

Skywork-R1V3 Technical Report

Multimodal Team
Skywork AI, Kunlun Inc



<https://huggingface.co/Skywork/Skywork-R1V3-38B>

<https://github.com/SkyworkAI/Skywork-R1V3>

Abstract

We introduce Skywork-R1V3, an advanced, open-source vision-language model (VLM) that pioneers a new approach to visual reasoning. Its key innovation lies in effectively transferring reasoning skills from text-only Large Language Models (LLMs) to visual tasks. The strong performance of Skywork-R1V3 primarily stems from our elaborate post-training RL framework, which effectively activates and enhances the model's reasoning ability, without the need for additional continue pre-training. Through this framework, we further uncover the fundamental role of the connector module in achieving robust cross-modal alignment for multimodal reasoning models. In addition, we introduce a unique indicator of reasoning capability, the entropy of critical reasoning tokens, which has proven highly effective for checkpoint selection during RL training. Skywork-R1V3 achieves state-of-the-art results of 76.0% on MMMU. This performance matches entry-level human expert capabilities. Remarkably, our RL-powered post-training approach enables even the 38B parameter model to rival top closed-source VLMs. The implementation successfully transfers mathematical reasoning to other subject-related reasoning tasks. We also include an analysis of curriculum learning and reinforcement finetuning strategies, along with a broader discussion on multimodal reasoning. Skywork-R1V3 represents a significant leap in multimodal reasoning, showcasing RL as a powerful engine for advancing open-source VLM capabilities.

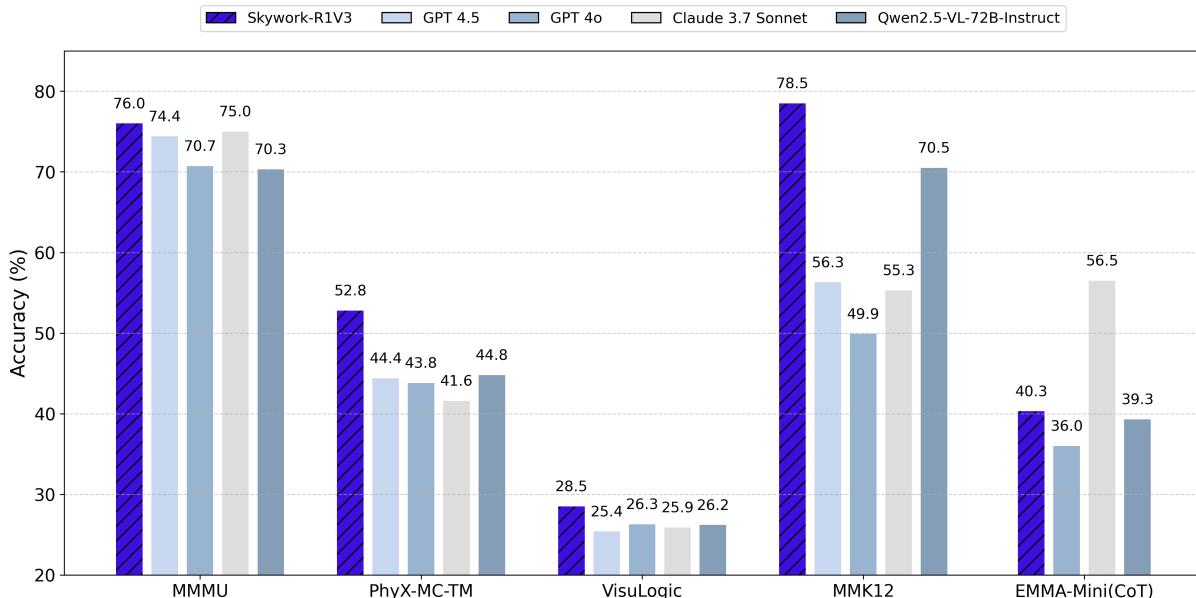


Figure 1: Benchmark performance of Skywork-R1V3.

