
Skywork R1V2: Visual Hybrid Reinforcement Learning for Multimodal Reasoning

Chris*, Yichen Wei*, Peng Yi, Xiaokun Wang, Weijie Qiu, Wei Shen,
Tianyidan Xie, Jiangbo Pei, Jianhao Zhang, Yunzhuo Hao, Xuchen Song[†],
Yang Liu[†], Yahui Zhou

Skywork AI, Kunlun Inc.
chris@kunlun-inc.com, xuchen.song@kunlun-inc.com

Abstract

We present Skywork R1V2, a next-generation multimodal reasoning model and a major leap forward from its predecessor, Skywork R1V. At its core, R1V2 introduces a hybrid reinforcement learning paradigm that harmonizes reward-model guidance with rule-based strategies, thereby addressing the long-standing challenge of balancing sophisticated reasoning capabilities with broad generalization. To further enhance training efficiency, we propose the Selective Sample Buffer (SSB) mechanism, which effectively counters the "Vanishing Advantages" dilemma inherent in Group Relative Policy Optimization (GRPO) by prioritizing high-value samples throughout the optimization process. Notably, we observe that excessive reinforcement signals can induce visual hallucinations, an important trade-off to monitor during training.

Empirical results affirm the exceptional capability of R1V2, with benchmark-leading performances such as 62.6 on OlympiadBench, 79.0 on AIME2024, 64.0 on LiveCodeBench, and 74.0 on MMMU. These results underscore R1V2's superiority over existing open-source models and demonstrate significant progress in closing the performance gap with premier proprietary systems, including Gemini 2.5 and OpenAI o4.

1 Introduction

Recent advancements in "slow-thinking" multimodal models—such as OpenAI-o1, Gemini-Thinking, Kimi-1.5, and Skywork-R1V—have led to substantial progress in complex reasoning tasks across mathematics and science. These models emulate reflective cognitive processes, achieving stronger performance than their "fast-thinking" counterparts, including GPT-o4 and Claude-3.5, which prioritize rapid generation over deliberate reasoning. For instance, slow-thinking models demonstrate over 30% improvement on benchmarks like AIME24 and AMC23, along with approximately 10% gains on science-related benchmarks such as GPQA.

However, extending slow-thinking strategies to multimodal domains introduces new challenges. While improvements are observed on visual reasoning tasks such as MMMU and MathVision, performance degrades on general perception benchmarks like AI2D and OCRBench. This decline is often accompanied by a rise in visual hallucinations, highlighting a critical issue: *How can we effectively promote slow-thinking behavior in vision-language models (VLMs) without compromising their generalization capabilities?*

*Equal contribution

[†]Corresponding author

To tackle this problem, we introduce **Skywork R1V2**, a next-generation vision-language model that acquires multimodal reasoning skills directly via reinforcement learning (RL), bypassing the need for teacher model distillation. R1V2 is trained on a comprehensive dataset spanning visual perception, scientific inquiry, and abstract reasoning, encompassing both general-purpose and reasoning-intensive tasks. The training process utilizes hybrid optimization signals derived from the R1V-RM reward model and structured rule-based feedback.

At its core, R1V2 builds upon the Group Relative Policy Optimization (GRPO) algorithm, which computes relative advantages by comparing candidate responses within the same query group. As training progresses, however, responses may converge (i.e., become uniformly correct or incorrect), resulting in vanishing advantage signals and limited reasoning diversity. To address this, we propose a *Selective Sample Buffer* (SSB), which caches high-quality training examples with non-zero advantages and reintroduces them during policy updates. This mechanism increases gradient density, enhances training stability, and encourages deeper reasoning paths.

In addition to GRPO, we incorporate a *Mixed Preference Optimization* (MPO) strategy that fuses preference-based and rule-based reinforcement signals. This hybrid scheme enables a more effective balance between reasoning specialization and generalization. Notably, we find that overly aggressive optimization of reasoning capabilities may lead to increased hallucination in visual tasks, emphasizing the importance of careful reward calibration.

Extensive evaluations demonstrate the effectiveness of R1V2 across multiple authoritative multimodal benchmarks, including 62.6% on OlympiadBench, 79.0% on AIME2024, 64.0% on LiveCodeBench, and 74.0% on MMMU. These results not only establish new open-source baselines but also substantially reduce the performance gap with proprietary state-of-the-art models such as Gemini 2.5 and GPT-o4.

In summary, Skywork R1V2 offers a promising and scalable framework for training robust and reflective multimodal reasoning agents via direct reinforcement learning. It highlights the potential of unifying high-level reasoning and perceptual understanding, paving the way for the next generation of general-purpose AI systems.

2 Related Work

2.1 Multimodal Reasoning Models

Recent advances in multimodal AI have increasingly focused on enhancing reasoning capabilities across different modalities. Several approaches have emerged to advance this frontier, broadly categorized into distillation-based methods, reinforcement learning techniques, and specialized architectural innovations.

Distillation-based approaches like Claude-3 [1], GPT-4V [18], and Gemini [24] leverage larger text-only reasoning models to transfer their abilities to multimodal contexts. These models typically use a teacher-student framework where the multimodal model is trained to emulate the reasoning processes of powerful text-only predecessors. While effective, this approach relies heavily on the availability of large proprietary text models, limiting accessibility for open research.

