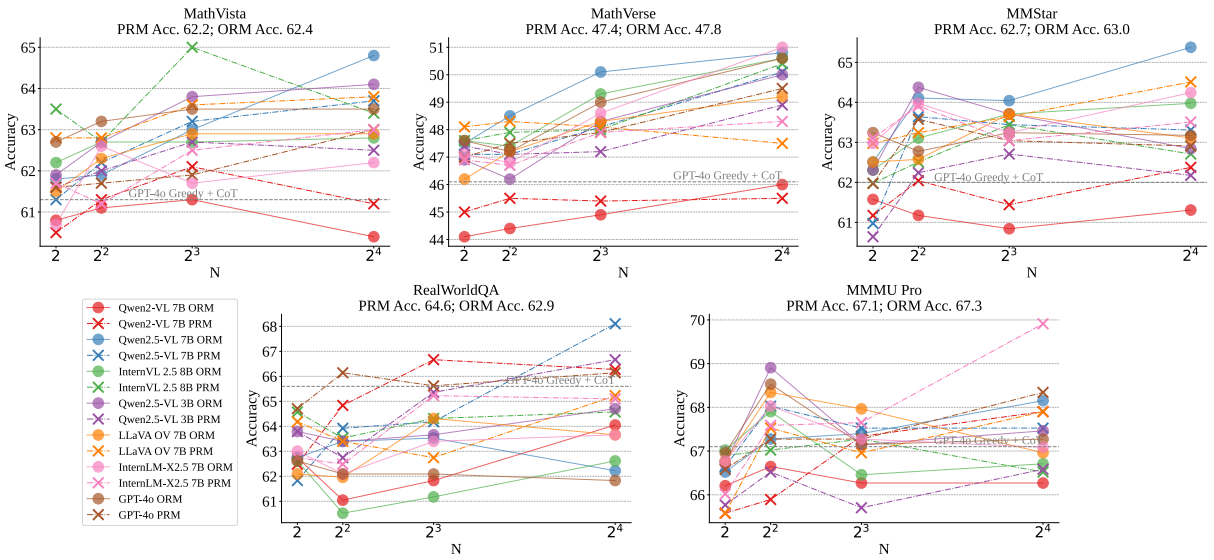

Figure 2: Benchmark results of 7 different VLLMs as reward models on 5 vision-language benchmarks. The base solution generator is GPT-4o. We report the average PRM and ORM scores in subtitles.

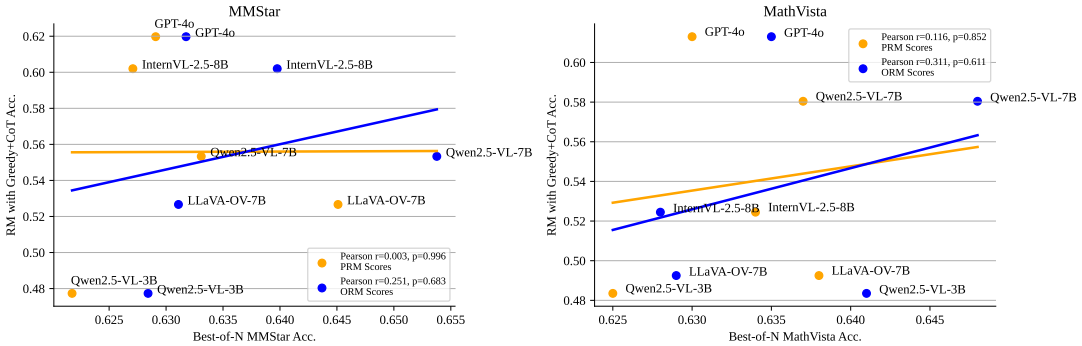

Figure 3: Correlations between the model performance and its reward performance on MMStar and MathVista. The rewarding performance is averaged over 4 different N in the BoN selection with GPT-4o as the generator.

lution sampler to sample 2^4 solutions given one question. Then we incorporate different VLLMs as the deterministic scorer to pick the best response among the candidates by assigning scores between 1 to 5 to each reasoning step. More details about prompt and model generation settings can be found in Appendix A. Through this approach, we uncover four key insights:

Findings 1: Neither ORM nor PRM excels across all vision-language tasks.

Among the five VL benchmarks in Figure 2, VLLMs as ORMs slightly outperform fine-grained PRMs in four cases, with an average margin of 0.3%. However, on RealWorldQA, a challenging VL task involving daily life images, knowledge, and reasoning, the PRM surpasses ORM by an average of 1.7%. Interestingly, the four datasets where ORM performs better (MathVista, MathVerse, MMStar, and MMMU Pro) primarily feature formal reasoning and mathematical problems, whereas RealWorldQA focuses on real-world scenarios. This contrasts with prior findings in the language domain, where PRMs have been shown to offer better guidance than ORMs for language-only math and reasoning tasks (Wang et al., 2023; Lightman et al., 2023; Wang et al., 2024b). One possible explanation is that current VLLMs are predominantly optimized on visual understanding tasks, rather than step-wise rewarding tasks.

Reward models consistently enhance performance across all five benchmarks compared to CoT greedy decoding. As the BoN candidate selection expands, RMs become increasingly effective in boosting performance. However, the impact varies across benchmarks. For example, in RealWorldQA, only four RMs at $N = 2^4$ improve the base model beyond CoT. In contrast, for the remaining benchmarks, most RMs outperform CoT when $N > 2^2$. Notably, on MathVerse, nearly all VLLMs enable GPT-4o to surpass its CoT decoding at $N \geq 2$, ex-

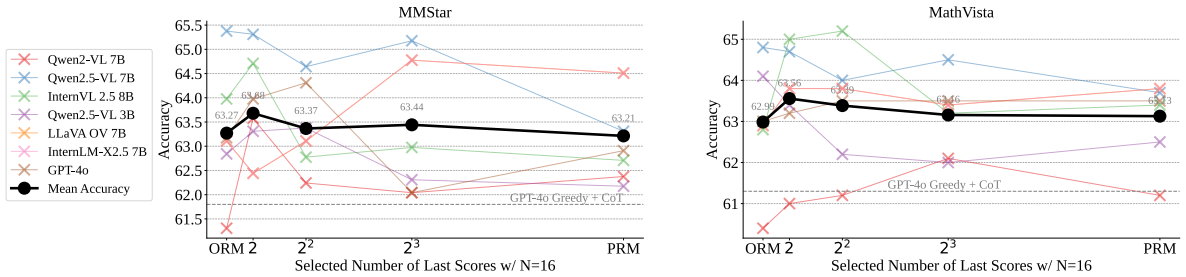

Figure 4: Model performance with the last n of step rewards selected under the Best-of-N paradigm.

cept for LLaVA-OneVision. This observation suggests that vision-language reward models may be less effective for complex visual perception tasks than for formal reasoning challenges.