1 Introduction

Vision-language models (VLMs) have emerged as a foundational paradigm for general-purpose AI, enabling systems to perceive, reason, and act within open-ended virtual and physical environments. By aligning visual and textual modalities within a unified framework, VLMs have driven significant progress across diverse domains, including multimodal reasoning (Team et al., 2025), image editing (Zhou et al., 2025), GUI agents (Nguyen et al., 2024), autonomous driving (Xu et al., 2024), and robotics (Wang et al., 2024a). Beyond research, VLMs are increasingly powering real-world applications in education, healthcare, conversational AI, and wearable technologies (Wang et al., 2024b). Despite this progress, current VLMs still lack human-level generality—particularly in tasks demanding robust spatial reasoning, precise object counting, imaginative visual synthesis, or complex interactive gameplay (Jian et al., 2025). These limitations highlight persistent challenges in VLM development (Peng et al., 2025). Furthermore, the inherent heterogeneity of multimodal data complicates both training and inference, introducing bottlenecks in data pipeline design, distributed training optimization, and standardized evaluation.

This challenge stands in stark contrast to the situation for large language models (LLMs) (Touvron et al., 2023; DeepSeek-AI, 2024). While LLMs benefit from vast, high-quality textual corpora encoding broad human knowledge, VLMs suffer from scarce and uneven vision-language annotations—particularly for perceptually-grounded concepts such as texture, lighting, and spatial relations. (Zhu et al., 2025). Consequently, the gap between closed-source and open-source models remains significantly wider for VLMs than for LLMs. In the LLM domain, strong open-source models like Deepseek R1(DeepSeek-AI, 2025) provide high-quality Chain-of-Thought (Wei et al., 2022) reasoning, narrowing the capability gap. However, the VLM landscape faces a much steeper challenge in closing this divide. While proprietary models like Gemini Pro 2.5 (DEEPMIND, 2025), OpenAI’s O-series (Jaech et al., 2024), and Seed-VL(Guo et al., 2025) demonstrate powerful visual reasoning across multiple domains, open-source alternatives such as QwenVL (Wang et al., 2024b), InternVL (Zhu et al., 2025), and LLaVA(Liu et al., 2023) struggle to match their performance. Moreover, these open VLMs often lack comparable reasoning abilities and structured thinking patterns, further widening the gap.

In this report, we take a significant step forward in advancing open-source VLMs, particularly in complex visual reasoning tasks. Additionally, we provide comprehensive training details on post-training methodologies for VLMs based on open-source frameworks. Key points we address include:

Cold Start Finetuning and RL can incentivize the main reasoning ability of the model. We construct our cold-start dataset using early version Skywork R1V2 (Wang et al., 2025b) and conduct SFT and RL experiments primarily with a instructed VLM. This strategy enables effective transfer of reasoning patterns from a language reasoning model to a vision-language model (Section 3).

Critical token entropy indicates reasoning ability. In the RL training stage, we distinguish the models with genuine reasoning capabilities from those merely mimicking reasoning patterns through our novel metric, entropy of critical token. Specifically, we monitor the entropy values at critical reasoning initiation points. This metric strongly correlates with the actual reasoning performance in validation sets, providing an efficient method to identify high-quality checkpoints during RL training. (Section 5.1).

Connector plays a central role in cross-modal alignment. Prior work (Guo et al., 2025) has primarily focused on achieving vision–text alignment during the pre-training phase and has demonstrated the effectiveness of the connector. In this study, we further identify the connector in VLMs as a critical component for maintaining cross-modal alignment during the reinforcement learning stage (Section 5.2). Furthermore, we find that performing connector-only tuning after RL serves as an effective strategy to re-balance the model’s knowledge distribution without compromising its reasoning ability (Section 5.6).