Reinforcement learning for multimodal reasoning has gained traction through models like Qwen-VL [3], SPHINX [14], and LLaVA [12]. These methods apply techniques such as RLHF, DPO, and other preference optimization strategies to align model outputs with human preferences for reasoning quality. However, most implementations have focused primarily on general instruction following rather than specialized reasoning capabilities.

Recent architectural innovations include "slow-thinking" approaches like OpenAI-o1 [17], Gemini-Thinking [23], and Kimi-1.5 [25], which introduce specialized mechanisms for extended deliberation. These models allocate additional computational resources to complex reasoning tasks, demonstrating substantial improvements on mathematical and scientific benchmarks. The first generation of Skywork-R1V [20] pioneered the direct application of text reasoning capabilities to vision through advanced adapter techniques, establishing a new paradigm for multimodal reasoning.

However, a common challenge across these approaches is maintaining balance between specialized reasoning abilities and general-purpose multimodal understanding. Models optimized heavily for mathematical reasoning often demonstrate degraded performance on everyday visual tasks, while

general-purpose models struggle with complex analytical reasoning. This fundamental tension motivates our current work.

2.2 Preference Optimization in Multimodal Models

Preference optimization has emerged as a powerful paradigm for aligning AI systems with human expectations, though its extension from language-only to multimodal reasoning presents significant challenges. While foundational techniques like RLHF [5] and DPO [21] have transformed text-based models, their direct application to multimodal contexts requires substantial adaptations to accommodate the complexity of cross-modal interactions.

Early multimodal preference work focused primarily on basic instruction following. Models like BLIP-2 [13] and InstructBLIP [6] pioneered instruction tuning approaches for vision-language alignment, while LLaVA-RLHF [22] and MM-REACT [29] explored human preference incorporation for general multimodal tasks. However, these efforts typically addressed straightforward visual question-answering scenarios rather than complex reasoning chains involving visual evidence.

The specific challenge of preference optimization for multimodal reasoning remains largely unexplored. Although recent models such as DeepSeek-VL [15] and Qwen-VL-Plus [2] have incorporated elements of preference-based training, neither specifically addresses the unique requirements of deep reasoning tasks where visual information must be integrated into extended logical deductions.

We identify two critical limitations in current approaches. First, the binary nature of typical preference pairs fails to capture the nuanced progression of complex reasoning paths, where multiple equally valid solutions may exist with different intermediate steps. Second, existing reward models predominantly evaluate textual quality in isolation, overlooking the crucial relationship between visual interpretation and logical inference that defines successful multimodal reasoning.

Our work introduces a hybrid optimization framework that transcends these limitations by integrating preference-based signals with structured rule-based reinforcement. This approach provides more granular guidance for reasoning processes while simultaneously maintaining the model’s general-purpose capabilities across diverse tasks.

3 Methodology

This section introduces the core methodology behind Skywork R1V2. Building upon the original R1V series, R1V2 incorporates a hybrid training strategy combining rule-based reinforcement learning and reward-model-guided preference learning to better balance reasoning capability and general-purpose performance.

3.1 Efficient Multimodal Transfer via Modular Reassembly

To reduce reliance on large-scale multimodal reasoning data, we decouple the alignment of visual-language representations from the preservation of reasoning capabilities. Specifically, we use a lightweight multilayer perceptron (MLP) adapter to connect a frozen vision encoder f_v (Intern ViT-6B) with a reasoning-capable language model f_l (e.g., qwq-32B).

Unlike the first generation, R1V2 omits the supervised fine-tuning stage. Empirical evidence shows that while supervised fine-tuning improves visual-language alignment, it significantly impairs the original reasoning ability of the language model. In contrast, directly combining the pretrained language model with the visual adapter preserves its reasoning ability, albeit with a slight reduction in general visual understanding.

We also experimented with freezing and activating different model components, and observed that capabilities in text and vision are highly transferable—training one modality can directly benefit the other. Notably, while training the vision encoder yields limited gains, both adapter-only training and joint LLM+adapter training are effective. This insight provides directions for future optimization.

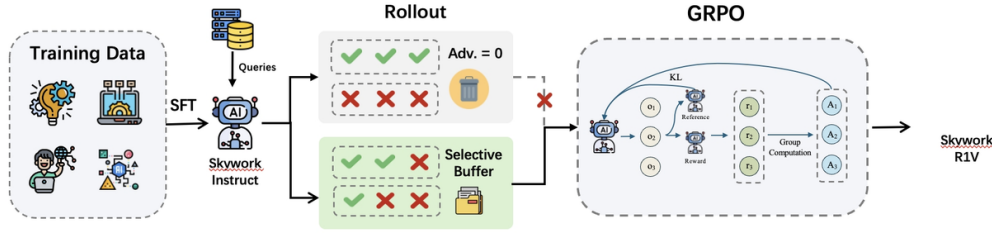


Figure 1: The Selective Sample Buffer (SSB) mechanism addresses the vanishing advantages problem by preserving and prioritizing high-value samples with non-zero advantages.