Method	Text-dominant	Visual-dominant
Greedy	58.9	51.0
ORM	62.2 +3.3	49.0 -2.0
PRM	62.0 +3.1	48.9 -2.1

Table 2: Average accuracy over 7 RMs on text or visual dominant examples in MathVerse using ORM or PRM.

Findings 2: Better vision-language models do not necessarily lead to better reward models.

Previous research has demonstrated that stronger VLLMs tend to produce better ORMs (Li et al., 2024b). However, this correlation does not necessarily hold for PRMs. To examine this, we plot the correlation between a model's greedy performance and its rewarding ability on MMStar and MathVista in Figure 3. In both tasks, ORMs exhibit higher Pearson correlation scores between general VLLM capability and rewarding ability, reinforcing the relationship between these two attributes. In contrast, under the PRM setting, the correlation is notably weak, averaging just 0.06%. This suggests that superior VLLM performance does not directly translate to stronger rewarding capabilities, particularly in process supervision. Notably, LLaVA-OneVision and Qwen2.5-VL rank highest as PRMs on these tasks. A surprising observation is that GPT-4o, the strongest VLLM among the tested models, underperforms as both an ORM and PRM in the deterministic scoring setting. This may be attributed to GPT-4o's tendency to overrate responses, introducing bias in certain reward tasks (Song et al., 2023; Herrera-Berg et al., 2023).

Findings 1 and 2 highlight the need for the development of more robust and generalizable PRMs in the vision-language domain.

Findings 3: The best practice for a vision-language reward model is to use rewards from the last few steps.

Beyond PRM and ORM, alternative RMs exist that balance between selecting only the final step (ORM) and considering all step rewards (PRM) by incorporating the last n step scores. We conduct experiments using the average of the last n step rewards (*i.e.*, $n \in [1, 2, 2^2, 2^3, \text{all}]$) as the final reward signal on MMStar and MathVista. As shown in Figure 4, the most effective approach consistently falls between ORM and PRM, with the optimal performance achieved by averaging the last 2 or 4 step rewards. Specifically, when using the last 2 step rewards as the final signal, the selected answer achieves the highest average accuracy, improving by 0.41% and 0.57% over the ORM setting on the two benchmarks. This finding suggests an improved strategy for selecting vision-language reward signals, striking a balance between ORM and PRM for enhanced performance.

Findings 4: VLLMs as reward models provide more benefits on text-dominant examples.

On MathVerse, certain examples require a stronger focus on textual reasoning (text-dominant), while others rely more on visual understanding (visual-dominant). We report the average performance of RMs on these two subsets in Table 2. The results indicate that vision-language RMs provide greater benefits to VLLMs on text-dominant examples but may negatively affect performance on visual-dominant ones. Since reward signals are integrated into the language generation process at the textual level, this finding suggests that current VLLMs exhibit stronger textual-level critique capabilities. This again, highlights the need for the development of specialized vision-language reward models to better handle visually intensive tasks.

	Dataset Souce	Size	Split	Ori. Size
ViLBench	MMStar	150	val	1,500
	MathVista	150	testmini	1,000
	MathVerse	100	testmini*	1,000
	MMMU Pro	100	test	1,592
	RealWorldQA	100	test	756
	Sum	600	test	5,848

Table 3: An overview of ViLBench. * means that we only sample 1000 entries from the testmini of MathVerse. Ori. Size is the original size of the dataset.

Dataset Source	Class	Size	PR Size
MAVIS-Geo (2024c)	Math	3,093	25,829
GeoQA170K (2021)	Math	8,063	31,406
CLEVR-Math (2022)	Math	957	1,425
A-OKVQA (2022)	General	2,044	9,241
ScienceQA (2022)	General	2,769	5,659
Sum	Math&General	16,926	73,560

Table 4: Statistics of ViLReward-73K, a vision-language process reward preference dataset. We show the initial size of the data source (Size) as well as the size of the process reward instance (PR Size).

2.1 ViLBench: A Vision-Language Benchmark Requiring Intensive Reward Feedback

As what we found previously, existing vision-language benchmarks do not require intensive feedback from RMs like vision-language PRMs. To address this, we leverage six open-weight VLLMs to filter samples where they perform well as PRMs but worse as ORMs under the BoN setting. To be concrete, we evaluate the average performance of ORMs and PRMs using 16 response candidates, providing the RMs with a broader selection. We introduce a PRM-preference indicator based on the average PRM and ORM scores, denoted as S_{prm} and S_{orm}, respectively. This indicator is calculated as $S = S_{\text{prm}} - S_{\text{orm}}$, allowing us to rank all samples across the five tested benchmarks according to their scores. In Table 3, we illustrate how we sample varying amounts of data from each task to construct our final dataset, ViLBench. For the evaluation metric, since each question is paired with a ground truth answer, we use accuracy between predicted answers and the ground truth as the final metric. We provide details of answer extraction and evaluations in Appendix C.

3 Part II: ViLPRM: A Vision-Language Process Reward Model

3.1 Vision-Language Preference Data Preparation

Data Selection and Filtering. Process preference data have been proven to be effective in training RMs in specific domains like math and logical reasoning. However, in scenarios that demand challenging visual perception understandings, the potential of PRMs remains underexplored. In order to generalize the vision-language PRM to subjects other than just math, we consider collecting challenging VL data from general visual perception and math datasets. We follow three rules to filter data for the process reward model training: (1) Unique image content for diverse visual features; (2) Challenging questions that elicit model reasoning for better process scoring; (3) Diverse source of the data for generalizing RM abilities in various domains.

In detail, we draw samples from 5 vision-language datasets consisting of 3 vision-language math data and 2 challenging visual perception tasks. For math domain:

- MAVIS-Geometry (Zhang et al., 2024c) is a dataset consisting of visual geometry questions that use GPT-4 to rewrite or generate geometry visual problems and solutions. There are four different difficulty levels, and we select the hardest two levels to sample 5,000 data as our metadata.