Beyond the core content, this report includes several additional topics: the application of curriculum learning to reinforcement learning (Section 5.3), and reproductions of reinforcement learning tricks (Yu et al., 2025) such as Clip-Higher and Dynamic Sampling (Section 5.5). We also explore the behavior of the visual reasoning model: the broader challenge of balancing memorization and generalization (Section 6.1), explore fast and slow thinking in inference (Section 6.2), and investigate the effectiveness of thinking token budgets (Section 6.3) and hallucination in the reasoning transfer pipeline (Section 6.4). We hope these areas spark further investigation and engagement from the research community.

2 Data Preparation

LongCoT Data for Cold Start We constructed a cold-start dataset consisting of approximately 20K instances collected from science-oriented practice questions of Chinese high-school difficulty, administered prior to 2024. The dataset covers four major disciplines: physics, chemistry, biology, and mathematics. It includes both multiple-choice diagnostic items and open-ended reasoning tasks.

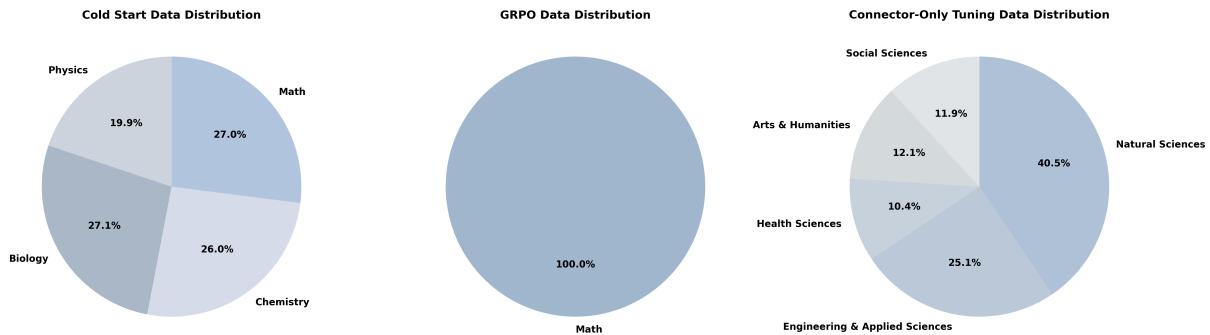


Figure 2: Data distribution across the three training stages.

To create the corresponding reasoning processes, we leveraged the Skywork-R1V2 model to generate step-by-step thought processes for each question. We then applied a rejection sampling strategy to select only those instances where the model’s final answer matched the ground-truth answer. This procedure resulted in a filtered dataset containing approximately 12k high-quality examples with reasoning chains.

Data for RL For the RL stage, we curated 15K high-quality multimodal **math** data with K12-level difficulty. The dataset consists entirely of multiple-choice and fill-in-the-blank items, where each instance is represented as a question q paired with its corresponding answer a , and no explicit reasoning steps are included.

Data for Connector-only Tuning To ensure domain diversity, we selected 10K examples from 20 distinct domains, including:

- **Natural Sciences:** Mathematics, Physics, Chemistry, Biology
- **Engineering & Applied Sciences:** Engineering, Agriculture, Environmental Science, Computer Science
- **Health Sciences:** Medicine, Pharmacy
- **Arts & Humanities:** Music, Literature, Philosophy, History
- **Social Sciences:** Economics, Psychology, Political Science, Sociology, Education, Law

The data distribution across the three training stages is illustrated in the Figure 2.

3 Post-Training Recipes

Our post-training recipes mainly consist of reward function design, cold start finetuning, reinforcement finetuning and connector-only finetuning. We will discuss details in this section.

3.1 Reward Function Design

Our reward function typically consists of a format reward and an accuracy reward. The format reward is defined as:

$$R_{\text{format}}(y) = \mathbb{1}(c = 1) \quad (1)$$

where $c = 1$ indicates that the generated response adheres to the specified chat template, which follows the structure “<think>...<think>...”.

