P1-VL: Bridging Visual Perception and Scientific Reasoning in Physics Olympiads

Yun Luo [1 * ✉] **Futing Wang** [1 *] **Qianjia Cheng** [1 *] **Fangchen Yu** [1] **Haodi Lei** [1] **Jianhao Yan** [1] **Chenxi Li** [1] **Jiacheng Chen** [2] **Yufeng Zhao** [1] **Haiyuan Wan** [1] **Yuchen Zhang** [1] **Shenghe Zheng** [1] **Junchi Yao** [1] **Qingyang Zhang** [1] **Haonan He** [1] **Wenxuan Zeng** [1] **Li Sheng** [1] **Chengxing Xie** [1] **Yuxin Zuo** [1] **Yizhuo Li** [1] **Yulun Wu** [1] **Rui Huang** [1] **Dongzhan Zhou** [1] **Kai Chen** [1] **Yu Qiao** [1] **Lei Bai** [1 ✉] **Yu Cheng** [2 1 ✉] **Ning Ding** [3 1 ✉] **Bowen Zhou** [1 3 ✉] **Peng Ye** [1 ✉] **Ganqu Cui** [1 ✉]

Abstract

The transition from symbolic manipulation to science-grade reasoning represents a pivotal frontier for Large Language Models (LLMs), with physics serving as the critical test anchor for binding abstract logic to physical reality. Physics demands that a model maintain physical consistency with the laws governing the universe, fundamentally requiring multimodal perception to ground abstract logic in reality. To bridge this visual-logical gap, we introduce P1-VL, a family of open-source vision-language models engineered for advanced scientific reasoning. Our method stabilizes post-training through *Curriculum Reinforcement Learning*, where task difficulty is progressively expanded, and further enhances inference with *Agentic Augmentation* for iterative self-verification. Evaluated on HiPhO, a rigorous benchmark of 13 physics olympiads exams from 2024–2025, our flagship `P1-VL-235B-A22B` becomes the first open-source Vision-Language Model (VLM) to secure 12 gold medals and achieves the state-of-the-art performance in the open-source models. Our agent-augmented system achieves the No.2 overall rank globally, trailing only Gemini-3-Pro. Beyond physics, P1-VL demonstrates remarkable scientific reasoning capacity and generalizability, establishing significant leads over base models in STEM benchmarks.

[1]Shanghai AI Laboratory, Shanghai, China [2]The Chinese University of Hong Kong, Hong Kong, China [3]Tsinghua University, Beijing, China. Correspondence to: Yun Luo <ylramirocs@gmail.com>, Ganqu Cui <chengyu@cse.cuhk.edu.hk>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Recent advances in Large Language Models (LLMs) (Yang et al., 2025; Comanici et al., 2025) have propelled artificial intelligence beyond symbolic manipulation toward science-grade reasoning (Gibney, 2025; Zhang et al., 2024; Bai et al., 2025a). Among scientific disciplines, **physics** stands as the critical test anchor for these emerging capabilities. Physics inherently demands a rigorous synthesis of multimodal perception and causal logic to model the behavior of the universe. Consequently, the ability to solve complex physics problems, a task that often requires mapping physical intuition to abstract laws, provides a rigorous test for genuine, first-principles reasoning.

Olympiad-level competitions, such as the International Physics Olympiad (IPhO) serve as a significant surrogate to evaluate the physics reasoning capacity. While recent works have begun to adapt LLMs for these challenges (Qiu et al., 2025; Chen et al., 2025b), they predominantly focus on textual reasoning, overlooking its inherent multimodality. In many Olympiad problems, diagrams contain essential geometric constraints, circuit topologies, and force interactions that are purposefully omitted from the text. Without the ability to perceive and interpret these visual schematics, a text-only model faces an insurmountable information gap. As illustrated in Figure 1, the task involves quantifying bubble radii and ascent velocities directly from visual inputs, effectively simulating the data analysis process in real-world physics experiments. Thus, Physics Olympiads serve as high-fidelity testbeds for Vision-Language Models (VLMs), verifying the ability to align visual observations with abstract physical laws (Yu et al., 2025a).

In this work, we introduce **P1-VL**, a new family of open-source vision-language models designed to advance the frontier of scientific AI. Our framework harmonizes two critical paradigms: *Curriculum Reinforcement Learning (RL) Training* and *Agentic Augmentation*. This dual approach ensures that the model not only internalizes robust reasoning capabilities through RL (Sutton et al., 1998), but also effectively

IPhO 2025 Q3 - Part A:

Champagne is a French sparkling wine. Fermentation of sugars produces carbon dioxide (CO_2) in the bottle. The molar concentration of CO_2 in the liquid phase c_ℓ and the partial pressure P_{CO_2} in the gas phase are related by $c_\ell = k_H P_{CO_2}$ known as Henry's law and where k_H is called Henry's constant.

- Surface tension of champagne $\sigma = 47\times10^{-3} J \cdot m^{-2}$,
- Density of the liquid $\rho_\ell = 1.0\times10^{3} kg \cdot m^{-3}$,

Fig. 1. A glass filled with champagne.

.....The bubbles of the train have the same initial radius and are emitted at a constant frequency fb = 20 Hz. For the range of velocities studied here, the drag force F on a bubble of radius a moving at velocity v in a liquid of dynamic viscosity η is given by Stokes' law $F = 6\pi\eta a v$ .....

Question: Give the expression of the main forces exerted on a vertically rising bubble. Obtain the expression of $v(a)$. Give a numerical estimate of η using ρ_ℓ, g_0 and quantities measured on **Fig. 2**.

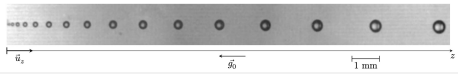

Fig. 2. A train of bubbles.

Noted: Requires measuring the **radius a of the bubbles** in the photograph (Fig. 2), and measuring the **ascent velocity v of the bubbles**, i.e. the vertical distance z between bubble n and bubble n-2 on the photograph to derive the velocity.

Figure 1. A question sample from the International Physics Olympiad 2025 (IPhO 2025), where the question requires measuring the radius of bubbles and estimating their velocity in Fig 2.

operationalizes them via adaptive agentic control during inference (Feng et al., 2026).

- **Curriculum RL Training.** P1-VL is developed exclusively through RL post-training (Guo et al., 2025; Cui et al., 2025a) built upon base vision-language models. To maximize the efficacy of this paradigm, we introduce a curriculum RL framework that systematically advances reasoning capabilities through progressive difficulty expansion, and robust stabilization mechanisms.
- **Agentic Augmentation.** During the inference phase, we integrate P1-VL models with the *PhysicsMinions* agent framework (Yu et al., 2025b) to enable iterative correction and self-verification. This mechanism facilitates multi-turn reflection, empowering the model to reason, critique, and refine its solutions—mirroring the rigorous cognitive process of human physicists.

We release two variants of the P1-VL family and evaluate them on HiPhO (Yu et al., 2025a), a rigorous benchmark aggregating 13 recent Olympiad exams from 2024–2025. Our flagship model, `P1-VL-235B-A22B`, marks a significant milestone for the open-source community: it is the first specialized physics vision-language model to secure **12 golds and 1 silver** across the full HiPhO suite. Remarkably, the standalone P1-VL surpasses even the agent-enhanced text baseline (`P1-235B-A22B+PhysicsMinions` (Chen et al., 2025b)), ranking **3rd** among all open and closed-source models. Meanwhile, our lightweight variant, `P1-VL-30B-A3B`, achieves 9 gold medals, surpassing the average score of most open-source models. When further augmented with the *PhysicsMinions* agent framework, P1-VL ascends to the **2nd overall spot** on the leaderboard.

Beyond its specialized domain, P1-VL demonstrates remarkable generalizability. In the multidisciplinary FrontierScience-Olympiad, both `P1-VL-235B-A22B` and `P1-VL-30B-A3B` models demonstrate robust gains across biology, chemistry, and physics, outperforming their respective baselines by margins of 8.0 and 9.1 points, respectively. In the mathematical reasoning tasks and multi-modal STEM reasoning tasks, our model could also consistently outperform their base models.

This work makes the following key contributions:

- We introduce **P1-VL**, a first family of open-source Vision-Language Model specialized in physics problems, achieving 12 gold medals in physics Olympiad competitions.
- We develop a **curriculum RL framework** with difficulty expansion, and training stabilization for sustained reasoning improvement.
- We open-source the P1-VL model family, filling the multimodal void within the P1 ecosystem (Chen et al., 2025b), and bridging the gap between visual perception and scientific reasoning.

Collectively, successfully bridging this visual-logical gap is more than an academic achievement; it lays the foundation for **general physical intelligence**. Mastering the rigorous constraints of Olympiad problems is a necessary precursor to achieving *machine scientific discovery* (Zheng et al., 2025c) and reliable *embodied AI*. Therefore, Olympiad-level reasoning is not an end, but a strategic step toward systems capable of understanding and eventually navigating the physical reality.

2. Physics Dataset

We introduce a systematically curated multimodal dataset of 8,033 problems designed to advance genuine scientific reasoning in VLMs. The dataset includes problems from physics Olympiads (4,126), undergraduate textbooks (2,968), and competition guides (939) (Table 1).

2.1. Dataset Construction

Data Collection. Our data construction is driven by a core objective: enabling VLMs to internalize structured reasoning aligned with physical laws and empirical consistency—a

prerequisite for genuine scientific intelligence. To achieve this, we prioritize problems that combine conceptual depth with rule-verifiable outcomes, traits shown to catalyze reasoning improvements (Guo et al., 2025; Yang et al., 2025; Wen et al., 2025; Chen et al., 2025b). Within the physics domain, physics Olympiad problems naturally contain such properties: they demand rigorous modeling and multi-step inference while remaining formally tractable. The value of this specific difficulty profile is underscored by empirical findings in PHYSICS (Zheng et al., 2025b), which demonstrate that LLMs face greater challenges in Olympiad tasks than in introductory undergraduate problems. Accordingly, we first curate our dataset from two complementary pillars: ten prestigious physics Olympiads (up to 2023, including IPhO and APhO) that capture a graded difficulty spectrum, and ten authoritative competition textbooks that provide systematically organized exercises with expert-verified solutions. To further augment the dataset's multi-modal richness, we extend our collection to include undergraduate textbooks. These sources serve as a vital repository of image-centric problems, significantly enhancing the overall density of visual schematics and diagrams within the corpus.

Table 1. Statistics of the multi-modal training data.