- GeoQA170K (Gao et al., 2023) contains over 170K geometric image-caption and question-answer pairs, building on GeoQA+ (Cao and Xiao, 2022) and GeoQA3K (Lu et al., 2021). We sample one question from each unique images from the data, resulting in 8,063 examples in total.

- CLEVR-Math (Lindström and Abraham, 2022) is a synthesized VQA dataset based on CLEVR (Johnson et al., 2017) that includes math word problem solving. We only consider 957 questions with distinct images and need multi-hop reasoning in the dataset.

And for the visual perception domain:

- A-OKVQA (Schwenk et al., 2022) contains question-answering problems about natural images, we select the questions that cannot be answered directly from the *difficult to direct answer* split of the dataset (1,544 exam-

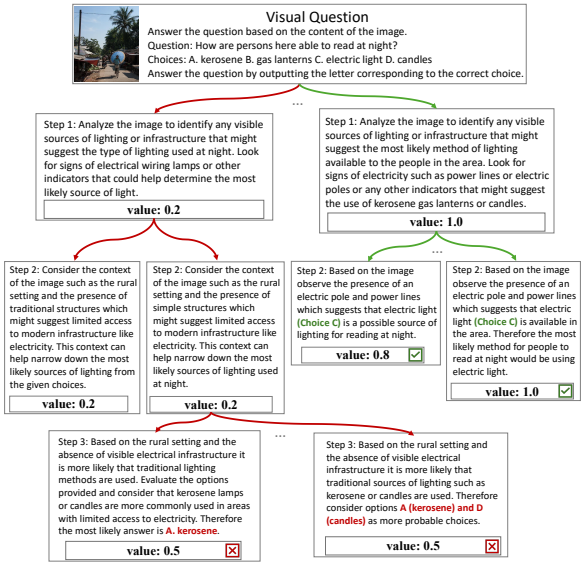

Figure 5: An example of partial MCTS tree we constructed for ViLReward-73K. The metadata is from A-OKVQA. We mark value scores from the preference data at each node.

ples) and draw 500 samples from the rest data points.

- ScienceQA (Lu et al., 2022) is a comprehensive dataset with 21K examples in science, the data is categorized into 12 grades based on the difficulty level. We use data that is harder than grade 7 as our metadata.

We present the detailed data volumes in Table 4.

MCTS Data Searching Engine. Instead of assigning coarse-grained binary scores (*e.g.*, "good" or "bad") to each process (Lightman et al., 2023; Wang et al., 2023), we adopt ReST-MCTS* (Zhang et al., 2025a) to assign fine-grained value v between 0 to 1 to each reasoning step as the reward value. Different from the vanilla ReST-MCTS*, we enable visual input for the policy model, allowing the model to answer questions based on visual inputs.

Following the standard MCTS tree construction, there are four major phases while tree node expansion during search: root node selection, node expansion, route simulation and value backpropagation. For the final answer evaluation, we use GPT-4o as the judge to evaluate the final output of tree search. We mainly introduce the calculation for node value and leave other details, including reasons for selecting different data sizes, in Appendix B. We define the quality value $v_k \in [0, 1]$ of a partial solution $p_k = [s_1, s_2, \ldots, s_k]$ to evaluate its progress toward a correct answer. v_k reflects

Model	3B	8B	GPT-4o	o1
ORM	30.8	27.8	28.1	34.9
PRM	31.1 (+0.3)	28.7 (+0.9)	31.1 (+3.0)	34.9 (+0.0)

Table 5: The average accuracy of 3 open-weight VLLMs as reward models on the proposed benchmark. PRMs have more advantage in enhancing model performance than ORMs on our ViLBench.

the correctness and contribution of each step s_i, higher v_k indicates a greater likelihood of being correct. The reasoning distance m_k is the minimum steps needed to reach the correct answer from p_k, estimated via simulations as it cannot be directly computed. By introducing a *weighted reward* w_{sk} for step s_k, incorporating m_k and the reward r_{sk}:

$$w_{sk} = \left(1 - \frac{v_{k-1}}{m_k + 1}\right)(1 - 2r_{sk}), \quad k = 1, 2, \ldots \tag{1}$$

The quality value updates iteratively:

$$v_k = \begin{cases} 0, & k = 0, \\ \max(v_{k-1} + w_{sk}, 0), & \text{otherwise.} \end{cases} \tag{2}$$

Here, $m_k = K - k$, where K is the total steps in solution s. Since reasoning for visual problems is often simpler, we modified the prompt to require the model's output steps to be more granular, ensuring that the model does not directly output the answer at the very beginning. We present one example of the reward value tree in Figure 5.

3.2 Process Reward Model Training

Based on the derived ViLReward-73K preference data, we train a 3B vision-language PRM, `ViLPRM`.

Model Architecture. `ViLPRM` is built upon a 3B VLLM Qwen2.5-VL (Bai et al., 2025). We follow the common practice of PRM to use the pre-trained weights of Qwen2.5-VL for most of the parts, such as the visual encoder and the MLP projector, but append a linear layer to output a scalar score after the language head (Dong et al., 2024a; Wang et al., 2024a). We do not consider generative score modeling due to efficiency concerns. Since the base model Qwen2.5-VL has been aligned with massive visual-language data, our reward model only requires learning to classify good or bad steps in vision-language reasoning trajectories and avoids using other pre-training data for modality alignment. We formalize the model input and output as: given the input question x and the model reasoning response y, the score head f transforms the logits

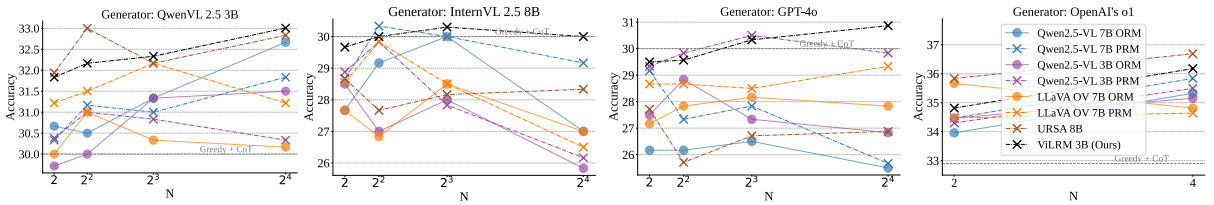

Figure 6: Accuracy results of different VLLMs using various RMs under the best-of-n strategy on ViLBench.

feature of the last token into a scalar $r(x, y)$. This scalar value $r(x, y)$ serves as the predicted reward score for the inputs.