The accuracy reward is the primary objective of our Reinforcement Learning (RL) process, which we aim to maximize:

$$R_{\text{accuracy}}(\hat{a}, a) = \begin{cases} 1, & \text{if } R_\theta(\hat{a}, a) = 1 \text{ or } R_{\text{rule}}(\hat{a}, a) = 1 \\ 0, & \text{else} \end{cases} \quad (2)$$

Here, $R_\theta(\hat{a}, a)$ likely represents an accuracy score derived from a language model-based verifier (eg. Qwen2.5-Instruct 32B), and $R_{\text{rule}}(\hat{a}, a)$ represents an accuracy score based on predefined rules.

Therefore, our final reward is defined as:

$$R = \epsilon R_{\text{accuracy}} + (1 - \epsilon) R_{\text{format}} \quad (3)$$

Empirically, we set ϵ to 0.8 in this report, emphasizing accuracy as the primary reward component.

3.2 Cold Start Finetuning

To initiate training, we collected thousands of cold-start samples from an early internal version of Skywork-R1V2. While demonstrating strong reasoning capabilities, this model suffered from issues like poor readability and language mixing. To enhance reasoning clarity and enable open sharing, we employed the Skywork-VL-Reward (Wang et al., 2025d) alongside GPT-4o to filter rambling and overly lengthy samples, resulting in a refined cold-start dataset.

Significantly, Skywork-R1V2 is created by directly stitching InternViT-6B-448px-V2.5 (Zhu et al., 2025) and QwQ-32B (Team, 2025). Its visual reasoning capabilities originate from textual reasoning, illustrating a successful transfer of reasoning skills and the potential for cross-modal reasoning inheritance.

Building on this, we integrate R1V2’s reasoning style and validated visual reasoning traces into the SFT stage using InternVL-38B (Zhu et al., 2025) as the base model. Following the cold-start phase, the model naturally adopted the “think before answering” (R1-style) approach—an inherent capability present from the LLM’s pretraining stage. By preserving the valid visual reasoning distribution identified in R1V2, we further enhance the model’s reasoning proficiency.

3.3 Reinforcement Finetuning

The RL objective is to learn a reasoning policy π_θ that maximizes the expected reward across prompts sampled from distribution \mathcal{D} , i.e.:

$$\max_{\pi} J(\pi) := \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi(\cdot|x)} [r(x, y)]. \quad (4)$$

PPO Proximal Policy Optimization (PPO) (Schulman et al., 2017) is a state-of-the-art policy gradient algorithm that ensures training stability by constraining the size of policy updates. It optimizes a clipped surrogate objective function, which is a lower bound on the policy performance improvement. The core objective for PPO is given by:

$$\mathcal{L}^{\text{PPO}}(\theta) = \hat{\mathbb{E}}_t [\min (\rho^t(\theta) \hat{A}^t, \text{clip}(\rho^t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}^t)]. \quad (5)$$

Here, $\rho^t(\theta) = \frac{\pi_\theta(a^t|s^t)}{\pi_{\text{ref}}(a^t|s^t)}$ is the importance sampling ratio between the new policy π_θ and the old policy π_k used for data collection. \hat{A}^t is an estimator of the advantage function at timestep t , and ϵ is a hyperparameter that defines the clipping range.

GRPO Group-normalized Reward Policy Optimization (GRPO) (Shao et al., 2024) is a variant of PPO tailored for learning from sparse, binary rewards ($r(x, y) \in \{0, 1\}$), often provided by a verifier. The key innovation of GRPO lies in its advantage estimation method.