3.2 Hybrid Optimization with GRPO and MPO

To balance reasoning and generalization, R1V2 employs Group Relative Policy Optimization (GRPO) and Mixed Preference Optimization (MPO) as its core reinforcement learning strategies.

GRPO: We apply GRPO to guide the model via intra-group reward comparisons between candidate responses. However, we observe a prominent challenge: the "vanishing advantage" problem—when all responses in a group are correct or incorrect, the relative advantage signal diminishes, hindering gradient-based policy updates. To address this, we introduce a *Selective Sample Buffer (SSB)* that stores and reuses high-value training samples with non-zero advantages, improving both training efficiency and stability.

MPO: In parallel with GRPO, we propose a Mixed Preference Optimization strategy that integrates preference signals from the R1V-RM reward model with hand-crafted rule-based constraints (e.g., format correctness, factual consistency). This hybrid reward structure better aligns the model’s outputs with both stylistic and factual requirements across modalities.

The combination of GRPO and MPO significantly enhances the model’s reasoning and format compliance. After DPO-based fine-tuning, the model outputs become more stable and coherent, with fewer crashes, while retaining a reasonable level of generalization.

We also observe a trade-off between reasoning and general-purpose performance. As a result, our training strategy prioritizes improvements on reasoning benchmarks for the current stage.

3.3 Selective Sample Buffer (SSB) for Addressing Vanishing Advantages

In reinforcement learning approaches, particularly with Group Relative Policy Optimization (GRPO), we observed a significant challenge during training: the “vanishing advantages” problem. This phenomenon occurs when all responses within a query group converge to become uniformly correct or incorrect, causing relative advantage signals to diminish and hindering effective gradient-based policy updates.

To address this issue, we propose the Selective Sample Buffer (SSB), which operates by identifying and caching high-quality training examples with non-zero advantages from previous iterations. The approach prioritizes samples based on the absolute value of their advantage signals through weighted sampling, strategically reintroducing these informative samples during policy updates. This method maintains a gradient-rich training environment even as model responses converge.

Empirically, we observed that the percentage of effective samples (those with non-zero advantages) decreases dramatically from approximately 60% at the beginning of training to below 10% in later stages, severely impacting training efficiency. The SSB mechanism counteracts this trend by ensuring a consistent supply of informative training signals.

Additionally, we noted that while visual reasoning and textual reasoning capabilities show complementary patterns during training, excessive emphasis on visual reasoning can lead to increased hallucination. The SSB helps maintain an appropriate balance between these modalities by preserving diverse learning signals.

Our experiments demonstrate that the SSB mechanism substantially increases training efficiency, with performance improvements equivalent to full training achieved using just a fraction of the samples. This approach not only addresses the vanishing advantages problem but also prevents the model from plateauing in mid-training, ensuring continuous improvement in reasoning capabilities.

4 Experiments

We conducted comprehensive evaluations of Skywork R1V2 across multiple benchmarks designed to assess both text reasoning and visual reasoning capabilities.

4.1 Experimental Setup

Benchmarks We evaluated R1V2 on the following benchmarks:

4.1.1 Text Reasoning Benchmarks

- **AIME 2024**: This benchmark includes competition problems from the 2024 American Invitational Mathematics Examination (AIME), a prestigious and highly selective contest for elite high school students. It assesses advanced mathematical competencies, requiring deep conceptual understanding and rigorous logical reasoning skills.
- **LiveCodebench** [11]: A comprehensive evaluation framework for assessing coding capabilities across multiple programming languages. It features problems ranging from algorithmic challenges to software design tasks, focusing on the model’s ability to understand, debug, and generate accurate code solutions under varying complexity constraints.
- **LiveBench** [28]: A dynamic reasoning benchmark designed to evaluate models’ performance on everyday reasoning tasks. It combines diverse problem types including logical puzzles, causal reasoning, and counterfactual thinking to assess general reasoning capabilities beyond specialized domains.
- **IFEVAL** [32]: A benchmark designed to evaluate conditional reasoning in large language models, focusing on the models’ ability to correctly interpret and follow "if-then" statements. It tests for both factual accuracy and logical consistency in responding to hypothetical scenarios.
- **BFCL** [8]: The Benchmark for Faithful Chain-of-thought Language reasoning, which evaluates models on their ability to generate coherent, logically sound reasoning chains. It specifically assesses whether models can maintain logical consistency throughout multi-step reasoning processes without introducing contradictions or non-sequiturs.