Statistics	**Number**
Data Composition	
Total Problems	8,033
- Problems from Olympiads	4,126 (51%)
- Problems from textbooks	3,907 (49%)
Total Answers	10,516
Total Fields	5
Total Subfields	25
Total Answer Types	5
Data Sources	
Textbooks	20
Olympiad Types	9
- Sets Collected	199
Token Statistics	
Average Question Tokens	943
Max Question Tokens	6,604
Max Solution Tokens	5,519
Images Statistics	
Questions with Images	5,513
Questions without Images	2,520
Max Images per Question	8
Average Images per Question	0.82

2.2. Quality Control.

To ensure data reliability, we implement a rigorous multi-stage pipeline integrating automated model-based filtering with human verification. **(1) OCR Correction.** Texts parsed from complex layouts or low-quality scans are manually validated against original PDFs to rectify OCR artifacts. **(2) Answer Cross-Validation.** Three models — Gemini-2.5-Flash, Claude-3.7-Sonnet, and GPT-4o — independently extract answers from each question–solution pair. Consensus is established via majority vote (at least two models agree); non-consensus items are discarded. **(3) Task Filtering.** Problems requiring diagram generation or involving unverifiable open-ended answers (e.g., proofs) are removed. **(4) Visual Consistency Check.** We utilize Gemini-2.5-Flash to detect and filter out samples where the question text references a figure that is missing from the input. **(5) Expert Review.** A final consistency check is performed by Claude-3.7-Sonnet, followed by targeted manual refinement by experts. Through this filtering process, the dataset is refined from 13,432 to 8,033 items, yielding a high-fidelity, multi-modal bilingual corpus well-suited for RLVR.

3. Approach

3.1. RL Formulation

We formulate the challenge of solving Physics Olympiad problems through the lens of reinforcement learning (RL) (Sutton et al., 1998) process. Formally, let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$ denote the underlying Markov Decision Process (MDP), where:

- $\mathcal{S}$ represents the state space, corresponding to the model context, including the problem statement and all previously generated reasoning tokens.
- $\mathcal{A}$ is the action space, which constitutes the discrete action space of possible tokens in the vocabulary.
- $P(s' \mid s, a)$ defines a (deterministic) state transition function, which appends the newly chosen action a to the state s, resulting in an updated context s'.
- $r(s, a)$ provides a scalar reward indicating the correctness and quality of the finalized reasoning trajectory.

The learning objective is to optimize the policy π_θ by maximizing the expected return:

$$J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} r(s_t, a_t)\right], \tag{1}$$

where $\pi_\theta(a_t|s_t)$ is the policy parameterized by model parameters θ, and $\tau = (s_0, a_0, \ldots, s_T)$ denotes a trajectory sampled from π_θ.

Policy Gradient. The policy gradient (Sutton et al., 1999) method optimizes π_θ by ascending the gradient of the expected return:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\, A^\pi(s_t, a_t)\right], \tag{2}$$

where $A^\pi(s_t, a_t)$ is the advantage function estimating the relative value of action a_t in state s_t. We adopt this standard form and instantiate it via GSPO below.

Group Sequence Policy Optimization (GSPO). GSPO (Zheng et al., 2025a) elevates optimization from the token level (Shao et al., 2024; Yu et al., 2025c) to the sequence level, employing length-normalized sequence likelihood importance ratios:

$$\rho_i(\theta) = \exp\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\log\frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})}\right), \quad (3)$$

where $|y_i|$ denotes the sequence length, and the $1/|y_i|$ term implements length normalization to reduce variance. The advantage function is computed at the sequence level:

$$\hat{A}_i^{\text{GSPO}} = \frac{R_i - \text{mean}(\{R_j\}_{i=1}^G)}{\text{std}(\{R_j\}_{i=1}^G)}, \quad (4)$$

with the objective function:

$$J^{\text{GSPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(\rho_i(\theta)\hat{A}_i^{\text{GSPO}}, \text{clip}(\rho_i(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i^{\text{GSPO}}\right)\right]. \quad (5)$$

3.2. Instantiation

Reward Design. We employ a binary reward scheme based on answer correctness, leveraging the fact that our physics dataset contains verifiable ground-truth outcomes:

$$r = \begin{cases} 1, & \text{if the predicted answer matches the ground truth,} \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

We adopt the test-case-style reward aggregation, defining the final reward as:

$$R = \frac{1}{N}\sum_{i=1}^{N} r_i, \quad (7)$$

where N is the number of required sub-answers in the problem, and r_i denotes the correctness indicator for the i-th sub-answer. Furthermore, we apply the system prompt as shown in Appendix A.1 to extract the multi-box answers.

Verifier Design. Physics problems frequently necessitate solutions in the form of complex symbolic expressions rather than scalar values, posing a challenge for standard verification metrics. We implement the hybrid verification framework that harmonizes rigorous symbolic logic with semantic model evaluation: (1) Rule-Based Verifier, we integrate symbolic computation libraries (SymPy (Meurer et al., 2017)) with heuristic strategies from math-verify (Kydlíček); (2) To capture correct answers that elude strict symbolic matching, we adopt the XVerify (Chen et al., 2025a) paradigm. We employ a specialized LLM (Qwen3-30B-A3B-Instruct) as a semantic judge.

3.3. Technical Design

3.3.1. Curriculum Training

Preliminary Difficulty Calculation and Filtering Strategy. The distribution of data difficulty and quality plays a pivotal role in the convergence of RL training (Zhang et al., 2025a; Cui et al., 2025a). To construct an effective training set, we adopt an inference-based metric to estimate the solvability. Let $D(x_i, y_i)$ denote the empirical pass rate:

$$D(x_i, y_i) = \frac{1}{N}\sum_{k=1}^{N}\mathbb{I}(\hat{y}_k = y_i), \quad (8)$$

where $\hat{y}_k$ denotes the k-th response generated by the model. We utilize the Qwen3-VL-30B-A3B model to perform extensive rollouts (with $N = 72$) on the training dataset. The correctness of inference rollouts is determined only by the rule-based verifier. Based on the distribution of $D(x_i, y_i)$, we apply a dual-end filtering strategy:

1. **Pruning Trivial Samples ($D(x_i, y_i) > 0.7$):** Samples that the model can already solve with high confidence contribute minimal gradient variance during optimization, accelerating entropy collapse of the policy model (Cui et al., 2025b). To improve training efficiency, we categorize samples with $D > 0.7$ as "trivial" and remove them from the training set following (Chen et al., 2025b).

2. **Recovering Zero-Shot Failures ($D(x_i, y_i) = 0.0$):** Samples where the pass rate is strictly 0.0 present a challenge; they represent either hard-core reasoning steps or malformed questions. Training on the malformed questions can lead to reward hacking or instability. Thus, we filter these malformed samples and introduce a *recovery pipeline* using Gemini-2.5-Flash. We prompt such a stronger model to (a) verify the alignment between the question and image, and (b) verify the completeness of the question, including physical factors or numerical values, (c) refine ambiguous problem descriptions. The unrecoverable tasks are finally excluded to maintain the overall quality. This process converts previously "unsolvable" noise into effective, clear RL tasks.

Curriculum RL Training. Directly training with RL on difficult problems for the VLMs could cause unstable policy gradients and low efficiency because of the low success rates, thus, we expand the difficulty of the tasks step by step. In the initial stage, we train the model with samples that exhibit a pass rate below an upper threshold of 0.7. By filtering out trivial instances (i.e., those with success rates $D(x_i, y_i) > 0.7$) early in the process, we prevent the model from wasting computation on tasks it has already mastered. Subsequently, we further intensify the training challenge by lowering this inclusion threshold. This progressive tightening of the difficulty constraint effectively prunes moderately

easy samples from the dataset, compelling the model to concentrate its optimization capacity exclusively on the harder tail of the distribution, where its performance is still lacking.

As the curriculum shifts toward the harder tail of the distribution, static rollout configurations become restrictive, failing to provide the search depth required for these escalated challenges (Cui et al., 2025b; Luo et al., 2025a). Thus, we follow P1(Chen et al., 2025b) to dynamically expand the exploration space, such as group size and generation window, in pace with the difficulty expansion, ensuring the search budget scales to sustain long-term learnability.

3.3.2. Training Stabilization Mechanism

RL frameworks (Zhu et al., 2025b; Sheng et al., 2024) often adopt different engines for rollout (e.g. vllm (Kwon et al., 2023) and SGLang (Zheng et al., 2024)) and training (e.g. FSDP (Merry et al., 2021) and Megatron (Shoeybi et al., 2019)), which introduce a mismatch between $\pi_\theta^{rollout}$ and π_θ^{train} due to differences in numerical precision, and computation optimization strategies, leading to biased gradient estimates and instability (Yao et al., 2025; Liu et al., 2025).

Masked Importance Sampling. To mitigate the instability caused by the training-inference mismatch, we adopt Sequence-level Masked Importance Sampling (Seq-MIS) (Liu et al., 2025). Unlike token-level clipping methods, Seq-MIS employs a "hard trust region" by rejecting entire out-of-distribution trajectories. We adopt the Geometric Mean of the importance weights to normalize the metric across different lengths. The gradient estimator is formulated as:

$$\nabla_\theta J_{\text{Geo-MIS}}(\theta) = \mathbb{E}_{\tau\sim\pi_{\text{old}}}\Bigg[\underbrace{\mathbb{I}\left(\rho(\tau)^{1/T} \le C\right)}_{\text{Geo-Mask}} \cdot \rho(\tau) \cdot \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) A^\pi(s_t, a_t)\Bigg]. \tag{9}$$

where $\tau = (s_0, a_0, \ldots, s_T, a_T)$ represents the sampled trajectory, and T is the sequence length. The sequence-level importance weight is defined as $\rho(\tau) = \prod_{t=0}^{T} \frac{\pi_\theta^{\text{trainer}}(a_t|s_t)}{\pi_\theta^{\text{rollout}}(a_t|s_t)}$. $\mathbb{I}(\cdot)$ denotes the indicator function, rejecting trajectories whose geometric mean of importance weights exceeds the threshold C.

3.4. Implementation

For training P1-VL, we adopt the VERL (Sheng et al., 2024) framework, which is an efficient VLMs post-training framework offering the implementation in Megatron. We adopt `Qwen3-VL-30B-A3B-Thinking` and `Qwen3-VL-235B-A22B-Thinking` as backbone VLMs. The setting and analysis of curriculum training are shown in Appendix A.2 and A.6. For questions containing images, the special token `<image>` is inserted at the corresponding position within the text. To facilitate unified RL training across modalities, we standardize the input format by padding all text-only samples with a blank image of resolution 448×448, and appending the `<image>` token to the end of the question. Since our training data is not specifically curated for image alignment, we freeze the parameters of both the vision encoder and the projection layer during RL training. The analysis of data composition is shown in Appendix A.5.

3.5. Agentic Augmentation

At inference time, we enhance test-time reasoning with a unified agentic augmentation framework, `PhysicsMinions` (Yu et al., 2025b), a pioneering coevolutionary multi-agent system for multimodal scientific problem solving. The system is organized into three tightly coupled studios: the *Visual Studio*, the *Logic Studio*, and the *Review Studio*. The Visual Studio performs structured perception and converts raw visual evidence into symbolic representations. The Logic Studio generates and iteratively refines candidate solutions through solver and introspector collaboration. The Review Studio validates solutions using both domain-specific and general verification. Together, these studios form a closed critique and refinement loop that enables scalable test-time reasoning and improves robustness on complex scientific tasks.