Instead of using the raw binary reward, GRPO first samples M independent and identically distributed candidate traces $\{y_{i,1}, \dots, y_{i,M}\}$ from the current policy π_k for each input x_i . The token-level advantage A_{ij}^t for each trace y_{ij} is then estimated by normalizing its final reward within the sampled group:

$$A_{ij}^t := \frac{r(x_i, y_{ij}) - \text{mean}(r(x_i, y_{i,1}), \dots, r(x_i, y_{i,M}))}{\text{std}(r(x_i, y_{i,1}), \dots, r(x_i, y_{i,M}))}, \quad (6)$$

This group normalization converts the sparse binary signal into a dense, continuous advantage estimate, providing a richer learning signal for the policy.

The final GRPO policy loss incorporates this group-normalized advantage into the PPO objective, adding a KL-divergence penalty that regularizes the policy π_θ towards a reference policy π_{ref} :

$$\begin{aligned} \mathcal{L}^{\text{GRPO}}(\theta) = -\hat{\mathbb{E}}_{x_i, \{y_{ij}\}} & \left[\frac{1}{M} \sum_{j=1}^M \frac{1}{|y_{ij}|} \sum_{t=0}^{|y_{ij}|-1} \left(\min \left(\rho_{ij}^t(\theta) A_{ij}^t, \text{clip} \left(\rho_{ij}^t(\theta), 1 - \epsilon, 1 + \epsilon \right) A_{ij}^t \right) \right. \right. \\ & \left. \left. - \beta D_{ij}^t(\theta) \right) \right]. \end{aligned} \quad (7)$$

In this equation, $\rho_{ij}^t(\theta) = \frac{\pi_\theta(a_{ij}^t|s_{ij}^t)}{\pi_{\text{ref}}(a_{ij}^t|s_{ij}^t)}$ is the importance sampling ratio, ϵ is inherited from PPO clipping parameter, β controls the strength of the KL regularization, and $D_{ij}^t(\theta) = \log \frac{\pi_\theta(a_{ij}^t|s_{ij}^t)}{\pi_{\text{ref}}(a_{ij}^t|s_{ij}^t)}$ is the KL-divergence penalty.

RL Implementation Details We use the VERL framework (Sheng et al., 2024) for RL training, progressively increasing the context and output lengths from 4,096 to 8,192. Each episode samples 1,024 rollouts, with 32 verifier-rewarded samples per prompt. Training employs a mini-batch size of 32, 1 gradient step per episode (on-policy finetuning for performance maximization), and a PPO clipping range of 0.2. No KL divergence penalty is applied during long CoT training .

The actor’s learning rate is set to 1e-5, differing from costumed GRPO configurations (discussed in Section 5.4). Besides, we keep all modules trainable during RL—Section 5.2 analyzes this choice.

3.4 Connector-Only Tuning

RL training encourages the model to explore and optimize for reward-related behaviors such as correctness and format coherence, which has proven to be a highly effective approach for enhancing the reasoning abilities of VLM. However, since the training data during the reinforcement learning stage were primarily composed of mathematical problems, the model’s knowledge acquisition exhibited a certain bias: while its reasoning ability improved, its knowledge base became skewed towards the mathematics domain.

To rebalance the knowledge base and enhance the model’s cross-disciplinary reasoning capabilities, we introduced an additional finetuning step targeting the cross-modal connector after the reinforcement learning stage. Specifically, we constructed a high-quality, multi-domain, multimodal dataset of 10K samples (see Section 2) to retrain the connector in a targeted manner, optimizing its ability to integrate knowledge across diverse domains.

This step effectively balanced the model’s knowledge distribution, significantly improving its perception and understanding in non-mathematical fields such as humanities, medicine, and the arts, while maintaining its original strengths in reasoning. As a result, the model achieved stronger generalization and reasoning performance across a wide range of disciplines.

Implementation Details Training was conducted for two epochs using the 10K multi-domain multimodal dataset, with a global batch size of 64 and a learning rate of 1e-5, scheduled using a cosine decay with a 0.03 warmup ratio. The maximum sequence length was set to 16,384 tokens. Early stopping is applied based on validation loss to avoid overfitting.