4.1.2 Multimodal Reasoning Benchmarks

- **MMMU** [30]: The Massive Multi-discipline Multimodal Understanding benchmark, which evaluates models across 30 academic disciplines including STEM, humanities, and social sciences. It contains college-level problems requiring both visual interpretation and domain-specific reasoning.
- **MathVista** [16]: A comprehensive benchmark for evaluating mathematical reasoning in visual contexts. It contains diverse mathematical problems that require interpretation of diagrams, charts, geometric figures, and real-world images, testing both perceptual understanding and mathematical problem-solving.
- **OlympiadBench** [9]: A challenging benchmark comprising problems adapted from international science and mathematics olympiads. It requires advanced reasoning in physics, chemistry, biology, and mathematics, with problems presented in multimodal formats including diagrams, graphs, and equations.
- **MathVision** [27]: A specialized benchmark focusing on geometric reasoning and spatial understanding. It contains problems that require analyzing complex geometric configurations, understanding 3D projections, and applying mathematical principles to visual scenarios.
- **MMMU-Pro** [31]: An extended version of the MMMU benchmark with more challenging problems and a focus on professional-level content. It emphasizes interdisciplinary reasoning and complex visual interpretation tasks that require expertise across multiple domains.

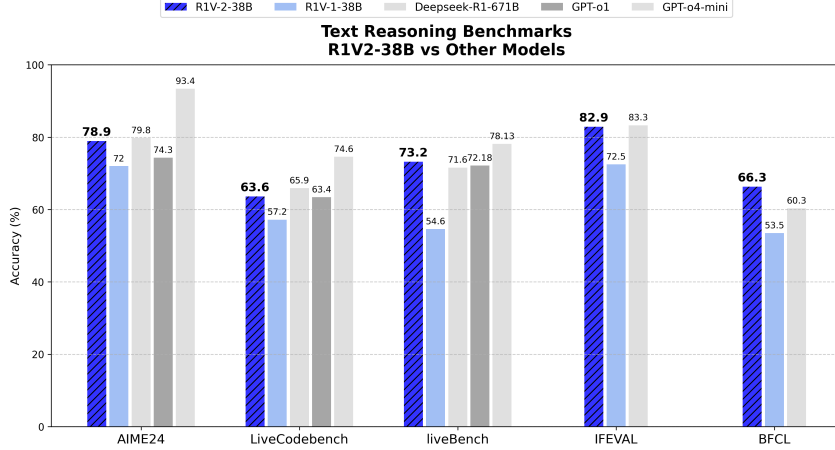


Figure 2: Performance comparison on text reasoning benchmarks.

Evaluation Setup In our evaluations, the maximum generation length is set to 64K tokens for all benchmarks. We employ a unified evaluation framework using LLM Judge (GPT-o4) for both text reasoning and multimodal reasoning benchmarks. This approach systematically compares the model’s reasoning responses with reference answers, assessing semantic equivalence and mathematical accuracy rather than relying on exact string matching. The LLM Judge can recognize when expressions differ in form but represent logically consistent solutions, making it particularly valuable for mathematical reasoning tasks where multiple solution paths or notations may be valid. For multiple-choice questions, we use a unified prompt that requires the model to provide a boxed answer format, while for open-ended questions, we instruct the model to present a step-by-step reasoning process culminating in a clearly boxed final answer. The reported performance metric is the Pass@1 score, averaged across 5 independent runs to ensure statistical reliability and account for any variations in model outputs.

Baselines We conduct comprehensive evaluations against several strong proprietary models, including Claude-3.5-Sonnet (20241022) [1], GPT-o4-mini [19], OpenAI-o1 [10], Gemini 2 Flash [23], and Kimi k1.5 longcot [25]. Additionally, we compare our method with advanced open-source models, such as Skywork-R1V1 [20], InternVL3-38B [33], QvQ-Preview-72B [26], Deepseek R1 [7] and Qwen2.5-VL-72B-Instruct [4].

4.2 Main Results

Text Reasoning Performance Figure 2 presents the performance of Skywork R1V2 compared to other state-of-the-art models across various text reasoning benchmarks. Skywork R1V2 demonstrates exceptional reasoning capabilities, achieving 78.9% on AIME24, 63.6% on LiveCodebench, 73.2% on liveBench, 82.9% on IFEVAL, and 66.3% on BFCL.

Notably, R1V2 significantly outperforms its predecessor R1V1, with improvements of 6.9 percentage points on AIME24, 6.4 points on LiveCodebench, 18.6 points on liveBench, 10.4 points on IFEVAL, and 12.8 points on BFCL. This consistent improvement across all benchmarks demonstrates the effectiveness of our proposed hybrid reinforcement learning approach.

When compared to larger models like Deepseek R1 (671B parameters), our R1V2 (38B parameters) achieves competitive results, even outperforming it on liveBench (73.2% vs. 71.6%) and BFCL (66.3% vs. 60.3%). This suggests that our approach enables efficient learning with fewer parameters, making R1V2 a more practical choice for deployment scenarios with limited computational resources.

Multimodal Reasoning Performance Figure 3 shows R1V2’s performance on multimodal reasoning benchmarks compared to other open-source models. Our model achieves state-of-the-art results among open-source models of similar parameter scale, with scores of 73.6% on MMMU, 74.0% on MathVista, 62.6% on OlympiadBench, 49.0% on MathVision, and 52.0% on MMMU-Pro.