Training. Unlike previous works that only assign binary scores to the RM input, we have detailed scores for each input that can classify step responses better. We use the Mean Square Error (MSE) loss between the ground truth reward and the predicted one to update the ViLPRM. We add more details about training in Appendix D.

4 Validating ViLPRM on ViLBench

To validate the effectiveness of ViLPRM, we conduct experiments under various settings on ViLBench.

4.1 Experimental Setups

We choose the test-time scaling strategy to verify the functionality of different RMs. On our filtered ViLBench, we first use 4 different VLLMs with various model size as the solution sampler, *i.e.*, Qwen2.5-VL-3B (Bai et al., 2025), InternVL-2.5-8B (Chen et al., 2024c), GPT-4o (OpenAI, 2024) and o1 (OpenAI, 2025). For all models except o1, we sample 16 candidate responses for BoN selection. While we sample 4 solutions for o1 due to the high cost of this sampling process.

For PRMs, we include four VLLMs — Qwen2.5-VL-3B (Bai et al., 2025), Qwen2.5-VL-7B (Bai et al., 2025), LLaVA-OneVision-7B (Li et al., 2024a) and a recent vision-language PRM URSA-RM (URSA for short) (Luo et al., 2025) as baselines, where URSA is a concurrent vision-language PRM developed using a base model of 8B parameters and trained on over 1000K carefully designed preference reward data. It has shown great improvements in the multimodal math problems, but lacks the capacity to generalize to more general vision-language tasks.

4.2 Results and Analysis

ViLBench requires more intensive reward feedback than other VL benchmarks. To confirm that our filtered ViLBench requires more fine-grained step rewards beyond simple output rewards, we present the average accuracy across three different VLLMs (QwenVL2.5 3B, LLaVA-OneVision 7B, and QwenVL2.5 7B) used as ORMs or PRMs with four solution samplers in Table 5. In most cases, PRMs enhance model performance more effectively than ORMs, yielding an average improvement of 1.4%. However, one notable exception occurs when selecting o1's responses, where no significant difference is observed between ORM and PRM. This may be due to o1's final output steps lacking sufficient detail, with its internal reasoning process hidden from users, making it less effective for prompt engineering (OpenAI, 2025).

Model	2	4	8	16	Avg.
QwenVL 2.5 3B (2025)	30.7	31.5	29.7	28.8	30.2
LLaVA OV 7B (2024a)	30.7	31.2	29.7	29.0	30.2
QwenVL 2.5 7B (2025)	29.0	30.8	29.0	26.7	28.9
URSA (2025)	31.0	30.8	30.0	30.3	30.6
ViLPRM (Ours)	**31.5**	**32.0**	**31.0**	**31.3**	**31.5**

Table 6: The average accuracy over four solution generators using different PRMs under different BoN setups.

ViLPRM performs better than other VL PRMs. Figure 6 presents the performance of RMs across four Best-of-N (BoN) settings with different solution samplers. As the number of response candidates increases, RMs generally enhance model performance. Specifically trained vision-language PRMs consistently improve results over the CoT strategy for $N \geq 2^3$. However, VLLM-based RMs may negatively impact performance when N becomes too large. For instance, among three sets of responses, three VLLMs acting as RMs exhibit varying degrees of degeneration in candidate selection from 2^3 to 2^4. In contrast, the two vision-language PRMs demonstrate greater consistency in identifying superior responses. This finding reinforces our claim that current VLLMs are not yet robust enough to serve effectively as reward models.

Table 6 reports the average performance of different RMs. Compared to the larger PRM URSA, which is trained on over ten times more data, our ViLPRM achieves a superior average performance

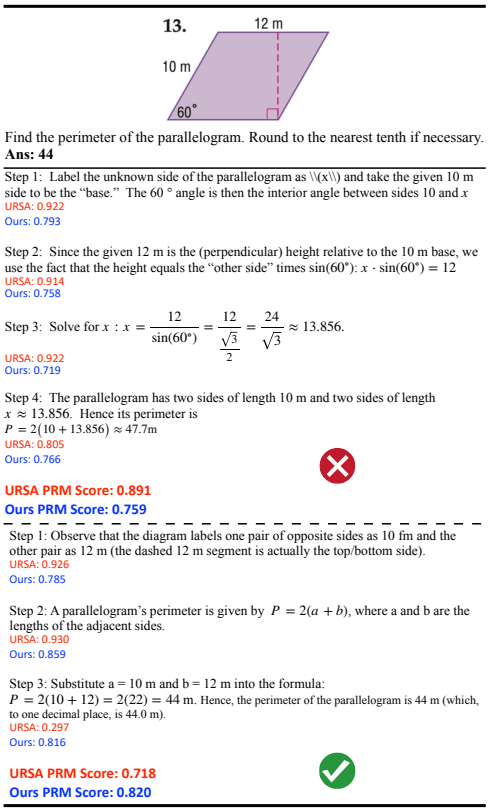

Figure 7: An example of process scores provided by URSA (Luo et al., 2025) and our ViLPRM. We mark different scores with different colors.

by 0.9% and outperforms its untrained 3B VLLM counterpart by 1.3%. Additionally, RMs consistently enhance the performance of o1, a model known for leveraging an internal thinking process as an efficient test-time scaling technique. This further underscores the importance of developing reliable reward models, even for models with built-in reasoning capabilities. We provide more evaluation results covering a wider range of reward models in Appendix F.

Examples for the Reward Task. In Figure 7, we present an example of using our ViLPRM and URSA for selecting the best response from OpenAI o1's responses. In the example, URSA prefers more steps in the reasoning and even with wrong trajectories, while our ViLPRM can choose accurate solutions. This is likely because URSA was trained on a massive amount of math reasoning data and may develop the preference for complex rather than accurate reasoning steps (Liu et al., 2025). We also present another example about medical reasoning in the Appendix E, which verifies that the proposed vision-language PRM has the capacity to also perform well beyond just math or reasoning tasks.

Model	Size	Accuracy
InternLM-XComposer2.5-Reward	7B	33.8
LLaVA OV	7B	36.5*
Qwen2 VL	7B	33.9*
LLaVA-Critic	7B	47.4*
VisualPRM	8B	27.4
GPT-4o-mini (2024-07-18)	-	44.8*
Qwen-VL-Max	-	48.1*
ViLPRM (Ours)	3B	46.3

Table 7: Model performance on VL-RewardBench, results with * are taken directly from its open leaderboard.