4. Experiment

4.1. Experimental Setup

Test Dataset. To rigorously evaluate performance on advanced physical reasoning tasks, we introduce the HiPhO benchmark (Yu et al., 2025a), the first dataset dedicated exclusively to High School Physics Olympiads. HiPhO aggregates 13 major exams administered between 2024 and 2025, spanning seven prestigious international and regional competition series: IPhO, APhO, EuPhO, NBPhO, PanPhO, PanMechanics, and F=MA. The inclusion criteria prioritized both global influence and the availability of granular human performance data to facilitate direct human-AI comparison[1].

Model Configurations. We compared against 39 representative models, including 13 closed-source and 26 open-source models, selected to reflect the current frontier of LLMs and VLMs (Appendix A.3). All models were evaluated under a standardized inference setup following the HiPhO protocol. The temperature was fixed at 0.6, and the maximum token limit was set as large as permitted by each

[1]Notable competitions such as CPhO, USAPhO, and APhO-2024 were excluded due to the unavailability of complete official contestant score distributions.

Table 2. Evaluation results on the HiPhO benchmark (13 physics Olympiads from 2024–2025) using the *exam score* metric. Gold, Silver and Bronze indicate scores above the respective thresholds. Models are ranked by average scores; **bold** is the highest score, and underline is the second highest. Here, only the theoretical parts of exams are used, hence Full Mark (Model) $\leq$ Full Mark (Human).

Physics Olympiad	IPhO		APhO	EuPhO		NBPhO		PanPhO		PanMechanics		F=MA		Avg.	Medal Table		
Year	2025	2024	2025	2025	2024	2025	2024	2025	2024	2025	2024	2025	2024				
Full Mark (Human)	30.0	30.0	30.0	30.0	30.0	72.0	72.0	100.0	100.0	100.0	100.0	25.0	25.0	57.2			
Full Mark (Model)	29.4	29.3	30.0	29.0	28.0	43.5	50.0	100.0	98.0	100.0	100.0	25.0	25.0	52.9			
Top-1 Score (Human)	29.2	29.4	30.0	27.0	30.0	53.2	40.8	81.0	66.5	62.0	51.0	25.0	24.0	42.2			
Top-1 Score (Compared Models)	25.2	25.9	28.4	21.0	23.9	37.1	43.8	66.3	82.5	86.0	82.3	23.8	23.9	43.9			
Gold Medal	19.7	20.8	23.3	16.5	20.4	28.6	26.5	41.5	52.0	52.0	51.0	15.0	14.0	29.3			
Silver Medal	12.1	11.1	18.7	9.8	14.2	20.1	19.4	28.5	37.5	36.0	26.0	11.0	12.0	19.7			
Bronze Medal	7.2	3.6	13.1	5.8	8.9	15.2	13.5	14.5	16.0	20.0	12.0	9.0	10.0	11.4			
Gemini-3-Pro (high)	**25.2**	25.2	28.3	**21.0**	**23.9**	**37.1**	**43.8**	66.3	76.2	**86.0**	74.8	**23.8**	**23.9**	**42.7**	13	0	0
P1-VL-235B-A22B+PhysicsMinions	22.1	25.8	27.1	13.7	22.9	31.8	32.6	**66.5**	**83.3**	76.0	**84.8**	23.1	21.8	40.9	12	1	0
GPT-5.2 (high)	22.3	25.5	**28.4**	15.7	**23.9**	35.8	37.1	59.8	82.5	67.6	82.3	23.4	23.5	40.6	12	1	0
P1-VL-235B-A22B	21.0	25.3	26.9	10.7	21.5	32.5	31.9	61.7	79.8	73.3	82.0	22.5	21.9	39.3	12	1	0
P1-235B-A22B+PhysicsMinions	23.2	25.2	28.0	12.4	23.5	31.9	35.4	57.7	67.0	77.5	74.8	21.5	20.5	38.4	12	1	0
Gemini-2.5-Pro	22.7	**25.9**	27.9	14.9	21.8	32.3	35.9	60.3	64.1	69.5	70.2	22.8	22.0	37.7	12	1	0
GPT-5	22.3	20.2	27.0	10.3	21.7	32.9	32.8	55.9	69.8	69.4	79.0	22.4	22.4	37.4	11	2	0
DeepSeek-V3.2-Thinking	20.2	24.9	26.1	6.7	18.9	33.3	29.7	57.3	64.9	78.6	78.0	21.5	19.9	36.9	11	1	1
P1-235B-A22B	21.2	24.7	27.4	10.8	23.0	31.8	28.4	54.7	56.7	74.7	72.9	20.9	19.4	35.9	12	1	0
Grok-4	18.7	23.5	25.0	11.5	20.5	25.8	29.3	45.0	75.4	72.1	78.6	19.8	19.8	35.8	10	3	0
o3	15.7	23.7	25.9	11.4	21.6	34.1	33.5	47.3	55.9	71.4	75.6	22.0	20.6	35.3	11	2	0
P1-VL-30B-A3B	17.9	22.5	24.6	10.6	19.0	30.2	24.6	50.8	58.6	79.1	75.5	20.8	20.3	35.0	9	4	0
Qwen3-VL-235B-A22B-Thinking	19.8	23.4	25.4	10.8	19.0	29.2	25.9	48.3	54.0	64.3	79.9	21.4	19.9	33.9	10	3	0
Qwen3-235B-A22B-Thinking-2507	17.1	23.0	26.2	10.9	20.4	33.6	28.1	44.7	51.8	69.1	72.9	18.5	18.9	33.5	10	3	0
P1-30B-A3B	18.5	22.3	25.4	7.4	18.8	29.2	24.0	47.0	51.4	69.5	69.1	19.6	20.0	32.5	8	4	1
GPT-OSS-120B (high)	20.5	22.1	25.8	12.8	23.0	32.4	32.4	48.8	52.4	54.7	56.1	20.5	19.2	32.4	12	1	0
o4-mini (high)	16.0	23.7	22.9	12.0	20.1	27.4	29.8	41.4	50.9	69.1	67.3	18.6	18.8	32.2	6	7	0
Gemini-2.5-Flash-Thinking	20.2	23.9	27.4	13.2	21.9	29.0	29.3	44.6	54.9	60.5	55.9	17.8	19.1	32.1	12	1	0
DeepSeek-R1	18.5	24.6	25.4	10.8	21.4	26.3	20.5	42.2	47.4	65.4	72.5	18.3	18.5	31.7	8	5	0
o4-mini	15.4	22.9	22.8	10.1	20.9	26.9	27.3	39.4	47.1	64.2	62.5	18.6	18.5	30.5	7	6	0
Kimi-K2-Thinking*	19.1	22.8	22.4	5.2	20.9	26.0	20.6	52.5	51.3	47.3	64.8	20.6	19.7	30.2	6	6	0
Claude-4-Sonnet-Thinking	19.0	22.0	24.8	9.7	20.5	28.1	25.6	43.1	39.3	57.4	61.8	19.2	20.1	30.0	8	4	1
Qwen3-30B-A3B-Thinking-2507	15.6	19.7	23.5	7.4	16.4	28.6	22.2	40.5	43.8	67.7	66.5	18.3	18.0	29.9	6	6	1
Qwen3-VL-30B-A3B-Thinking	15.7	19.3	25.3	11.7	20.2	29.7	28.3	41.6	45.9	57.6	51.4	18.9	20.7	29.7	8	5	0
Qwen3-32B	15.7	19.3	23.9	9.8	21.2	28.9	24.1	36.6	41.8	67.0	59.2	18.9	16.6	29.5	7	6	0
Kimi-K2-Instruct	16.5	19.8	24.2	11.0	16.9	26.5	26.2	35.9	41.8	65.9	58.9	16.0	18.2	29.1	5	8	0
GPT-OSS-120B	16.9	21.4	22.8	9.1	19.9	26.0	25.8	37.4	41.8	57.1	59.7	17.8	17.6	28.7	5	7	1
Intern-S1	15.9	14.2	21.7	9.0	16.6	23.0	20.5	41.1	50.3	60.4	57.4	18.4	19.5	28.3	4	8	1
Claude-4-Sonnet	15.7	19.2	22.8	9.5	16.5	27.5	21.3	40.4	43.3	46.5	48.5	16.8	16.5	26.5	2	10	1
DeepSeek-V3	13.6	16.4	22.1	7.1	17.2	21.1	17.3	37.2	35.0	48.4	46.5	14.1	15.6	24.0	1	9	3
Mistral-Medium-3	14.2	14.1	19.9	8.5	12.2	20.4	19.6	30.8	28.6	32.9	36.1	13.9	14.1	20.4	1	8	4
InternVL3-78B-Instruct	12.9	12.5	17.7	7.5	15.2	22.3	22.5	26.2	27.4	21.1	27.1	12.0	13.0	18.3	0	8	5
GLM-4.5V	11.9	4.4	16.2	8.7	14.1	19.5	14.0	18.5	16.0	47.8	39.0	13.0	13.8	18.2	0	4	9
LLaMA4-Scout-17B	9.7	9.5	13.1	5.4	10.4	22.8	12.8	26.6	24.1	35.4	34.5	8.1	6.4	16.8	0	2	7
GPT-4o	10.2	9.4	15.1	6.8	9.2	16.4	11.7	27.8	22.8	28.2	26.5	15.0	10.9	16.2	1	1	10
Qwen3-8B	10.6	12.7	11.5	7.1	11.9	20.1	17.3	26.3	22.3	21.8	22.8	10.8	10.0	15.8	0	2	10
Qwen2.5-VL-32B-Instruct	9.9	8.2	16.5	6.9	10.0	15.3	14.4	22.5	22.4	28.1	29.9	7.6	4.6	15.1	0	1	10
Qwen2.5-VL-72B-Instruct	10.6	7.2	13.6	6.1	8.1	13.3	11.4	26.8	18.2	24.0	28.5	13.5	9.8	14.7	0	2	7
InternVL3-38B-Instruct	8.9	7.8	12.3	6.1	8.3	14.0	10.6	24.1	20.4	27.5	24.8	8.2	6.8	13.8	0	0	7
InternVL3-9B-Instruct	4.7	3.7	7.2	4.2	4.1	9.4	6.0	12.4	11.0	11.6	16.3	6.4	6.2	7.9	0	0	2
Qwen2.5-VL-7B-Instruct	3.5	2.5	5.7	4.4	3.6	7.3	5.5	14.7	7.6	9.8	11.5	4.4	3.5	6.5	0	0	1
Phi-4-multimodal	2.0	1.6	4.2	3.6	3.6	5.0	4.5	8.3	9.0	10.0	10.1	4.4	5.0	5.5	0	0	0
DeepSeek-VL2	1.8	0.5	2.5	3.4	3.4	5.0	3.2	5.6	4.8	7.3	6.4	5.0	3.9	4.0	0	0	0

* For `Kimi-K2-Thinking`, the inference timeout is set to 2 hours, and 9.58% of cases exceed this limit.

Table 3. Evaluation results on the FrontierScience-Olympiad benchmark. The state-of-the-art performance of all the models is marked in bold, and that of open-source models is underlined. * refers to the evaluation results reported in (Wang et al., 2026).