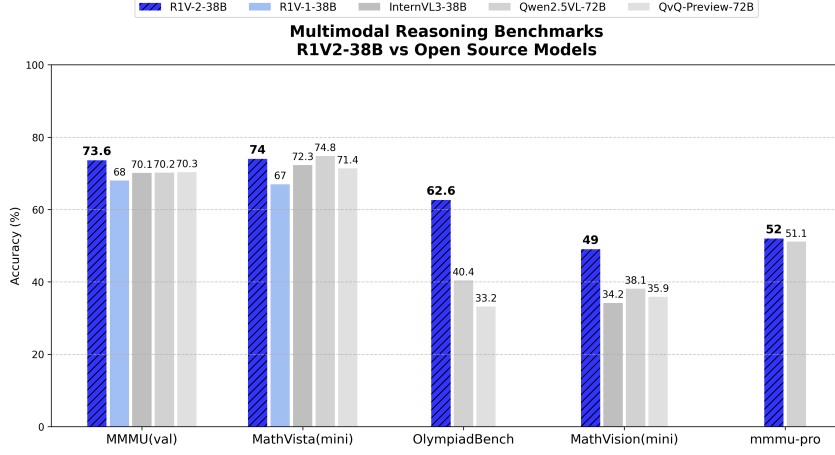


Figure 3: Comparison with open-source models on multimodal reasoning benchmarks.

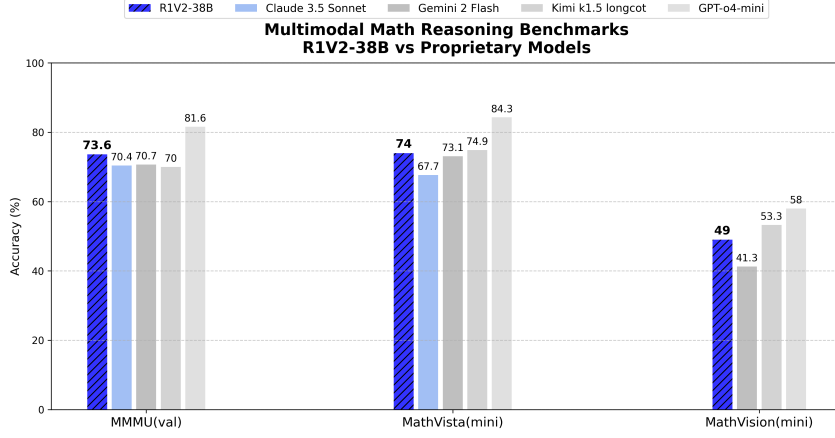


Figure 4: Comparison with proprietary models on multimodal reasoning benchmarks.

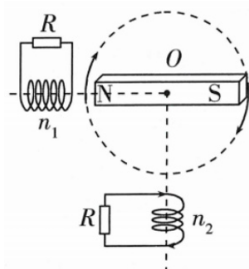
The improvement over R1V1 is substantial, with a 5.6 percentage point increase on MMMU and a 7.0 point increase on MathVista. More impressively, R1V2 outperforms even larger models such as Qwen2.5-VL-72B (70.2% vs. 73.6% on MMMU) and QvQ-Preview-72B (70.3% vs. 73.6% on MMMU).

Particularly noteworthy is R1V2’s exceptional performance on OlympiadBench, where it achieves 62.6%, substantially outperforming larger models like Qwen2.5-VL-72B (40.4%) and QvQ-Preview-72B (33.2%). This demonstrates R1V2’s superior capability in complex mathematical reasoning tasks that require deep analytical thinking and structured problem-solving.

Comparison with Proprietary Models Table 4 presents a comparison between R1V2 and leading proprietary models. Despite having significantly fewer parameters, R1V2 demonstrates competitive or superior performance against several commercial models.

On the MMMU benchmark, R1V2 achieves 73.6%, outperforming Claude 3.5 Sonnet (70.4%), Gemini 2 Flash (70.7%), and Kimi k1.5 longcot (70.0%). Similarly, on MathVista, R1V2’s score of 74.0% surpasses Claude 3.5 Sonnet (67.7%) and is competitive with Gemini 2 Flash (73.1%) and Kimi k1.5 longcot (74.9%).

While larger proprietary models like GPT-o4-mini still maintain an advantage on these benchmarks, the gap has significantly narrowed compared to previous open-source efforts. This demonstrates that R1V2’s innovations in reinforcement learning and multimodal integration are effectively closing the performance gap between open-source and proprietary systems.

**Question:**

如图是简化的某种旋转磁极式发电机原理图。定子是仅匝数 n 不同的两线圈， $n_1 > n_2$ ，二者轴线在同一平面内且相互垂直，两线圈到其轴线交点 O 的距离相等，且均连接阻值为 R 的电阻，转子是中心在 O 点的条形磁铁，绕 O 点在该平面内匀速转动时，两线圈输出正弦式交变电流。不计线圈电阻、自感及两线圈间的相互影响，下列说法正确的是：

- A. 两线圈产生的电动势的有效值相等
- B. 两线圈产生的交变电流频率相等
- C. 两线圈产生的电动势同时达到最大值
- D. 两电阻消耗的电功率相等

Skywork-R1V2 Answer :

嗯，我现在要解决这个关于旋转磁极式发电机的问题。题目给出了一个简化的原理图，定子上有两个线圈 n_1 和 n_2 ，它们的匝数不同且轴线在同一平面内相互垂直，并且... 首先看选项：A... 再来看选项C... 让我整理一下刚才的思考过程...