Discussions We provide a detailed discussion of PRM limitations in Appendix G. PRMs work well in structured tasks but falter with unclear step segmentation or uniform step weighting; adaptive evaluation and better segmentation could help. We also discuss the point that multimodal RMs also lack cross-task robustness, calling for more diverse training and broader evaluation beyond accuracy.

5 ViLPRM on Reward Benchmarks

To further validate ViLPRM, we conduct experiments on VL-RewardBench[1] (Li et al., 2024b) and VisualProcessBench (Wang et al., 2025).

In Table 7, we provide results on VL-RewardBench, a benchmark for output-level reward modeling. We segment responses by sentence and average the scores to adapt it to stepwise evaluation. Despite the benchmark not being tailored for PRMs, ViLPRM (3B) outperforms GPT-4o-mini (2024-07-18) and remains competitive with Qwen-VL-Max and LLaVA-Critic (7B), demonstrating strong generalization and effectiveness.

Model	Size	Accuracy
Qwen2.5 VL	3B	11.7
LLaVA OV	7B	4.5
Qwen2.5 VL	7B	63.9
VisualPRM	8B	66.5
ViLPRM (Ours)	3B	68.8

Table 8: Model performance on VisualProcessBench. We only consider correct and incorrect reasoning steps for evaluation following the original implementation.

Table 8 presents results on VisualProcessBench, which explicitly targets process-level visual reasoning. Following the official setup, we threshold final scores at 0 to classify steps as correct or incorrect. We observe that ViLPRM shows similar performance to VisualPRM with 8B parameters and over 5 times more training data (400K for VisualPRM and 73K for ViLPRM). For three VLLMs

[1] https://huggingface.co/spaces/MMInstruction/VL-RewardBench

we tested, we simply ask the model to judge if the current solution step logically follows and is factually consistent with the image and question.

Across both VL-RewardBench and VisualProcessBench, ViLPRM demonstrates robust multimodal reward capabilities by providing more accurate step-wise feedback and more practical guidance for improving VLLM performance.

6 Related Works

Reward Benchmark. There are plenty of works put their emphasis in the text-only reward benchmarks (Lambert et al., 2024b; Liu et al., 2024; Zhou et al., 2024a), and some of them are specifically designed for PRMs (Song et al., 2025; Zhang et al., 2025b). When shifting to vision-language domain, traditional VLLM evaluation mainly focuses on the general abilities of the model, including multiple aspects like knowledge, reasoning, fairness, and safety (Lee et al., 2024b; Tu et al., 2023; Yue et al., 2024; Lu et al., 2024; Zhang et al., 2024b). VL-RewardBench (Li et al., 2024b) is the first work that sources reinforcement learning preference data and rewrites knowledge-intensive vision-language samples to form a diverse benchmark. Our proposed VILBENCH fills the gap of benchmarking PRMs in vision-language domain.

Reward Modeling. Reward models are important for guiding AI models at both training and inference stages. There are typically three different forms of RMs: (1) discriminative RM treats the rewarding task as token classification. It usually leverages a linear head to fit the reward score via a regression loss (Stiennon et al., 2020; Ouyang et al., 2022). (2) LLM-as-a-judge leverages the generative ability of language models to output feedback in the form of text, often a critique or explanation of why a certain output is good or bad (Xiong et al., 2024; Lee et al., 2024a; Zheng et al., 2023). (3) Implicit RMs that are models optimized using DPO that the predicted log probabilities are interpreted as implicit reward signal (Lambert et al., 2024a; Ivison et al., 2023; Zhou et al., 2024b). In our work, the proposed ViLPRM is a discriminative RM.

7 Conclusion

We introduce a comprehensive suite for vision-language process reward modeling (PRM), and evaluate seven VLLMs as reward models. Our results show VLLM-based PRMs improve stepwise reasoning in structured vision-language tasks but falter in visual-dominant scenarios, underscoring the need for adaptive step evaluation. We collect 600 examples as VILBENCH where PRMs perform better than ORMs and curate ViLReward-73K with 73.6K step-wise rewards, enabling ViLPRM to outperform other models on VILBENCH by 3.3%.

Acknowledgment

We thank the Microsoft Accelerate Foundation Models Research Program for supporting our computing needs.

Limitations

VILBENCH extends beyond traditional language-only or multimodal output reward models by introducing a multimodal process reward model, along with new benchmarks, datasets, and a dedicated reward model. While we address the gap in current benchmarking efforts for evaluating VLLMs as reward models (especially PRMs) and conduct experiments on five widely-used tasks, we acknowledge that other vision-language tasks may have been overlooked. For preference data collection, we adopt a fully automated pipeline to gather 73K vision-language samples, but do not explore integrating language-only process preference data such as PRM800K (Lightman et al., 2023), RLHFlow (Dong et al., 2024b) or Math Shepherd (Wang et al., 2023), which have proven effective for training general-purpose vision-language models. Regarding the reward model, although our 3B model outperforms mainstream RMs, larger models may further improve performance—albeit at the cost of greater computational overhead.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Jie Cao and Jing Xiao. 2022. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th international conference on computational linguistics*, pages 1511–1520.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

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

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024b. Are we on the right way for evaluating large vision-language models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, and 1 others. 2025. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024a. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024b. Rlhf workflow: From reward modeling to online rlhf. *Transactions on Machine Learning Research*.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and 1 others. 2023. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Wentao Ge, Shunian Chen, Guiming Hardy Chen, Junying Chen, Zhihong Chen, Nuo Chen, Wenya Xie, Shuo Yan, Chenghao Zhu, Ziyue Lin, and 1 others. 2023. Mllm-bench: evaluating multimodal llms with per-sample criteria. *arXiv preprint arXiv:2311.13951*.

Grok-1.5 Team. 2024. Grok-1.5 vision preview.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Eugenio Herrera-Berg, Tomás Vergara Browne, Pablo León-Villagrá, Marc-Lluís Vives, and Cristian Buc Calderon. 2023. Large language models are biased to overestimate profoundness. *arXiv preprint arXiv:2310.14422*.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A Smith, Iz Beltagy, and 1 others. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *arXiv preprint arXiv:2311.10702*.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024a. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2024b. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. 2024a. 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.

Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, and 1 others. 2024b. Vhelm: A holistic evaluation of vision language models. *NeurIPS*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, and 1 others. 2024b. Vl-rewardbench: A challenging benchmark for vision-language generative reward models. *arXiv preprint arXiv:2411.17451*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*.

Runze Liu, Junqi Gao, Jian Zhao, Kaiyan Zhang, Xiu Li, Biqing Qi, Wanli Ouyang, and Bowen Zhou. 2025. Can 1b llm surpass 405b llm? rethinking compute-optimal test-time scaling. *arXiv preprint arXiv:2502.06703*.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2024. Rm-bench: Benchmarking reward models of language models with subtlety and style. *arXiv preprint arXiv:2410.16184*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. 2025. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*.

OpenAI. 2024. Hello GPT-4o.

OpenAI. 2025. Introducing openai o1.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

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

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025. Prmbench: A fine-grained and challenging benchmark for process-level reward models. *arXiv preprint arXiv:2501.03124*.

Ziang Song, Tianle Cai, Jason D Lee, and Weijie J Su. 2023. Reward collapse in aligning large language models. *arXiv preprint arXiv:2305.17608*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. 2023. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *arXiv preprint arXiv:2311.16101*.

Chenglong Wang, Yang Gan, Yifu Huo, Yongyu Mu, Murun Yang, Qiaozhi He, Tong Xiao, Chunliang Zhang, Tongran Liu, Quan Du, and 1 others. 2024a. Rovrm: A robust visual reward model optimized via auxiliary textual preference data. *arXiv preprint arXiv:2408.12109*.

Jun Wang, Meng Fang, Ziyu Wan, Muning Wen, Jiachen Zhu, Anjie Liu, Ziqin Gong, Yan Song, Lei Chen, Lionel M Ni, and 1 others. 2024b. Openr: An open source framework for advanced reasoning with large language models. *arXiv preprint arXiv:2410.09671*.

Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2023. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024c. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, and 1 others. 2025. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2024. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*.

Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. 2025. Multimodal rewardbench: Holistic evaluation of reward models for vision language models. *arXiv preprint arXiv:2502.14191*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 2024. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, and 1 others. 2025. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2025a. Rest-mcts*: Llm self-training via process reward guided tree search. *NeurIPS*.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, and 1 others. 2024a. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024b. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, and 1 others. 2024c. Mavis: Mathematical visual instruction tuning. *arXiv e-prints*, pages arXiv–2407.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025b. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Enyu Zhou, Guodong Zheng, Binghai Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, and 1 others. 2024a. Rmb: Comprehensively benchmarking reward models in llm alignment. *arXiv preprint arXiv:2410.09893*.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024b. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*.

A Experimental Settings for VLLM-as-a-Judge

In this section, we demonstrate the details of how we benchmark VLLMs as reward models on existing vision-language (VL) benchmarks. Following (V)LLM-as-a-judge paradigm, we input the predefined scoring rule, the question, as well as the solution steps for VLLMs to judge the score. We show our prompt below:

> You are a highly capable multimodal AI assistant tasked with evaluating the quality of intermediate reasoning steps provided for visual questions. The input answer may represent an incomplete step in the larger reasoning process. Assign a score from 1 to 5 based on how well the step contributes to addressing the question.
> Question: question
> Answer: answer
> Your score should reflect the overall quality of the answer, focusing on its relevance, coherence, accuracy, and clarity Scoring Scale (1-5):
> 5 (Excellent): The reasoning step is highly relevant, accurate, detailed, and exceptionally clear, making a strong contribution to addressing the question.
> 4 (Good): The reasoning step is relevant, mostly accurate, and clear, with logical progression and only minor flaws.
> 3 (Fair): The reasoning step is somewhat relevant and partially accurate, demonstrating basic logic but lacking detail, clarity, or precision.
> 2 (Poor): The reasoning step is partially relevant but contains major errors, lacks coherence, or is difficult to understand.
> 1 (Very Poor): The reasoning step is irrelevant or nonsensical, showing no meaningful connection to the question or image.
> After your evaluation, please: 1. Assign one overall score from 1 to 5 based on the descriptions above. 2. Explain your reasoning in detail, highlighting specific strengths and weaknesses of the answer.
> Example Response: Reasoning: [Explanation of the evaluation]. Overall Score: [1-5]

B Data Collection Details for ViLReward-73K

B.1 Data Selection

We detail the five datasets that we leverage for ViLReward-73K.

- MAVIS-Geometry (Zhang et al., 2024c) is a mathematical geometry problem dataset that includes 4 different difficulty levels, marked as depth0, depth1, depth2, and depth3. We found that depth0 and depth1 are relatively simple, while depth3, compared to depth2, mostly just increases the number of composite bodies without significantly increasing the difficulty. We chose depth2 as the source for our synthetic data, selecting 5000 examples from it as our question data. MAVIS-Geometry categorizes question types into three classes: text-dominant questions, text-lite questions, and vision-dominant questions. We use vision-dominant questions to enhance the model's visual capabilities. We demonstrate the MCTS tree constructed on MAVIS-Geometry in Figure 8.

- A-OKVQA (Schwenk et al., 2022) mainly contains question-answering problems about natural images, while the majority of these problems are relatively straightforward, requiring only basic visual recognition and common knowledge. In our study, we focused specifically on the more challenging questions within this dataset. We kept 9% of the more difficult ones which were labeled as "difficult direct answer". To generalize to the general visual perception domain, we also sample 500 examples as metadata from the questions that can be directly answered.

- GeoQA170K (Gao et al., 2023) contains over 170K geometric image-caption and question-answer pairs, building on GeoQA+ (Cao and Xiao, 2022) and GeoQA3K (Lu et al., 2021). We sample one question from each unique images from the data, resulting in 8,063 examples in total.

- CLEVR-Math (Lindström and Abraham, 2022) is a synthesized VQA dataset based on CLEVR (Johnson et al., 2017) that includes math word problem solving. They incorporate addition/subtraction types of math problems.