Model	Biology/10	Chemistry/40	Physics/50	Total/100
GPT-5.2*	**43.5**	**89.0**	74.3	**77.1**
Gemini-3-Pro*	41.0	85.6	**75.5**	76.1
Claude-Opus-4.5*	24.0	81.8	72.5	71.4
GPT-5*	35.5	81.5	67.0	69.7
P1-VL-235B-A22B+PhysicsMinions	26.3	77.2	67.3	67.1
Grok-4*	33.0	73.2	67.2	66.2
DeepSeek-V3.2-Thinking	26.3	74.1	67.3	65.9
P1-235B-A22B+PhysicsMinions	30.0	71.0	68.0	65.4
Kimi-K2-Thinking	20.0	76.6	65.0	65.1
GLM-4.7	20.0	70.6	69.5	65.0
P1-VL-235B-A22B	30.0	71.3	65.5	64.3
o3*	30.0	76.1	59.0	62.9
P1-235B-A22B	22.5	67.2	65.8	62.0
GPT-OSS-120B (high)	38.8	63.4	61.5	60.0
Qwen3-VL-235B-A22B-Thinking	26.3	61.9	57.8	56.3
Qwen3-235B-A22B-Thinking-2507	26.3	58.1	57.3	54.5
P1-30B-A3B	15.0	61.9	56.3	54.4
P1-VL-30B-A3B	20.0	58.8	54.0	52.5
Qwen3-VL-30B-A3B-Thinking	18.8	49.4	43.5	43.4
Qwen3-30B-A3B-Thinking-2507	10.0	47.8	45.3	42.8
GPT-4o*	3.0	12.4	14.1	12.3

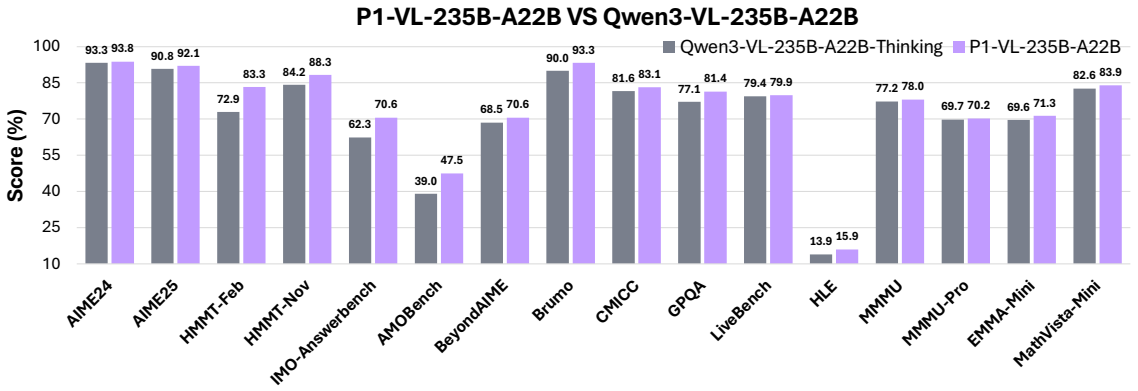

Figure 2. Evaluation of P1-VL-235B-A22B models against their respective base models on out-of-domain benchmarks. The evaluation suite encompasses both text-only and multi-modal tasks, spanning mathematical reasoning and broader STEM disciplines.

model. For every problem, we conducted 8 independent inference runs and computed the average score per problem. These averages were then aggregated across problems to yield the final exam score for each Olympiad.

Evaluation Method. Adhering to the HiPhO protocol (Yu et al., 2025a), we employ `Gemini-2.5-Flash` as the automated evaluator. The evaluation framework mirrors official Olympiad grading by integrating both *answer-level* verification and *step-level* reasoning analysis. For each problem, the final score is determined as the maximum of the answer-based and step-based evaluations. Consequently, we report the cumulative *exam score*, facilitating a direct comparison between model performance and official human medal thresholds.

4.2. Evaluation on Physics Olympiads

Superior Single-Model Performance. Evaluation results are presented in Table 2, verifying the effectiveness of our RL training strategy. As noted above, the P1-VL model series is trained purely through reinforcement learning.

- **P1-VL-235B-A22B** ranks **3rd** among all evaluated models, following only Gemini-3-Pro and GPT-5.2. With 12 gold and 1 silver medals (average score: 39.3), it surpasses the closed-source models such as Gemini-2.5-Pro (37.7), GPT-5 (37.4), and Grok-4 (35.8). Notably, the standalone P1-VL outperforms the agent-augmented baseline (`P1-235B-A22B` + `PhysicsMinions`, 38.4), proving that intrinsic visual reasoning capabilities can surpass purely agentic workflows in complex physical scenarios,

which incorporates image caption information. It shows that our models could effectively perceive the image information and bridge it to the scientific reasoning (case study for Text+Variable Figure is shown in Appendix A.8). Compared with its base model (Qwen3-VL-235B-A22B-Thinking, 33.9), it improves 5.4 points on average.

- **P1-VL-30B-A3B** achieves 9 gold and 4 silver medals with an average score of 35.0. It ranks third among open-source models, behind only DeepSeek-V3.2-Thinking and P1-235B-A22B. Despite its smaller size, it outperforms larger baselines such as Qwen3-VL-235B-A22B-Thinking (33.9) and Qwen3-235B-A22B-Thinking-2507 (33.5), demonstrating significant parameter efficiency. It also outperforms some closed-source models such as o4-mini and Gemini-2.5-Flash-Thinking, showing excellent performance on scientific reasoning.

Agentic Augmentation. With the agentic augmentation of *PhysicsMinions*, the average score of `P1-VL-235B-A22B` improves from 39.3 to 40.9, propelling it to the **2nd rank globally** and surpassing the closed-source GPT-5.2 (40.6). This combined system establishes new state-of-the-art performance on three physics Olympiads, including PanPhO 2025 (66.5 vs 66.3), PanPhO 2024 (83.3 vs 82.5), and PanMechanics 2024 (84.8 vs 82.3). These gains underscore the efficacy of the "model + system" paradigm, demonstrating how multi-agent collaboration significantly amplifies complex scientific reasoning capabilities.

5. Discussion

5.1. Generalizability of P1-VL

We conduct specialized post-training to enhance physics problem-solving; however, a critical question remains: *Does this domain-specific optimization compromise or catalyze general STEM reasoning?*

To investigate the model's generalization capabilities, we evaluate P1-VL on the open-source benchmark *FrontierScience-Olympiad* (Wang et al., 2026). As detailed in Table 3, the results demonstrate robust positive transfer: both P1-VL-235B-A22B and P1-VL-30B-A3B achieve significant gains over their base counterparts across all three scientific domains, yielding a total score improvement of 8.0 and 9.1, respectively. Furthermore, when augmented with the *PhysicsMinions* agent framework, `P1-VL-235B-A22B+PhysicsMinions` attains a total score of 67.1, securing state-of-the-art performance among all evaluated open-source models.

Besides, we compare the P1-VL series with their respective base models across diverse benchmarks. Figure 2 summarizes the comparative results of P1-VL-235B-A2B, illustrating that our models consistently surpass their base counterparts across both text-only and multi-modal benchmarks. For example, in the domain of advanced mathematics, P1-VL-235B-A22B outperforms their respective baselines on AMOBench by margins of 8.5. The results of P1-VL-30B-A3B are shown in Appendix A.4. This pattern confirms that P1-VLs not only retain but actively enhance reasoning capabilities beyond the target physics domain, extending even to rigorous challenges such as IMO-AnswerBench and AMOBench. The advantage is not limited to text-based reasoning but extends significantly to visual perception. Taking the multi-image reasoning task EMMA-Mini as a prime example, P1-VL-235B-A22B achieves a substantial gain of 1.7 points over their baseline. This result specifically highlights our model's enhanced capacity to synthesize information across multiple images and little catastrophic forgetting issue (Luo et al., 2025b) during RL training.

6. Related Work

Recent work in physics-grounded reasoning has focused on leveraging LLM-guided reinforcement learning to tackle scientifically rigorous and competition-level physics problems. Modern physics reasoning benchmarks such as PhysReason (Zhang et al., 2025b), PhysUniBench (Wang et al., 2025), UGPhysics (Xu et al., 2025), Phybench (Qiu et al., 2026), and SeePhys (Xiang et al., 2026) evaluate models' capabilities in long-horizon reasoning, diagram interpretation, and step-by-step problem solving, revealing persistent gaps in current LLM reasoning and highlighting the need for interactive, verifiable scientific reasoning. In this direction, P1 demonstrated that RLVR can train models to achieve Olympiad-level physics problem solving (Chen et al., 2025b), and Sim2Reason explored RL-guided physics reasoning using structured interactions (Prabhudesai et al., 2026). Building on these lines of work, P1-VL integrates visual-language grounding and multimodal reasoning to tackle olympiad physics.

7. Conclusion

In this work, we presented **P1-VL**, the first open-source Vision-Language Model family capable of achieving Olympiad-level proficiency in physics. By addressing the critical visual-logical gap inherent in physical problem solving, P1-VL overcomes the limitations of text-only reasoning. We proposed to improve the reasoning capacity by Curriculum RL Training to internalize robust reasoning patterns and Agentic Augmentation to enable rigorous test-time reflection. The evaluation results on the HiPhO benchmark, where P1-VL-235B-A22B secured 12 gold medals and demonstrated superior generalization across broader scientific disciplines, underscore the potential of P1-VLs in modeling complex scientific reasoning systems. By releasing P1-VL to the research community, we aim to accelerate

the transition from symbolic manipulation to genuine machine scientific discovery, paving the way for AI systems that can truly understand and navigate the physical world.

8. Acknowlegement

This work is supported by Shanghai AI Laboratory and a locally commissioned task from the Shanghai Municipal Government. We would like to extend our special thanks to the developers and maintainers of the following open-source projects, which have been critical to the implementation of this work. This includes Qwen3-VL (Bai et al., 2025b) and Qwen (Yang et al., 2025), which provided the foundational base models for our research; slime (Zhu et al., 2025b), whose innovative framework enabled efficient reinforcement learning in our training pipeline; and verl (Sheng et al., 2024), which offered a versatile reinforcement learning framework to support model training. We also thank sglang (Zheng et al., 2024) for its efficient infrastructure for LLM and VLM serving and inference, and Megatron-LM (Shoeybi et al., 2019) for providing the large-scale model training framework.

Bibliography

Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., Bach, N., Bao, J., Benhaim, A., Cai, M., Chaudhary, V., Chen, C., et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

An, S., Cai, X., Cao, X., Li, X., Lin, Y., Liu, J., Lv, X., Ma, D., Wang, X., Wang, Z., and Zhou, S. Amo-bench: Large language models still struggle in high school math competitions, 2025. URL https://arxiv.org/abs/2510.26768.

Anthropic. Claude 3.7 sonnet system card. URL https://www.anthropic.com/news/visible-extended-thinking.