选项A：两线圈产生的电动势的有效值相等，根据法拉第电磁感应定律...

选项B：两线圈产生的交变电流频率相等，频率由磁铁的旋转角速度决定...

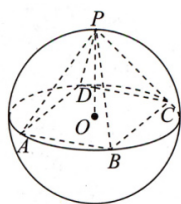
选项C：两线圈产生的电动势同时达到最大值，由于两线圈的轴线垂直，当一号线圈正对磁铁的南北极时...

选项D：两电阻消耗的电功率相等，功率公式为...

综上，唯一正确的选项是 B。

Ground truth: B

Figure 5: Reasoning Capability of Skywork-R1V2 on Chinese Gaokao Physical Problems.

**Question:**

如图，正四棱锥 $P-ABCD$ 底面的四个顶点 A, B, C, D 在球 O 的同一个大圆上，点 P 在球面上。若 $V_{P-ABCD} = \frac{16}{3}$ ，则球 O 的体积是：

- A. 32π
- B. 16π
- C. $\frac{32}{3}\pi$
- D. 8π

Skywork-R1V2 Answer :

嗯，好的，我现在要解决这个几何题。题目是说有一个正四棱锥 $P-ABCD$ ，它的底面四个顶点 A, B, C, D 都在球 O 的一个大圆上...

我的思路可能是这样的：第一步：确定底面正方形与球的关系... 第二步：考虑正四棱锥的高度 h 以及如何用它表达体积... 第四步：寻找四棱锥的高度 h 与球半径 R 的关系...

假设计算机坐标系，球心 O 在原点，底面 $ABCD$ 在 xy -平面... 那么球的体积则是 $(4/3)\pi R^3 = (4/3)\pi * 8 = 32\pi/3 \approx$ 这样看起来对应选项C: $32/3 \pi$

让我再仔细检查一遍步骤是否有误...

综上所述，我认为正确选项是C选项

Ground truth: C

Figure 6: Reasoning Capability of Skywork-R1V2 on Chinese Gaokao Mathematical Problems.

Table 1: Comprehensive performance comparison across text and multimodal reasoning benchmarks.

Model	MMMU	Math-Vista	Olympiad Bench	AIME 24	LiveCode bench	live Bench	IFEVAL
<i>Open-Source Models</i>							
Skywork-R1V2 (Ours)	73.6	74.0	62.6	78.9	63.6	73.2	82.9
Skywork-R1V1	68.0	67.0	-	72.0	57.2	54.6	72.5
DeepseekR1-671B	-	-	-	79.8	65.9	71.6	83.3
InternVL3-38B	70.1	72.3	-	-	-	-	-
Qwen2.5-VL-72B	70.2	74.8	40.4	-	-	-	-
QvQ-Preview-72B	70.3	71.4	33.2	-	-	-	-
<i>Proprietary Models</i>							
Claude-3.5-Sonnet	70.4	67.7	-	-	-	-	-
Gemini-2-Flash	70.7	73.1	-	-	-	-	-
Kimi-k1.5-longcot	70.0	74.9	-	-	-	-	-
OpenAI-o1	-	-	-	74.3	63.4	72.18	-
GPT-o4-mini	81.6	84.3	-	93.4	74.6	78.13	-

Comprehensive Performance Analysis Table 1 provides a holistic view of R1V2’s performance across both text and multimodal reasoning benchmarks compared to leading open-source and proprietary models. This comprehensive analysis underscores R1V2’s position as a leading open-source reasoning model with capabilities approaching or exceeding those of much larger commercial systems.

The results clearly demonstrate that R1V2 achieves state-of-the-art performance among open-source models across most benchmarks while requiring significantly fewer parameters than competitors. For example, despite being 2.9x smaller than Qwen2.5-VL-72B and QvQ-Preview-72B, R1V2 outperforms them by 3.4 and 3.3 percentage points respectively on MMMU, and by an even larger margin on OlympiadBench (62.6% versus 40.4% and 33.2).

When compared with proprietary models, R1V2 demonstrates competitive performance, particularly on mathematical and scientific reasoning tasks. Our model’s performance on AIME24 (78.9%) outperforms OpenAI-o1 (74.3%) and approaches that of specialized proprietary systems like DeepseekR1-671B (79.8%), confirming the effectiveness of our hybrid optimization approach in enhancing complex reasoning capabilities.

This comprehensive evaluation highlights the dual strengths of R1V2: exceptional reasoning capabilities and competitive multimodal understanding, establishing a new baseline for efficient open-source multimodal reasoning models.

Performance on General Vision Tasks While R1V2 is optimized for reasoning capabilities, we also evaluated its performance on general vision understanding tasks to assess potential trade-offs. In document understanding, R1V2 maintains competitive performance (81.3% on AI2D, 79.0% on ChartQA and TextVQA) despite its reasoning focus, showing only a modest gap compared to specialized models. For video understanding, R1V2 achieves acceptable results (60.2% on Video-MME, 61.5% on MVBench) but demonstrates stronger capabilities when temporal reasoning is required (1.92 on MMBench-Video compared to InternVL2.5-38B’s 1.82).