Table 1: Comparison of models on visual-language benchmarks

Benchmark	Metrics (Avg@5)	Skywork-R1V3-38B	QVQ-72B Preview	Internvl3-78B	Qwen2.5 VL 72B	Claude 3.7 Sonnet	GPT-4o
<i>General</i>							
MMMU ^{val}	Acc.	76.0	70.3	72.2	70.3	75.0	70.7
EMMA ^{mini-cot}	Acc.	40.3	32.0	38.3	39.3	56.5	36.0
MMMU-pro	Acc.	55.4	46.9*	48.6	51.1	50.0	54.5
MMK12	Acc.	78.5	62.7*	67.4*	70.5*	55.3	49.9
MMstar	Acc.	70.6	60.8	72.5	70.8	68.8	65.1
MBench-en-1.1	Acc.	85.7	72.6*	87.7	88.0	82.0	84.3
HallusionBench	Acc.	61.3	55.3*	59.1	55.2	58.3	56.2
<i>Mathematics</i>							
MathVista ^{mini}	Acc.	77.1	71.4	72.2	74.8	66.8	62.9
MathVerse ^{vision-only}	Acc.	59.6	45.1	51.0	57.6	49.9*	49.9
MathVision	Acc.	52.6	35.9	43.1	38.1	58.6	31.2
WeMath ^{strict}	Acc.	56.5	37.7	46.1	50.6	48.9*	50.6
<i>Logic</i>							
Visulogic	Acc.	28.5	23.5*	27.7	26.2	25.9	26.3
LogicVista	Acc.	59.7	53.8	55.9	57.1	60.6*	64.4
MME-reasoning	Acc.	42.8	35.2	32.1	34.1	34.1	30.2
<i>Physics</i>							
PhyX ^{mc-text-minimal}	Acc.	52.8	35.2*	40.5	44.8	41.6	43.8
SeePhy	Acc.	31.5	22.5	19.0*	24.2	34.6	21.9

* indicates results from our evaluation framework.

4 Evaluation

To comprehensively evaluate the reasoning capabilities of Skywork-R1V3, we assess its performance on a diverse suite of publicly available multimodal benchmarks. These benchmarks encompass both

general-purpose understanding tasks and advanced multimodal reasoning tasks, thereby providing a broad and rigorous evaluation setting. Furthermore, we position Skywork-R1V3-38B relative to several mainstream multimodal models on representative public benchmarks to offer a more holistic perspective on its overall performance.

Evaluation Settings For evaluation, we use a generation token budget of 16,384 and a decoding temperature of 1.0. Results are averaged over five runs. We adopt VLMEvalKit (Duan et al., 2024) as the primary evaluation framework, with targeted modifications (i.e. add rule-based check for MMMU (Yue et al., 2024)) to better support long chain-of-thought reasoning models. Additionally, we refine task-specific evaluation protocols to improve consistency and reliability across benchmarks. To promote open research and ensure reproducibility, we plan to release the complete evaluation framework along with all associated prompt sets. Our evaluations compare against a diverse set of representative models, including open-source models (InternVL3-78B (Zhu et al., 2025), Qwen2.5-VL-72B (Wang et al., 2024b) and QVQ-72B Preview (Team, 2024)) and leading proprietary models (Claude 3.7 Sonnet (Anthropic, 2024), GPT-4o (OpenAI, 2024)), to ensure a comprehensive and rigorous benchmarking.

General Benchmarks On general vision-language benchmarks, Skywork-R1V3-38B demonstrates strong and consistent performance, achieving state-of-the-art among open-source models on MMMU (Yue et al., 2024) (76.0%), MMMU-pro (Yue et al., 2025) (55.4%), and MMK12 (Meng et al., 2025) (78.5%). On EMMA (Hao et al., 2025), Skywork-R1V3-38B surpasses larger open-source models such as InternVL3-78B and Qwen2.5-VL-72B. On MMStar (Chen et al., 2024), Skywork-R1V3-38B reaches 70.6%, which is comparable to the best open-source model (InternVL3-78B at 72.5%). Similarly, on MMBench-en-1.1 (Liu et al., 2024), Skywork-R1V3-38B scores 85.7%, ranking among the top performers and closely matching the larger InternVL3-78B (87.7%) and Qwen2.5-VL-72B (88.0%). In HallusionBench (Guan et al., 2024) (which evaluates multimodal hallucination robustness), Skywork-R1V3-38B achieves 61.3%, outperforming other advanced models and demonstrating solid reliability in this challenging scenario.