Bai, L., Cai, Z., Cao, M., Cao, W., Chen, C., Chen, H., Chen, K., Chen, P., Chen, Y., Chen, Y., Cheng, Y., Cheng, Y., Chu, P., Chu, T., Cui, E., Cui, G., Cui, L., Cui, Z., Deng, N., Ding, N., Dong, N., Dong, P., Dou, S., Du, S., Duan, H., Fan, C., Gao, B., Gao, C., Gao, J., Gao, S., Gao, Y., Gao, Z., Ge, J., Ge, Q., Gu, L., Gu, Y., Guo, A., Guo, Q., Guo, X., He, C., He, J., Hong, Y., Hou, S., Hu, C., Hu, H., Hu, J., Hu, M., Hua, Z., Huang, H., Huang, J., Huang, X., Huang, Z., Jiang, Z., Kong, L., Li, L., Li, P., Li, P., Li, S., Li, T., Li, W., Li, Y., Lin, D., Lin, J., Lin, T., Lin, Z., Liu, H., Liu, J., Liu, J., Liu, J., Liu, K., Liu, K., Liu, K., Liu, S., Liu, S., Liu, W., Liu, X., Liu, Y., Liu, Z., Lu, Y., Lv, H., Lv, H., Lv, H., Lv, Q., Lv, Y., Lyu, C., Ma, C., Ma, J., Ma, R., Ma, R., Ma, R., Ma, X., Ma, Y., Ma, Z., Mi, S., Ning, J., Ning, W., Pang, X., Peng, J., Peng, R., Qiao, Y., Qiu, J., Qu, X., Qu, Y., Ren, Y., Shang, F., Shao, W., Shen, J., Shen, S., Song, C., Song, D., Song, D., Su, C., Su, W., Sun, W., Sun, Y., Tan, Q., Tang, C., Tang, H., Tang, K., Tang, S., Tong, J., Wang, A., Wang, B., Wang, D., Wang, L., Wang, R., Wang, W., Wang, W., Wang, Y., Wang, Z., Wu, L.-I., Wu, W., Wu, Y., Wu, Z., Xiao, L., Xing, S., Xu, C., Xu, H., Xu, J., Xu, R., Xu, W., Yang, G., Yang, Y., Ye, H., Ye, J., Ye, S., Yu, J., Yu, J., Yu, J., Yuan, F., Zhang, B., Zhang, C., Zhang, C., Zhang, H., Zhang, J., Zhang, Q., Zhang, Q., Zhang, S., Zhang, T., Zhang, W., Zhang, W., Zhang, Y., Zhang, Z., Zhao, H., Zhao, Q., Zhao, X., Zhao, X., Zhou, B., Zhou, D., Zhou, P., Zhou, Y., Zhou, Y., Zhu, D., Zhu, L., and Zou, Y. Intern-s1: A scientific multimodal foundation model, 2025a. URL https://arxiv.org/abs/2508.15763.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. Qwen3-vl technical report, 2025b. URL https://arxiv.org/abs/2511.21631.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025c.

Balunović, M., Dekoninck, J., Petrov, I., Jovanović, N., and Vechev, M. Matharena: Evaluating llms on uncontaminated math competitions, February 2025. URL https://matharena.ai/.

ByteDance-Seed. Beyondaime: Advancing math reasoning evaluation beyond high school olympiads. https://huggingface.co/datasets/ByteDance-Seed/BeyondAIME, 2025.

Chen, D., Yu, Q., Wang, P., Zhang, W., Tang, B., Xiong, F., Li, X., Yang, M., and Li, Z. xverify: Efficient answer verifier for reasoning model evaluations. *arXiv preprint arXiv:2504.10481*, 2025a.

Chen, J., Cheng, Q., Yu, F., Wan, H., Zhang, Y., Zheng, S., Yao, J., Zhang, Q., He, H., Luo, Y., Zhao, Y., Wang, F., Sheng, L., Xie, C., Zuo, Y., Li, Y., Zeng, W., Wu, Y., Huang, R., Zhou, D., Chen, K., Qiao, Y., Bai, L., Cheng, Y., Ding, N., Zhou, B., Ye, P., and Cui, G. P1: Mastering physics olympiads with reinforcement learning, 2025b. URL https://arxiv.org/abs/2511.13612.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., 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.

Cui, G., Yuan, L., Wang, Z., Wang, H., Li, W., He, B., Fan, Y., Yu, T., Xu, Q., Chen, W., et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025a.

Cui, G., Zhang, Y., Chen, J., Yuan, L., Wang, Z., Zuo, Y., Li, H., Fan, Y., Chen, H., Chen, W., et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025b.

Deepmind, G. Gemini-3-pro system card. URL https://deepmind.google/models/gemini/pro/.

DeepSeek-AI, Liu, A., Mei, A., Lin, B., Xue, B., Wang, B.-L., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., Lu, C., Zhao, C., Deng, C., Xu, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Li, E., Zhou, F., Lin, F., Dai, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Li, H., Liang, H., Wei, H., Zhang, H., sheng Luo, H., Ji, H., Ding, H., Tang, H., Cao, H., Gao, H., Qu, H., Zeng, H., Huang, J., Li, J., Xu, J., Hu, J., Chen, J., Xiang, J., Yuan, J., Cheng, J., Zhu, J., Ran, J., Jiang, J., Qiu, J., Li, J., Song, J.-M., Dong, K., Gao, K., Guan, K., Huang, K., Zhou, K., wei Huang, K., Yu, K., Wang, L., Zhang, L., Wang, L., Zhao, L., Yin, L., Guo, L., Luo, L.-L., Ma, L., Wang, L., Zhang, L., Di, M. S., Xu, M. Y., Zhang, M., Zhang, M., Tang, M., Zhou, M., Huang, P., Cong, P., Wang, P., Wang, Q., Zhu, Q., Li, Q., Chen, Q., Du, Q., Xu, R., Ge, R., Zhang, R., Pan, R., Wang, R., Yin, R., Xu, R., Shen, R., Zhang, R., Liu, S. H., Lu, S., Zhou, S., Chen, S., Cai, S., Chen, S., Hu, S., Liu, S., Hu, S., Ma, S., Wang, S., Yu, S., Zhou, S., Pan, S., Zhou, S., Ni, T., Yun, T., Pei, T., Ye, T., Yue, T., Zeng, W., Liu, W., Liang, W., Pang, W., Luo, W., Gao, W., Zhang, W., Gao, X., Wang, X., Bi, X., Liu, X., Wang, X., Chen, X., Zhang, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yu, X., Li, X., Yang, X., Li, X., Chen, X., Su, X., Pan, X., Lin, X., Fu, X., Wang, Y. Q., Zhang, Y., Xu, Y., Ma, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Qian, Y., Yu, Y., Zhang, Y., Ding, Y., Shi, Y., Xiong, Y., He, Y., Zhou, Y., Zhong, Y., Piao, Y., Wang, Y., Chen, Y., Tan, Y., Wei, Y., Ma, Y., Liu, Y., Yang, Y., Guo, Y., Wu, Y., Wu, Y., Cheng, Y., Ou, Y., Xu, Y., Wang, Y., Gong, Y., Wu, Y., Zou, Y.-H., Li, Y., Xiong, Y., Luo, Y.-W., mei You, Y., Liu, Y., Zhou, Y., Wu, Z. F., Ren, Z., Zhao, Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., guo Zhang, Z., Hao, Z., Gou, Z., Ma, Z., Yan, Z., Shao, Z., Huang, Z., Wu, Z., Li, Z., Zhang, Z., Xu, Z., Wang, Z., Gu, Z., Zhu, Z., Li, Z.-R., Zhang, Z., Xie, Z., Gao, Z., Pan, Z., Yao, Z., Feng, B., Li, H., Cai, J. L., Ni, J., Xu, L., Li, M., Tian, N., Chen, R. J., Jin, R., Li, S. S., Zhou, S., Sun, T., Li, X. Q., Jin, X., Shen, X., Chen, X., Song, X., Zhou, X., Zhu, Y. X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Huang, Z., Xu, Z., Zhang, Z., Ji, D.-L., Liang, J., Guo, J., Chen, J., Xia, L., Wang, M., Li, M., Zhang, P., Chen, R., Sun, S., Wu, S.-K., Ye, S.-Y., T.Wang, Xiao, W. L., An, W., Wang, X., Sun, X., Wang, X., Tang, Y., Zha, Y., Zhang, Z.-N., Ju, Z., Zhang, Z., and Qu, Z. Deepseek-v3.2: Pushing the frontier of open large language models. 2025. URL https://api.semanticscholar.org/CorpusID:283448719.

Feng, S., Ma, R., Yan, X., Fan, Y., Hu, Y., Huang, S., Zhang, S., Cao, Z., Peng, T., Yuan, J., Guo, Z., Zhong, Z., Du, S., Wang, W., Shi, J., Zhou, Y., He, X., Yu, Z., Yu, F., Zheng, Q., Wu, J., Liu, M., Zhang, C., Hou, S., Li, S., Jiang, Y., Lou, W., Wang, L., Wang, Z., Wang, J., Xu, W., Deng, Y., Liu, D., Wang, Y., Zhang, W., Ling, F., Zhang, S., Wang, X., Zheng, S., Huang, X., Sun, S., Hu, S., Ye, P., Song, C., Wang, B., He, C., Liu, Y., Li, X., Hou, Q., Chen, T., Yue, X., Wang, B., He, L., Lin, D., Zhou, B., Zhang, B., and Bai, L. Internagent-1.5: A unified agentic framework for long-horizon autonomous scientific discovery. *arXiv preprint arXiv:2602.08990*, 2026.

Gibney, E. Deepmind unveils 'spectacular'general-purpose science ai. *Nature*, 641(8064):827–828, 2025.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Hao, Y., Gu, J., Wang, H. W., Li, L., Yang, Z., Wang, L., and Cheng, Y. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark, 2025. URL https://arxiv.org/abs/2501.05444.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Kydlíček, H. Math-Verify: Math Verification Library. URL https://github.com/huggingface/math-verify.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Liu, J., Li, Y., Fu, Y., Wang, J., Liu, Q., and Shen, Y. When speed kills stability: Demystifying rl collapse from the training-inference mismatch. *Notion Blog*, 2025.

Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., and Gao, J. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

Luo, M., Tan, S., Wong, J., Shi, X., Tang, W. Y., Roongta, M., Cai, C., Luo, J., Li, L. E., Popa, R. A., and Stoica, I. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL, 2025a. Notion Blog.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., and Zhang, Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 2025b.

Luong, T., Hwang, D., Nguyen, H. H., Ghiasi, G., Chervonyi, Y., Seo, I., Kim, J., Bingham, G., Lee, J., Mishra, S., Zhai, A., Hu, C. H., Michalewski, H., Kim, J., Ahn, J., Bae, J., Song, X., Trinh, T. H., Le, Q. V., and Jung, J. Towards robust mathematical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://aclanthology.org/2025.emnlp-main.1794/.

Merry, A., Rajbhandari, S., Shoeybi, M., Puri, R., Fung, P., Anandkumar, A., and Catanzaro, B. Fully sharded data parallel: Reducing memory usage for large model training. *PyTorch Developer Blog*, 2021. URL https://pytorch.org/blog/.

Meta. Llama 4 system card. URL https://www.llama.com/docs/model-cards-and-prompt-formats/llama4/.