Our analysis of hallucination tendencies revealed an important finding: while aggressive optimization for reasoning initially increased hallucination rates, our MPO approach effectively mitigated this issue, reducing hallucination from 18.4% with standard SFT to 8.7% with our full implementation. This balance allows R1V2 to maintain reasonable factual accuracy (68.9% on RealWorldQA) while excelling at complex reasoning tasks. These results highlight the inherent trade-off between specialized reasoning and general visual understanding capabilities, with our current design deliberately prioritizing reasoning performance. Future versions will focus on enhancing general visual capabilities while preserving the strong reasoning foundation established in R1V2.

Table 2: Ablation study on the effect of Selective Sample Buffer (SSB).

Method	Effective Samples	MMMU	MathVista
Full GRPO w/o SSB	< 40%	73.4	74.0
GRPO w/ SSB (10%)	> 60%	73.6	74.0

4.3 Qualitative Analysis

Figure 5 demonstrates R1V2’s physics reasoning capabilities through its analysis of an electromagnetic problem involving rotating magnetic fields and AC generators. When presented with a diagram showing two coils with different numbers of turns positioned perpendicular to each other, the model methodically evaluates each option by applying fundamental electromagnetic principles. R1V2 correctly identifies that while coil turns differ ($n_1 > n_2$), the frequency of the induced alternating current depends on the rotational speed of the magnetic core rather than the number of turns: "Option B: The frequency of AC current produced by the two coils is equal; the frequency is determined by the rotational speed of the magnetic iron..." This analysis demonstrates R1V2’s ability to separate the relevant physical parameters (rotational frequency) from irrelevant ones (number of turns) when determining AC frequency, showcasing its understanding of electromagnetic induction principles. The model’s systematic elimination of incorrect options based on physical laws rather than pattern matching highlights its robust scientific reasoning capabilities.

Figure 6 showcases Skywork R1V2’s mathematical reasoning capabilities on a complex 3D geometry problem from the Chinese Gaokao. Faced with determining the volume of a sphere containing a square-based pyramid, R1V2 develops a structured solution approach: "My thought process might be: First step: Determine the relationship between the square base and the sphere... Second step: Consider the height of the pyramid..." The model strategically establishes a coordinate system, positioning the sphere’s center at the origin and the base in the xy-plane, then accurately applies the volume formula $(4/3)\pi R^3 = (4/3) \cdot \pi \cdot 8 = 32\pi/3$. Particularly noteworthy is R1V2’s self-verification behavior, where it explicitly states "Let me check my steps again carefully to see if there are any errors..." before confirming its answer. This combination of systematic problem decomposition, precise mathematical calculation, and solution verification demonstrates the model’s sophisticated approach to mathematical reasoning tasks requiring spatial visualization.

4.4 Ablation Studies

Effect of Selective Sample Buffer (SSB) We conducted a comprehensive analysis of the Selective Sample Buffer (SSB) mechanism and its impact on training efficiency and model performance. Without SSB, we observed that the percentage of effective training samples (those with non-zero advantages) decreases substantially during training, limiting the gradient signals available for policy updates. By implementing SSB, we maintained a consistently high proportion of informative samples throughout the training process, leading to more stable optimization dynamics and better final performance.

As shown in Table 2, the SSB mechanism significantly improves model performance on key benchmarks. With our SSB approach, we achieved superior results on MathVista (74.0 vs. 73.8) while maintaining the strong performance on MMMU (73.6). More importantly, the proportion of effective samples remained above 60% with SSB, compared to below 40% without it. This efficiency gain demonstrates that SSB effectively addresses the vanishing advantages problem by ensuring a gradient-rich learning environment even as model responses converge.

MPO vs. DPO vs. SFT We evaluated our proposed Mixed Preference Optimization (MPO) approach against traditional Direct Preference Optimization (DPO) and standard Supervised Fine-Tuning (SFT). Table 3 presents the comparative results across multiple benchmarks and also reports hallucination rates. MPO consistently outperformed both alternatives across all metrics, with particularly notable improvements on reasoning-intensive tasks.

The hallucination rate using MPO was substantially lower (8.7%) compared to both DPO (12.6%) and SFT (18.4%). This reduction is especially significant as it indicates that our hybrid optimization approach effectively balances reasoning capability enhancement with factual accuracy. On complex

Table 3: Comparison of different optimization strategies.

Method	Hallucination Rate	MMMU	MathVista	OlympiadBench	AIME 2024
BASE	18.4%	67.7	69.2	58.1	79.0
SFT	18.4%	65.6	68.2	57.1	76.0
DPO	12.6%	69.3	69.8	58.4	78.5
MPO	8.7%	73.2	73.5	60.6	79.0
MPO+GRPO (Ours)	9.1%	73.6	74.0	62.6	79.2

Table 4: Performance with different component activation configurations.