Mathematics Benchmarks In the mathematics reasoning benchmarks, Skywork-R1V3-38B exhibits robust multi-modal problem-solving abilities. For instance, it achieves 77.1% on MathVista (Lu et al., 2023), 59.6% on MathVerse (Zhang et al., 2024) and 56.5% on WeMath (Qiao et al., 2024) (strict), consistently outperforming all larger open-source models (72B/78B) on these tasks. Skywork-R1V3-38B also surpasses Claude 3.7 Sonnet and GPT-4o on most of these math benchmarks, highlighting the strength of its advanced reasoning capabilities in complex mathematics problems. Skywork-R1V3 achieves a comparable score on the 2025 GAOKAO math exam as depicted in Figure 4, demonstrating strong reasoning and generalization abilities on out-of-distribution (OOD) problems. Evaluated using a human and LLM mixed-judging approach, Skywork-R1V3 scored an impressive 142 out of 150. This performance surpasses several advanced models.

Logic Benchmarks Our Skywork-R1V3-38B continues to perform competitively on logical reasoning tasks, achieving 28.5% on VisuLogic (Xu et al., 2025) and narrowly outperforming the best open competitor (InternVL3-78B at 27.7%). On the LogicVista (Xiao et al., 2024) benchmark, Skywork-R1V3-38B reaches 59.7%, approaching the performance of GPT-4o (64.4%) while surpassing all other open-source models. Furthermore, Skywork-R1V3-38B achieves 42.8% on the MME-Reasoning (Yuan et al., 2025) benchmark, which exceeds the larger open-source model (QVQ-72B Preview, 35.2%) by over 7 points and outperforms GPT-4o (30.2%) by more than 12 points.

Physics Benchmarks On physics-related benchmarks, Skywork-R1V3-38B demonstrates strong generalization. It achieves 52.8% on PhyX (Shen et al., 2025), substantially outperforming competing models (detailed performance is shown in Figure 3). On SeePhy (Xiang et al., 2025), Skywork-R1V3-38B scores 31.5%, which is just below Claude 3.7 Sonnet's 34.6% but higher than other open-source models. These results indicate that our Skywork-R1V3-38B can effectively handle complex physics problems, outperforming much larger open-source models and even approaching the advanced proprietary model's performance.