Meurer, A., Smith, C. P., Paprocki, M., Čertík, O., Kirpichev, S. B., Rocklin, M., Kumar, A., Ivanov, S., Moore, J. K., Singh, S., et al. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, 2017.

Mistral. Mistral-medium-3 system card. URL https://mistral.ai/news/mistral-medium-3.

OpenAI. Gpt-4o system card, a. URL https://openai.com/index/gpt-4o-system-card/.

OpenAI. Gpt-5 system card, b. URL https://openai.com/index/gpt-5-system-card/.

OpenAI. Gpt-5.2 system card, c. URL https://openai.com/index/introducing-gpt-5-2/.

OpenAI. Openai o3 and o4-mini system card, d. URL https://openai.com/index/introducing-o3-and-o4-mini/.

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Prabhudesai, M., Satpathy, A., Li, Y., Qin, Z., Bhardwaj, N., Zadeh, A., Li, C., Fragkiadaki, K., and Pathak, D. Solving physics olympiad via reinforcement learning on physics simulators. *arXiv preprint arXiv:2604.11805*, 2026.

Qiu, J., Shi, J., Juan, X., Zhao, Z., Geng, J., Liu, S., Wang, H., Wu, S., and Wang, M. Physics supernova: Ai agent matches elite gold medalists at ipho 2025. *arXiv preprint arXiv:2509.01659*, 2025.

Qiu, S., Guo, S., Song, Z.-Y., Sun, Y., Cai, Z., Wei, J., Luo, T., Yin, Y., Haoxu, Z., Hu, Y., et al. Phybench: Holistic evaluation of physical perception and reasoning in large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., and Wu, C. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*, 2024.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Sutton, R. S., Barto, A. G., et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Team, G.-V., Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., Duan, S., Wang, W., Wang, Y., Cheng, Y., He, Z., Su, Z., Yang, Z., Pan, Z., Zeng, A., Wang, B., Chen, B., Shi, B., Pang, C., Zhang, C., Yin, D., Yang, F., Chen, G., Xu, J., Zhu, J., Chen, J., Chen, J., Chen, J., Lin, J., Wang, J., Chen, J., Lei, L., Gong, L., Pan, L., Liu, M., Xu, M., Zhang, M., Zheng, Q., Yang, S., Zhong, S., Huang, S., Zhao, S., Xue, S., Tu, S., Meng, S., Zhang, T., Luo, T., Hao, T., Tong, T., Li, W., Jia, W., Liu, X., Zhang, X., Lyu, X., Fan, X., Huang, X., Wang, Y., Xue, Y., Wang, Y., Wang, Y., An, Y., Du, Y., Shi, Y., Huang, Y., Niu, Y., Wang, Y., Yue, Y., Li, Y., Zhang, Y., Wang, Y., Wang, Y., Zhang, Y., Xue, Z., Hou, Z., Du, Z., Wang, Z., Zhang, P., Liu, D., Xu, B., Li, J., Huang, M., Dong, Y., and Tang, J. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025a. URL https://arxiv.org/abs/2507.01006.

Team, K., Bai, Y., Bao, Y., Chen, G., Chen, J., Chen, N., Chen, R., Chen, Y., Chen, Y., Chen, Y., et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025b.

Wang, L., Su, E., Liu, J., Li, P., Xia, P., Xiao, J., Zhang, W., Dai, X., Chen, X., Meng, Y., et al. Physunibench: an undergraduate-level physics reasoning benchmark for multimodal models. *arXiv preprint arXiv:2506.17667*, 2025.

Wang, M., Lin, R., Hu, K., Jiao, J., Chowdhury, N., Chang, E., and Patwardhan, T. Frontierscience: Evaluating ai's ability to perform expert-level scientific tasks. *arXiv preprint arXiv:2601.21165*, 2026.

Wen, X., Liu, Z., Zheng, S., Xu, Z., Ye, S., Wu, Z., Liang, X., Wang, Y., Li, J., Miao, Z., et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025.

White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B., Jain, S., Shwartz-Ziv, R., Jain, N., Saifullah, K., Naidu, S., et al. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*, 4, 2024.

Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

xAI. Grok 4 system card. URL https://x.ai/grok.

Xiang, K., Li, H., Zhang, T. J., Huang, Y., Liu, Z., Qu, P., He, J., Chen, J., Yuan, Y.-J., Han, J., et al. Seephys: Does seeing help thinking?–benchmarking vision-based physics reasoning. *Advances in Neural Information Processing Systems*, 38, 2026.

Xu, X., Xu, Q., Xiao, T., Chen, T., Yan, Y., Zhang, J., Diao, S., Yang, C., and Wang, Y. Ugphysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. *arXiv preprint arXiv:2502.00334*, 2025.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yao, F., Liu, L., Zhang, D., Dong, C., Shang, J., and Gao, J. Your efficient rl framework secretly brings you off-policy rl training, August 2025. URL https://fengyao.notion.site/off-policy-rl.

Yu, F., Wan, H., Cheng, Q., Zhang, Y., Chen, J., Han, F., Wu, Y., Yao, J., Hu, R., Ding, N., Cheng, Y., Chen, T., Bai, L., Zhou, D., Luo, Y., Cui, G., and Ye, P. Hipho: How far are (m)llms from humans in the latest high school physics olympiad benchmark? *arXiv preprint arXiv:2509.07894*, 2025a.

Yu, F., Yao, J., Wang, Z., Wan, H., Huang, Y., Zhang, B., Hu, S., Zhou, D., Ding, N., Cui, G., Bai, L., Ouyang, W., and Ye, P. Physicsminions: Winning gold medals in the latest physics olympiads with a coevolutionary multimodal multi-agent system, 2025b. URL https://arxiv.org/abs/2509.24855.

Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025c.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., and Chen, W. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024a.

Yue, X., Zheng, T., Ni, Y., Wang, Y., Zhang, K., Tong, S., Sun, Y., Yu, B., Zhang, G., Sun, H., Su, Y., Chen, W., and Neubig, G. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024b.

Zhang, Q., Wu, H., Zhang, C., Zhao, P., and Bian, Y. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *Advances in neural information processing systems*, 2025a.

Zhang, X., Dong, Y., Wu, Y., Huang, J., Jia, C., Fernando, B., Shou, M. Z., Zhang, L., and Liu, J. Physreason: A comprehensive benchmark towards physics-based reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16593–16615, 2025b.

Zhang, Y., Chen, X., Jin, B., Wang, S., Ji, S., Wang, W., and Han, J. A comprehensive survey of scientific large language models and their applications in scientific discovery. *arXiv preprint arXiv:2406.10833*, 2024.

Zheng, C., Liu, S., Li, M., Chen, X.-H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025a.

Zheng, L., Yin, L., Xie, Z., Sun, C. L., Huang, J., Yu, C. H., Cao, S., Kozyrakis, C., Stoica, I., Gonzalez, J. E., et al. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583, 2024.

Zheng, S., Cheng, Q., Yao, J., Wu, M., He, H., Ding, N., Cheng, Y., Hu, S., Bai, L., Zhou, D., et al. Scaling physical reasoning with the physics dataset. *arXiv preprint arXiv:2506.00022*, 2025b.

Zheng, Y., Koh, H. Y., Ju, J., Nguyen, A. T., May, L. T., Webb, G. I., and Pan, S. Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence*, pp. 1–11, 2025c.

Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a.

Zhu, Z., Xie, C., Lv, X., and slime Contributors. slime: An llm post-training framework for rl scaling. `https://github.com/THUDM/slime`, 2025b. GitHub repository. Corresponding author: Xin Lv.

A. Appendix

A.1. System Prompt

System Prompt of Multi-box Style

```
Please answer the problem adhering to the following rules:
1. Please use LaTeX format to represent the variables and formulas
   used in the solution process and results.
2. Please put the final answer(s) in \boxed{}, note that the unit
   of the answer should not be included in \boxed{}.
3. If the problem requires multiple answers, list them in order,
   each in a separate \boxed{}.
```

Figure 3. System prompt design for P1-VL training.

A.2. Curriculum Training Phrases.

Settings. We detail the stage-wise hyperparameter settings of the P1-VL post-training process in Table 4. Following our curriculum framework, we progressively amplify the training complexity by scaling the data difficulty (metric defined in Eq. 8) and expanding the exploration space via larger group sizes and extended generation windows. To bridge the train-inference gap, we consistently apply *Masked Importance Sampling*—a correction technique for off-policy stability—across all stages. Notably, our optimization backbone is built upon the GSPO algorithm to ensure stable convergence for MoE architectures. Furthermore, we exclusively employ rule-based verifiers during training to mitigate reward hacking.

Table 4. Configuration of different phrases in P1-VL training.

Model	**P1-VL-30B-A3B**			**P1-VL-235B-A22B**		
	Stage 1	Stage 2	Stage 3	Stage 1	Stage 2	Stage 3
Group size	8	8	16	8	8	16
Difficulty	(0.0,0.7]	(0.0,0.5]	[0.0,0.5]	(0.0,0.7]	(0.0,0.5]	[0.0.0.5]
Generation Window	40k	60k	80k	48k	54k	54k
Learning Rate	1×10^{-6}			1×10^{-6}		
Algorithm	GSPO w. MIS			GSPO w. MIS		
Verifier Choice	Rule-based Verifier Only			Rule-based Verifier Only		
Rollout Batch Size	2048			1024		
Update Batch Size	256			256		

Analysis. Figure 4 illustrates the training dynamics across the curriculum RL process. A key observation is the gradual increase in response length, indicating that the model is developing the capacity for deeper reasoning to tackle complex problems. To accommodate this evolving capability, we strategically expand the difficulty and meanwhile expand exploration space by increasing the generation window and group size at each stage transition. Quantitatively, validation results on HiPhO demonstrate a steady performance improvement throughout training, underscoring both the effectiveness and the stability of our algorithm.