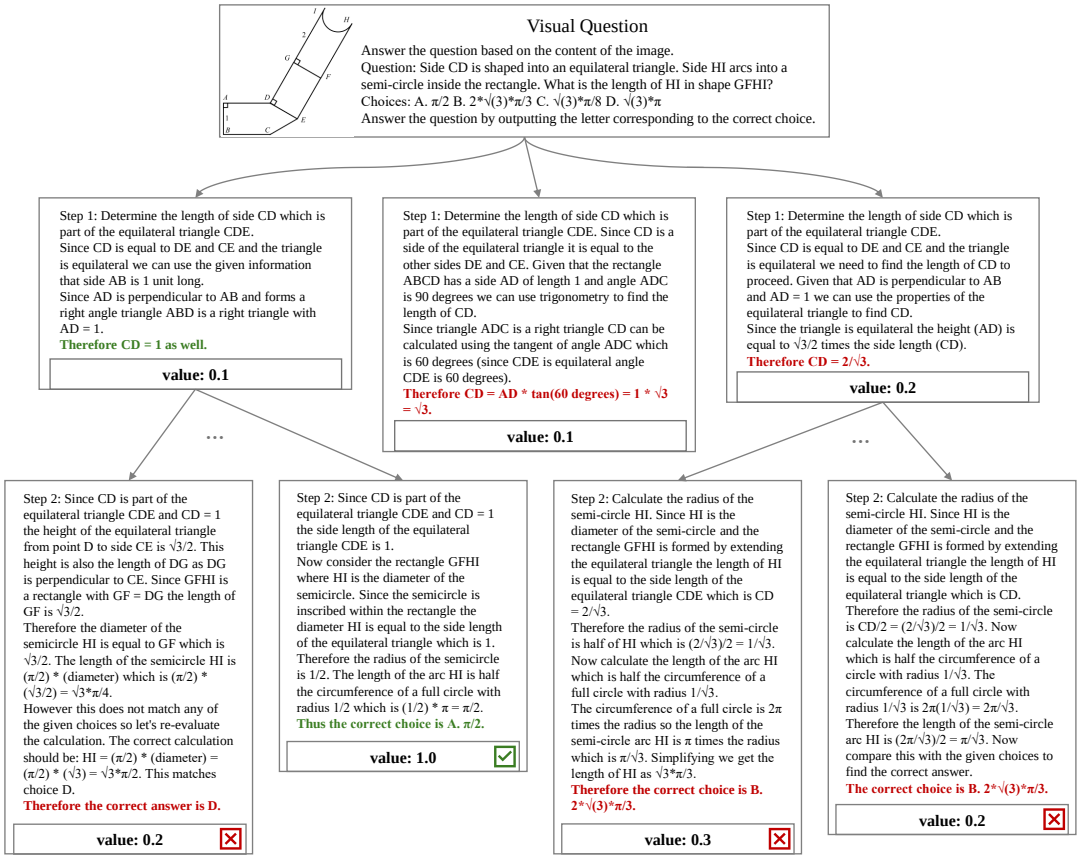

Figure 8: MCTS tree we have constructed for geometry problem datasets (*e.g.*, MAVIS-Geometry). One path in the tree yields a correct result, while the remaining paths result in incorrect answers. It is worth noting that we use ellipses to omit some nodes in the original MCTS tree for better presentation.

We only consider 957 questions with distinct images and need multi-hop reasoning in the dataset.

- ScienceQA (Lu et al., 2022) is a comprehensive datasets with 21K examples in science, the data is categorized into 12 grades based on the difficulty level. We use data that is harder than grade 7 as our metadata.

B.2 MCTS Searching Details

Our construction of the search tree is primarily based on the Monte Carlo Tree Search (MCTS). The process of building this search tree follows several key steps:

Search Tree Initialization. The search process begins with a root node that represents the initial state of the problem. This root node serves as the foundation for the entire search tree, with its parent node set to None.

Node Expansion During Search. During each iteration of the MCTS process, the tree undergoes expansion through four essential phases:

- **Selection Phase.** Starting from the root node, a path is selected based on a specific strategy (we use the Upper Confidence Bound algorithm) until a node that has not been fully expanded is identified.

- **Expansion Phase.** For a node that has not been fully expanded, potential child nodes are generated. Each child node represents a possible subsequent state and is incorporated into the search tree structure with appropriate depth and parent-child relationships.

- **Simulation Phase.** From each newly expanded node, a simulation is conducted using a predetermined strategy (often a random approach) until a terminal state is reached.

- **Backpropagation Phase.** The results obtained from the simulation are propagated backward through all nodes along the path from the root to the expanded node. This pro-

cess updates key node statistics including the visit count and value estimations.

Termination Criteria. The construction of the search tree continues until specific termination conditions are met. In our implementation, we set the iteration limit to 10, meaning the search process concludes after completing 10 iterations. Once this criterion is satisfied, the search process terminates, and the final tree structure is established.

Answer Evaluation. We follow the LLM-as-a-judge approach, using LLM to score the final answers and then propagate these scores from leaf nodes to previous nodes. We first prompt the LLM to extract the final answer from the model's output, then input the question, the model-generated answer, and the correct answer to the LLM to determine whether the final answer is correct.

C ViLBench Evaluation Details

We employ the accuracy between predicted answers and the ground truth as the metric for our ViLBench. To avoid inaccurate extraction of the answer, we follow previous works (Lu et al., 2024; Zhang et al., 2024b) to employ GPT-based extraction. In detail, we prompt GPT-3.5-turbo to compare the prediction with the ground truth, the input instruction shows below:

> Given the following:
> ### Generated Answer: model predicted answer
> ### Ground Truth Answer: ground truth answer
> Please compare the final answer in the generated response to the ground truth answer. Ignore any reasoning or intermediate steps and focus only on whether the final letter answer in the generated response matches the ground truth.
> Output True if the final answer aligns with the ground truth answer; otherwise, output False.

D Vision-Language PRM Training Details

We employ the value head architecture for PRM training. In detail, we train the model on our ViLReward-73K for 2 epochs with a constant learning rate of $2e^{-5}$. We randomly sample 300 instances as the validation set during training and save the model checkpoint with the lowest validation loss.

Model	QwenVL 2.5 3B	InternVL2.5-8B	GPT-4o	GPT-o1	Average
LLaVA-Critic	30.05	28.50	26.83	34.13	29.88
XComposser	32.83	28.50	28.33	36.35	31.50
URSA	32.17	28.17	26.71	35.84	30.72
ViLPRM (Ours)	32.33	30.50	30.33	34.39	**31.89**

Table 9: Model performance under the *Best-of-8* setting on ViLBench.