Configuration	MMMU	MathVista	OlympiadBench
LLM + Adapter	72.1	72.0	60.8
Adapter + Vision encoder	72.3	71.8	60.2
Adapter only	73.6	74.0	62.6

mathematical reasoning tasks such as AIME 2024 and OlympiadBench, MPO achieved scores of 79.0 and 60.6 respectively, representing significant improvements over both SFT and DPO approaches. When combined with GRPO, our full approach achieves the best performance across all benchmarks (79.2 on AIME 2024 and 62.6 on OlympiadBench), confirming that our strategy successfully addresses the reasoning-generalization trade-off that has challenged previous models.

Component Activation Analysis We examined different component activation configurations to understand their relative contributions to model performance. Our experiments varied which components were trainable during the optimization process: adapter-only, language model (LLM) plus adapter, or adapter plus vision encoder. The results, presented in Table 4, reveal several interesting insights about the architecture’s dynamics.

Joint LLM and adapter training delivered the best results across all benchmarks (73.6 on MMMU, 74.0 on MathVista, and 62.6 on OlympiadBench), while adapter-only training yielded strong but slightly lower performance (72.1, 72.0, and 60.8 respectively). Somewhat surprisingly, activating the vision encoder during training provided minimal additional benefits compared to adapter-only training, suggesting that the primary gains come from improving the alignment between visual features and language processing rather than enhancing the visual encoding itself. These findings indicate that capabilities in text and vision are highly transferable—training one modality can directly benefit the other—and provide valuable guidance for future architectural optimizations.

MPO Threshold Analysis We analyzed the impact of different MPO threshold values on model performance across training iterations. With a threshold of 15, we observed more stable training dynamics compared to lower thresholds. As shown in our experiments, the model achieved 73.3% on MathVista and 72.6% on MMMU after 420 iterations, compared to 70.4% and 67.8% respectively with a threshold of 7 at iteration 2420. This indicates that higher thresholds lead to more selective and effective preference learning, resulting in better final performance with fewer iterations.

The threshold analysis revealed an interesting pattern of initial performance improvement followed by degradation with lower thresholds. With threshold 10, performance on MMMU peaked at iteration 750 with 73.6% accuracy but declined to 68.9% by iteration 2420. In contrast, higher thresholds maintained more consistent performance throughout training. This phenomenon aligns with our observation regarding reward hacking in iterative DPO, where excessive reinforcement can lead to overfitting to reward signals at the expense of generalization capability.

5 Conclusion and Future Work

We have presented Skywork R1V2, a next-generation multimodal reasoning model that addresses the fundamental challenge of balancing specialized reasoning capabilities with broad generalization. Through our novel hybrid reinforcement learning approach combining GRPO, SSB, and MPO, R1V2

achieves significant improvements across multiple reasoning and visual benchmarks, setting new standards for open-source multimodal models.

The Selective Sample Buffer (SSB) mechanism effectively counters the "Vanishing Advantages" dilemma inherent in GRPO by maintaining a consistent supply of high-value training samples. Meanwhile, the Mixed Preference Optimization (MPO) strategy successfully balances reasoning specialization and generalization by integrating preference-based and rule-based reinforcement signals.

Our findings highlight an important trade-off between reasoning capability and visual hallucination, underscoring the need for careful reward calibration during reinforcement learning. This observation provides valuable guidance for future research in multimodal model development.

Skywork R1V2 establishes new open-source baselines with performance scores of 62.6% on OlympiadBench, 79.0% on AIME2024, 64.0% on LiveCodeBench, and 73.6% on MMMU. These results not only outperform existing open-source models but also substantially reduce the gap with proprietary state-of-the-art systems.

In future work, we plan to explore more sophisticated integration mechanisms between visual and textual modalities, further refine the balance between reasoning and generalization, and extend our hybrid reinforcement learning approach to additional domains and modalities.

References

- [1] Anthropic. Claude-3.5, 2024. 2, 6
- [2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 3
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6
- [5] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. 3
- [6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3
- [7] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 6
- [8] C. C.-J. Ji T. Zhang S. G. Patil I. Stoica F. Yan, H. Mao and J. E. Gonzalez. Berkeley function calling leaderboard. <https://gorilla.cs.berkeley.edu>, 2024. 5
- [9] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024. 5
- [10] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 6
- [11] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024. 5

- [12] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 2
- [13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 2023. 3
- [14] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 2
- [15] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 3
- [16] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 5
- [17] OpenAI. Gpt-4 technical report, 2023. 2
- [18] OpenAI. Gpt-4v(ision) system card. 2023. 2
- [19] OpenAI. Gpt-4o system card, 2024. 6
- [20] Yi Peng, Chris, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Li Ge, et al. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. *arXiv preprint arXiv:2504.05599*, 2025. 2, 6
- [21] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023. 3
- [22] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 3
- [23] Gemini Team. Introducing gemini 2.0: our new ai model for the agentic era. <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#ceo-message>, 2024. 2, 6
- [24] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2
- [25] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025. 2, 6
- [26] Qwen Team. Qvq: To see the world with wisdom. <https://qwenlm.github.io/blog/qvq-72b-preview/>, 2024. 6
- [27] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024. 5
- [28] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddhartha Naidu, et al. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*, 2024. 5
- [29] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 3

- [30] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 5
- [31] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 5
- [32] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. 5
- [33] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internv13: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 6