A.3. Evaluated Models

- **Closed-source models:** Gemini-3-Pro (Deepmind), GPT-5.2 (OpenAI, c), GPT-5 (OpenAI, b), o3 (OpenAI, d), o4-mini (high) (OpenAI, d), o4-mini (OpenAI, d), GPT-4o (OpenAI, a), Gemini-2.5-Pro (Comanici et al., 2025), Gemini-2.5-Flash-Thinking (Comanici et al., 2025), Grok-4 (xAI), Claude-4-Sonnet-Thinking (Anthropic), Claude-4-Sonnet (Anthropic), Mistral-Medium-3 (Mistral).
- **Open-source models:** DeepSeek-V3.2-Thinking (DeepSeek-AI et al., 2025), Qwen3-VL Series (Bai et al., 2025b), Intern-S1 (Bai et al., 2025a), InternVL3 Series (Zhu et al., 2025a), Qwen2.5-VL Series (Bai et al., 2025c), GLM-4.5V (Team et al., 2025a), DeepSeek-VL2 (Wu et al., 2024), LLaMA4-Scout-17B (Meta), Phi-4-multimodal (Abouelenin et al.,

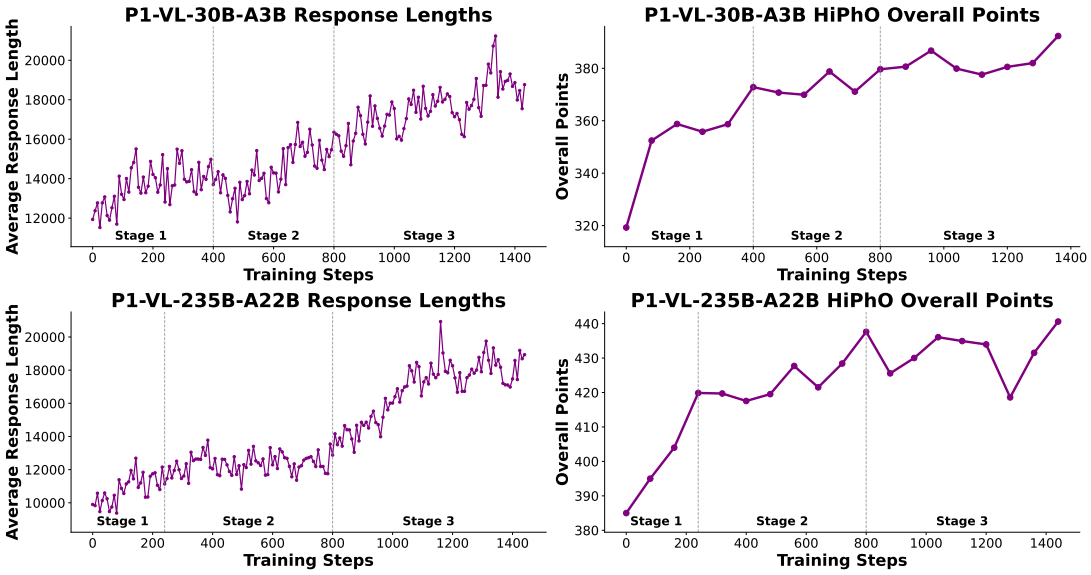

Figure 4. The training dynamics of P1-VL curriculum RL. The upper side shows the dynamics of P1-VL-30B-A3B, and the lower shows those of P1-VL-235B-A22B. The training is set in three stages with curriculum difficulty expansion, group size expansion, and generation window expansion. The average response length represents the generation length during training, and the overall points refer to the sum score of all the exams in HiPhO with model-based verifier (Qwen-3-30B-Instruct-2507).

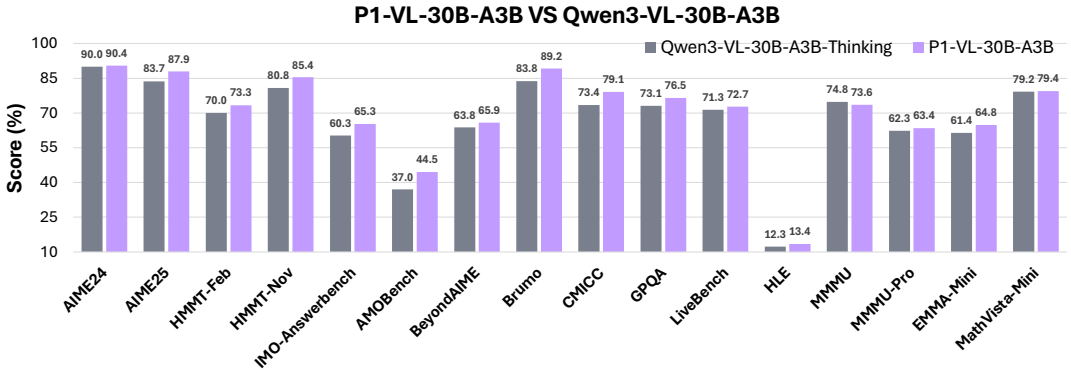

Figure 5. Evaluation of P1-VL-30B-A3B models against their respective base models on out-of-domain benchmarks. The evaluation suite encompasses both text-only and multi-modal tasks, spanning mathematical reasoning and broader STEM disciplines.

2025), P1 Series (Chen et al., 2025b), GPT-OSS-120B (Agarwal et al., 2025), Kimi-K2-Thinking (Team et al., 2025b), Kimi-K2-Instruct (Team et al., 2025b), DeepSeek-R1 (Guo et al., 2025), DeepSeek-V3 (Liu et al., 2024), Qwen3 Series (Yang et al., 2025).

A.4. Evaluated OOD Benchmarks

- **Text Benchmarks**: AIME24, AIME25, HMMT-Nov, HMMT-Feb, Brumo, CMICC (Balunović et al., 2025), IMO-AnswerBench (Luong et al., 2025), AMOBench (An et al., 2025), BeyondAIME (ByteDance-Seed, 2025), GPQA (Rein et al., 2024)), and a general reasoning task (LiveBench (White et al., 2024).
- **Multi-modal Benchmarks**: five STEM-oriented benchmarks, including HLE (Phan et al., 2025)), MMMU (Yue et al., 2024a), MMMU-Pro (Yue et al., 2024b), EMMA-Mini (Hao et al., 2025), and MathVista-Mini (Lu et al., 2024)[2].

The out-of-domain evaluation results of P1-VL-30B-A3B are shown in Figure 5.

[2]For MathVista-Mini and MMMU, we adopt the evaluation scripts posted on the Qwen3-VL repository, and for MMMU-Pro we adopt the setting in MMMU repository.

A.5. Effect of Training Data Composition

We also empirically analyze the impact of training data composition. Figure 6 illustrates the performance of Qwen3-VL-4B-Thinking (using XVerifier) when trained exclusively on questions containing images. Crucially, we observe that incorporating text-only data (padded with blank images) induces no negative transfer; rather, it yields superior performance almost all the time. It demonstrates the feasibility and superiority of using mixed multi-modal data. Consequently, we train our P1-VL models using a combination of both the text-only and image-text data.

Figure 6. Comparison of Qwen3-VL-4B-Thinking training performance on the image-text data versus the mixed data (text-only + image-text).

A.6. Effect of Curriculum Training

Figure 7 illustrates the impact of curriculum training on model evolution. Each line includes a period of observation after the parameters are fixed. It shows that in the absence of difficulty expansion, the response length fluctuates stagnantly across Stages 1 and 2, resulting in negligible performance gains. Conversely, introducing the curriculum strategy yields a marked increase in both average response length and reasoning accuracy. These results underscore the critical role of our curriculum RL framework in driving deep reasoning.

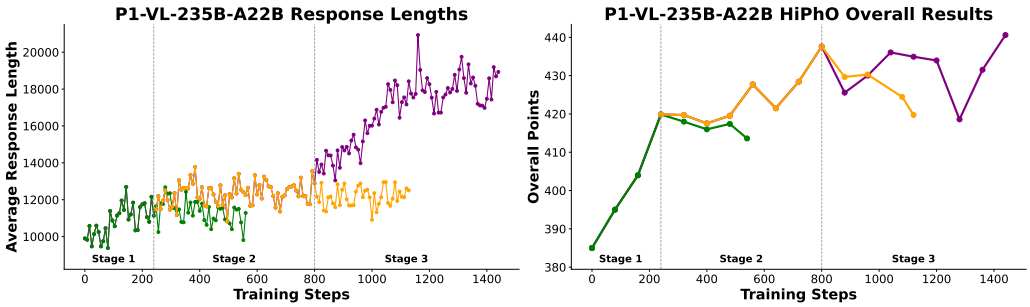

Figure 7. Training dynamics of RL training with and without curriculum difficulty expansion.

A.7. RL Training on Intern Series

We further verify the effectiveness of our RL training on the InternVL-3.5-30B-A3B and Intern-S1-mini models. On the HiPhO benchmark, InternVL-3.5-30B-A3B performance improves from 18.3 to 22.6, while Intern-S1-mini sees an increase from 8.5 to 12.4 (utilizing Qwen3-30B-A3B-Instruct as the model-based verifier). Using Gemini-2.5-Flash as the verifier, the trained Intern-S1-mini achieves one silver and ten bronze medals, an improvement over the baseline of ten bronze medals. These consistent improvements demonstrate that our RL training strategy is effective in eliciting the models' latent capabilities in scientific reasoning.

A.8. Case Study I: Text+Variable Figure

This case study examines a complex problem from the 2025 International Physics Olympiad (IPhO), which investigates the physical principles of Cox's 18th-century timepiece—an ingenious device that harnesses atmospheric pressure fluctuations to generate energy. The problem requires understanding the figure information to determine the states of the tubes and the behaviors, and analyzing the table information. The P1-VL-235B-A22B model achieved a perfect score (1.0/1.0) on this problem, demonstrating strong capabilities in:

- **Physical intuition:** Correctly identifying the critical force balance constraint at the stop position.
- **Visual-Structural Alignment:** Accurately translating the schematic representation of the tube states (liquid levels and gas volumes) into precise hydrostatic equations, effectively bridging visual features with symbolic reasoning.
- **Tabular Information Integration:** Synthesizing discrete parameters from the provided table to constrain the variables, specifically extracting the dimensional specifications required to calculate the total work done.

- **Long-horizon Logical Reasoning:** Sustaining a coherent derivation chain to link atmospheric pressure fluctuations with mechanical energy storage, correctly applying the Ideal Gas Law without hallucinating non-existent constraints.

This performance demonstrates P1-VL's proficiency in tackling competition-level physics problems, requiring a synergy of deep physical understanding, rigorous analytical reasoning, and precise diagrammatic interpretation.

IPhO 2025 Question 2-A.1

Background:

In 1765, British clockmaker James Cox invented a clock whose only source of energy is the fluctuations in atmospheric pressure. Cox's clock used two vessels containing mercury. Changes in atmospheric pressure caused mercury to move between the vessels, and the two vessels to move relative to each other. This movement acted as an energy source for the actual clock.

We propose an analysis of this device. Throughout, we assume that

- the Earth's gravitational field $\vec{g} = -g\vec{u_z}$ is uniform with $g = 9.8\mathrm{m} \cdot \mathrm{s}^{-2}$ and $\vec{u_z}$ a unit vector;
- all liquids are incompressible and their density is denoted rho;
- no surface tension effects will be considered; - the variations of atmospheric pressure with altitude are neglected;
- the surrounding temperature T_a is uniform and all transformations are isothermal.

We first consider a bath of water that occupies the semi-infinite space $z \leq 0$. The air above it is at a pressure $P_\mathrm{a} = P_0$. A cylindrical vertical tube of length $H = 1\mathrm{m}$, cross-sectional area $S = 10\mathrm{cm}^2$ and mass $m = 0.5\mathrm{kg}$ is dipped into the bath. The bottom end of the tube is open, and the top end of the tube is closed. We denote h the altitude of the top of the tube and z_ℓ that of the water inside the tube. The thickness of the tube walls is neglected.

We start from the situation where the tube in Figure 8 contains no gas and its top is at the bath level: in other words, $h = 0$ and $z_l = 0$ (case a). The tube is then slowly lifted until its bottom end reaches the bath level. The pulling force exerted on the tube is denoted $\vec{F} = F\vec{u}_z$.