Model	QwenVL 2.5 3B	InternVL2.5-8B	GPT-4o	GPT-o1	Average
LLaVA-Critic	29.38	28.00	27.33	34.47	29.79
XComposser	33.17	28.17	28.50	38.91	32.19
URSA	32.83	28.33	26.88	36.69	31.18
ViLPRM (Ours)	33.17	30.00	30.87	36.07	**32.53**

Table 10: Model performance under the *Best-of-16* setting on ViLBench.

E Scoring Examples from RMs

In Figure 9, we present another example in the domain of medical reasoning task. As the PRM URSA (Luo et al., 2025) was not trained on the domain of general knowledge, it gives biased judges in this case. In the meanwhile, our ViLPRM is capable of providing more accurate and consistent step rewards in this domain.

F More Results of ViLPRM

To further demonstrate the effectiveness of ViLPRM, we compare it with another two reward models LLaVA-Critic (Xiong et al., 2024) and InternLM-XComposer2.5-Reward-7B (Zang et al., 2025). We randomly draw 200 samples from ViLBench and evaluate under *Best-of-8* and *Best-of-16* settings. Results in Table 9 and Table 10 show that ViLPRM consistently outperforms the larger InternLM-XComposer2.5-Reward (7B) by average margins of 0.39 and 0.34, respectively. LLaVA-Critic, though consuming over 5 times more computational resources as a generative reward model, performs suboptimally due to its coarse-grained, non-stepwise feedback.

G Discussions

Vision-Language PRM is Bounded by Clear Step Segmentation. How to best split the reasoning step for PRMs has always been a problem (Liu et al., 2025; Guo et al., 2025; Cui et al., 2025). In structured tasks like math problems, PRMs provide fine-grained feedback, improving step-by-step reasoning. However, when the segmentation of steps is unclear or reasoning is unnecessary, PRMs may harm the performance. For instance, text-heavy tasks saw a 3% accuracy boost with PRMs, while visual-dominant tasks suffered a 2% drop, likely due to PRMs overemphasizing irrelevant steps.

PRMs also struggle when all steps are treated equally. Previous works have proposed to use single step to represent all step rewards (Wang et al., 2024b; Liu et al., 2025). We found that rewarding only the last few critical steps improved accuracy more than using all steps, striking a balance between PRMs and ORMs. A major challenge is identifying which steps truly matter. Future improvements should focus on adaptive step evaluation, where PRMs automatically adjust reward weight based on step importance. Better segmentation strategies, such as enforcing clearer step structures during training or integrating step selection mechanisms can help PRMs generalize better across tasks.

Improved Training Paradigm is Required for Multimodal RMs. Current training approaches for multimodal reward models fail to generalize across diverse tasks. Many RMs, including PRMs, are task-sensitive (Zhang et al., 2025b; Liu et al., 2025), meaning they work well on specific domains but struggle elsewhere. For example, PRMs trained on math tasks such as URSA perform poorly on vision-heavy reasoning, suggesting that current methods do not equip RMs with broad evaluation skills. Besides, our results show that even advanced VLLMs like GPT-4o do not automatically become good reward models, often overrating responses.

To improve vision-language PRMs, training must diversify data sources, integrating both textual and visual-heavy reasoning tasks. Instead of relying solely on step-wise learning, future RMs should also consider incorporating adaptive reward mechanisms, adjusting considered step scores based on task complexity. Additionally, evaluation benchmarks for reward models should also go beyond accuracy, assessing consistency, bias, and generalization (Yasunaga et al., 2025).

H Declaration of AI Tool Usage

During the preparation of this manuscript, I used OpenAI's GPT-4o model for minor language refinement and smoothing of the writing. The AI tool was not used for generating original content, conducting data analysis, or formulating core scientific ideas. All conceptual development, experimentation, and interpretation were conducted independently without reliance on AI tools.

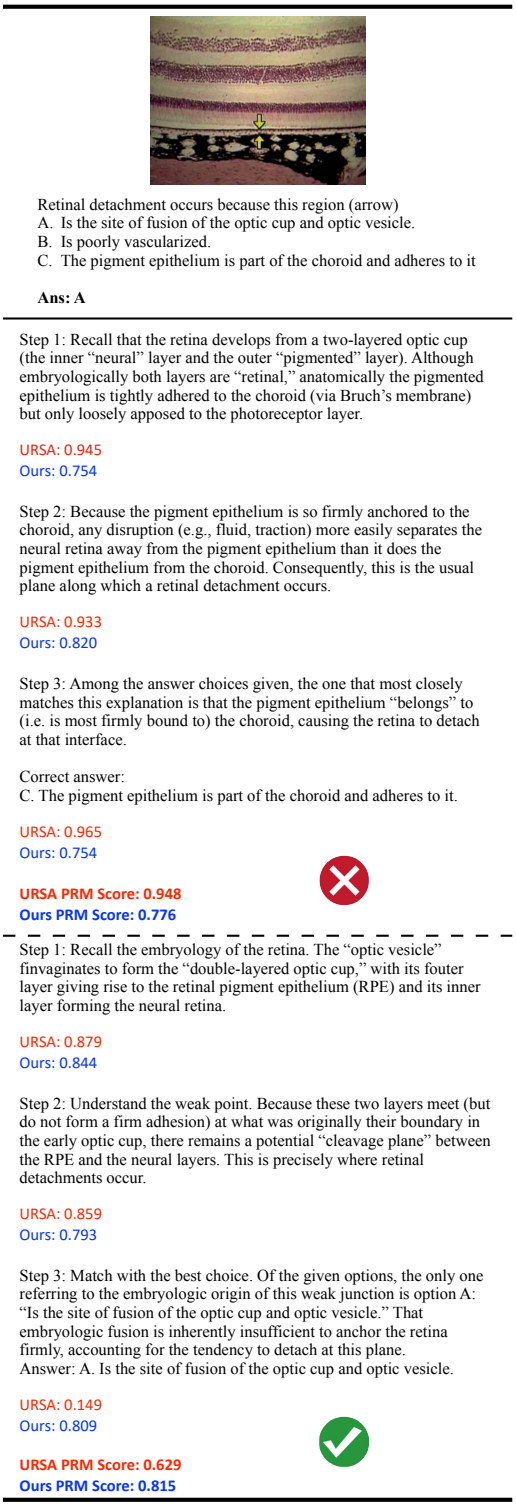

Figure 9: An example from o1's generation of a medical reasoning example from our VILBENCH.