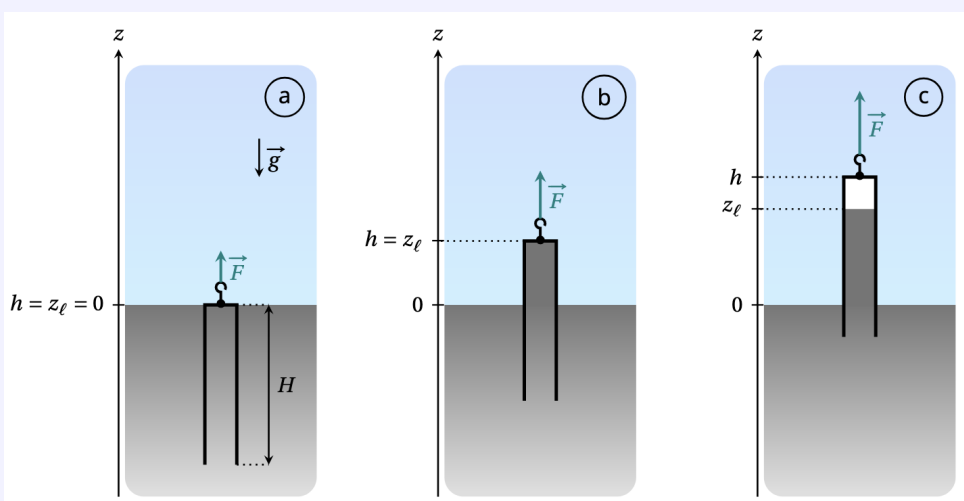

Figure 8. Sketch of the tube in different configurations

Question:

For the configuration shown in Figure 8 (case b), determine the expressions for:

1. The pressure P_w in the water at the top of the tube
2. The force $\vec{F}$ necessary to maintain the tube at this position

Given data: P_0, ρ, m, S, h, g and $\vec{u}_z$

P1-VL Solution Overview

To analyze the situation for case b in Cox's timepiece, we consider the vertical tube of length H, cross-sectional area S, and mass m, lifted so that its bottom end is at the bath surface ($z = 0$) and its top is at altitude h. In this configuration, the water level inside the tube is at the same height as the top of the tube, i.e., $z_\ell = h$, indicating that the tube is completely filled with water.

(A.1) P_w and $\vec{F}$

The tube is sealed at the top and open at the bottom, and the water inside is in hydrostatic equilibrium. The pressure at the bottom of the tube (at $z = 0$) must equal the atmospheric pressure P_0, as the tube is open to the bath at this level.
Using the hydrostatic pressure law, the pressure at any height z inside the water column is given by:

$$P(z) = P_0 - \rho g z \tag{10}$$

At the top of the tube, where $z = h$, the pressure in the water is:

$$P_\mathrm{w} = P_0 - \rho g h \tag{11}$$

This result reflects that the pressure at the top of the water column is reduced by the weight of the water column of height h.

To determine the force $\vec{F}$ necessary to hold the tube in place, we perform a force balance on the tube alone.
Forces on the Tube:

- **Gravitational force:** The tube has a mass m, so its weight is $-mg\vec{u_z}$.
- **Pressure forces:** The atmospheric pressure P_0 acts downward on the top cap of the tube, exerting a force $-P_0 S\vec{u_z}$.The water pressure P_w at the top of the tube acts upward on the inner surface of the cap, exerting a force $+P_\mathrm{w} S\vec{u_z}$.
- **Applied force:** $\vec{F}$ is applied upward to maintain equilibrium.

Thus, the **net pressure force** on the tube is:

$$(P_\mathrm{w} - P_0)S\vec{u_z} = -\rho g h S\vec{u_z} \tag{12}$$

This is a downward force due to the pressure difference between the outside atmosphere and the water column inside the tube.
The net force on the tube must be zero:

$$\vec{F} + (P_\mathrm{w} - P_0)S\vec{u_z} - mg\vec{u_z} = \vec{0} \tag{13}$$

Substituting $P_\mathrm{w} = P_0 - \rho g h$:

$$\vec{F} - \rho g h S\vec{u_z} - mg\vec{u_z} = \vec{0} \Rightarrow \vec{F} = (mg + \rho g h S)\vec{u_z} \tag{14}$$

This result is physically intuitive: the force required to hold the tube in place must counteract both the tube's own weight and the net downward pressure force due to the difference between atmospheric pressure and the reduced pressure at the top of the water column.

Answer: $\boxed{P_{\mathrm{w}} = P_0 - \rho g h}$ $\boxed{\vec{F} = (mg + \rho g h S)\vec{u_z}}$

IPhO 2025 Question 2-A.2

Background: Three experiments are performed. In each, the tube is lifted from the initial state shown in Figure 8(a) under the conditions specified in Figure 9.

Experiment	Liquid	T_{a} (°C)	ρ (kg·m^{-3})	P_{sat} (Pa)
1	Water	20	1.00×10^3	2.34×10^3
2	Water	80	0.97×10^3	47.4×10^3
3	Water	99	0.96×10^3	99.8×10^3

Figure 9. Experimental conditions and numerical values of physical quantities for each experiment (P_{sat} designates the saturated vapour pressure of the pure fluid)

In each case, we study the evolution of the force F that must be applied in order to maintain the tube in equilibrium at an altitude h, the external pressure being fixed at $P_{\mathrm{a}} = P_0 = 1.000 \times 10^5$ Pa. Two different behaviours are possible

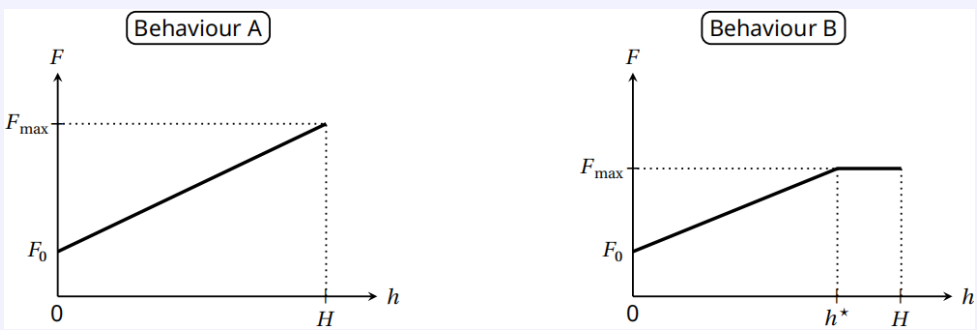

Figure 10. Two different behaviours

Question:
For each experiment, complete the table in the answer sheet to indicate the expected behaviour and the numerical values for $F_{\max}$ and for h^* (when pertinent), where $F_{\max}$ and h^* are defined in the figures illustrating the two behaviours.

P1-VL Solution Overview

To determine the behavior of the tube during lifting and the corresponding values of $F_{\max}$ and h^*, we analyze the pressure equilibrium and the conditions under which vaporization (cavitation) occurs.

(A.2) Answer Sheet

The vertical cylinder, sealed at the top and open at the bottom, is lifted from the bath (where $h = 0$) to the position

where its bottom end reaches the surface (i.e., $h = H = 1\,\mathrm{m}$). The pressure at the top of the tube is governed by:

$$P_{\mathrm{top}} = P_0 - \rho g h \tag{15}$$

Cavitation (formation of vapor) occurs when $P_{\mathrm{top}} = P_{\mathrm{sat}}$, i.e., when:

$$h^* = \frac{P_0 - P_{\mathrm{sat}}}{\rho g} \tag{16}$$

- **Behavior A** occurs when $h^* > H$, meaning the tube remains completely filled with liquid throughout the lifting process ($z_\ell = h$).
- **Behavior B** occurs when $h^* < H$, meaning the water column inside the tube stops rising at h^*, and the top of the tube continues to rise while the water level inside remains at $z_\ell = h^*$.

The force F required to hold the tube in equilibrium is given by:

$$F = mg + (P_0 - P_{\mathrm{top}})S \tag{17}$$

- For **Behavior A**, $P_{\mathrm{top}} = P_0 - \rho g h$, so:

$$F = mg + \rho g S h \tag{18}$$

  At $h = H$,

$$F_{\max} = mg + \rho g S H \tag{19}$$

- For **Behavior B**, once $h > h^*$, $P_{\mathrm{top}} = P_{\mathrm{sat}}$, so:

$$F_{\max} = mg + (P_0 - P_{\mathrm{sat}})S \tag{20}$$

Data Recap

- Cross-sectional area: $S = 10\,\mathrm{cm}^2 = 1 \times 10^{-3}\,\mathrm{m}^2$
- Mass of the tube: $m = 0.5\,\mathrm{kg}$
- Gravitational acceleration: $g = 9.8\,\mathrm{m/s}^2$
- Atmospheric pressure: $P_0 = 1.000 \times 10^5\,\mathrm{Pa}$

Experiment 1: Water at $20^\circ\mathrm{C}$

- $\rho = 1.00 \times 10^3\,\mathrm{kg/m}^3$
- $P_{\mathrm{sat}} = 2.34 \times 10^3\,\mathrm{Pa}$

$$h^* = \frac{1.000 \times 10^5 - 2.34 \times 10^3}{1.00 \times 10^3 \times 9.8} \approx 9.96\,\mathrm{m} > H \tag{21}$$

$\rightarrow$ **Behavior A**

$$F_{\max} = mg + \rho g S H = 0.5 \times 9.8 + 1.00 \times 10^3 \times 9.8 \times 1 \times 10^{-3} = 4.9 + 9.8 = 14.7\,\mathrm{N} \tag{22}$$

Experiment 2: Water at $80^\circ\mathrm{C}$

- $\rho = 0.97 \times 10^3\,\mathrm{kg/m}^3$
- $P_{\mathrm{sat}} = 47.4 \times 10^3\,\mathrm{Pa}$

$$h^* = \frac{1.000 \times 10^5 - 47.4 \times 10^3}{0.97 \times 10^3 \times 9.8} \approx 5.53\,\mathrm{m} > H \tag{23}$$

→ **Behavior A**

$$F_{\max} = mg + \rho g S H = 4.9 + 0.97 \times 10^3 \times 9.8 \times 1 \times 10^{-3} \approx 4.9 + 9.506 = 14.4\,\mathrm{N} \tag{24}$$

Experiment 3: Water at $99°\mathrm{C}$

- $\rho = 0.96 \times 10^3\,\mathrm{kg/m^3}$
- $P_{\mathrm{sat}} = 99.8 \times 10^3\,\mathrm{Pa}$

$$h^* = \frac{1.000 \times 10^5 - 99.8 \times 10^3}{0.96 \times 10^3 \times 9.8} \approx 0.0213\,\mathrm{m} = 2.1\,\mathrm{cm} < H \tag{25}$$

→ **Behavior B**

$$F_{\max} = mg + (P_0 - P_{\mathrm{sat}})S = 4.9 + (200)(1 \times 10^{-3}) = 4.9 + 0.2 = 5.1\,\mathrm{N} \tag{26}$$

Answer:

Experiment	Behaviour (A or B?)	h^* (cm)	$F_{\max}$ (N)
1	A	None	14.7
2	A	None	14.4
3	B	2.1	5.1