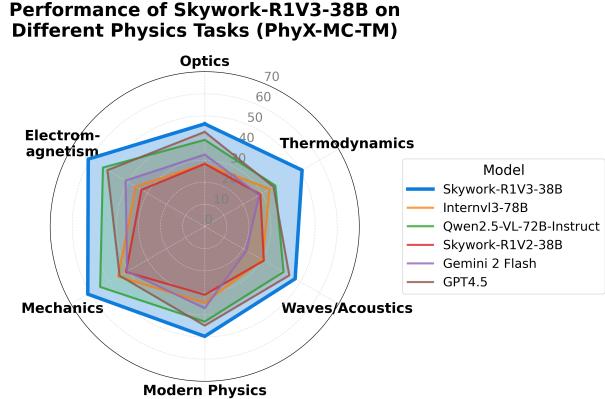


Figure 3: The Performance of Skywork-R1V3-38B on PhyX-MC-Text-Minimal

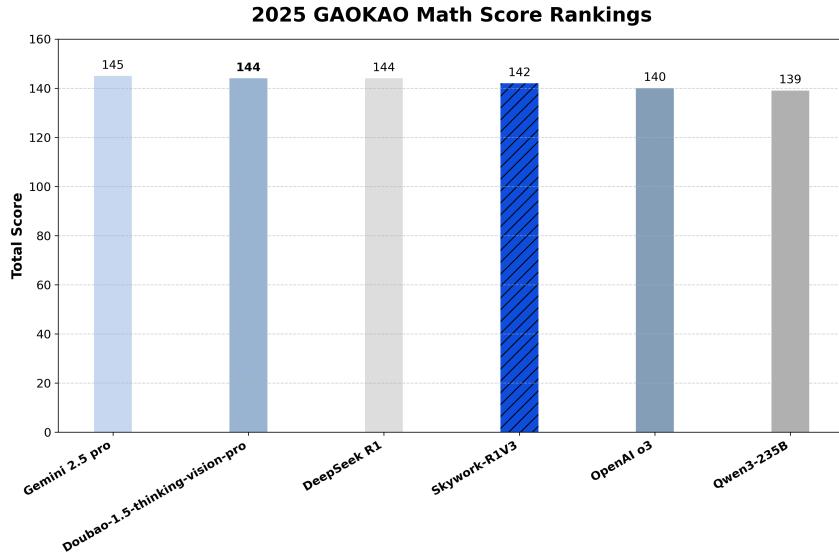


Figure 4: Model Rankings on 2025 GAOKAO Math

5 Empirical Analysis on Reinforcement Learning

This section details our empirical analysis of Reinforcement Learning (RL) applied to Skywork-R1V3. RL demonstrates great potential for enhancing model capabilities. Our findings aim to stimulate further research into equipping VLMs with advanced reasoning capabilities via RL. For this study, we utilized the GRPO algorithm.

We highlight that direct transfer of RL techniques from text-only Large Language Models (LLMs) to VLMs is insufficient due to the distinct challenges posed by visual modality integration. This necessitates the development of VLM-specific RL recipes. Ultimately, this section underscores the critical need for continued exploration to optimize RL for Visual Large Language Models (VLMs) and unlock their full potential in visual language understanding and reasoning.

5.1 Critical Token Entropy Indicates Reasoning Ability

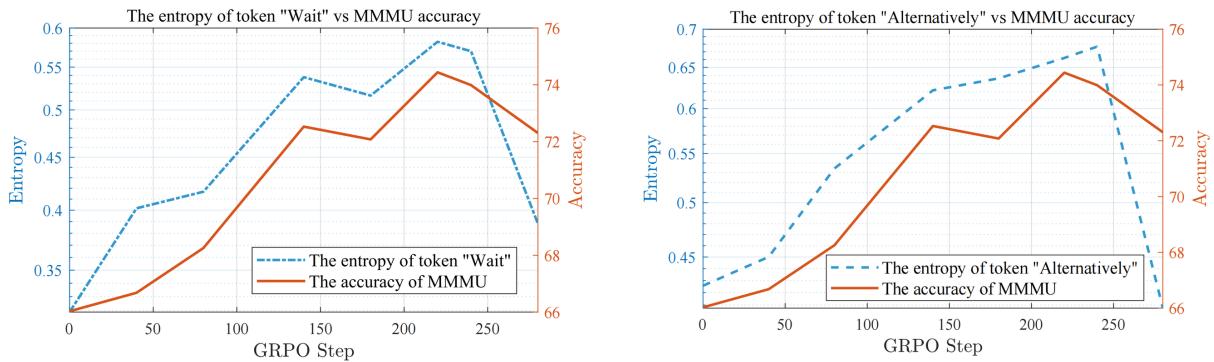


Figure 5: The entropy of critical token vs. MMMU accuracy

During the reinforcement learning training process, we observed that although models fine-tuned with cold start supervised learning can generate responses that appear to follow a reasoning style and achieve high reward values on in-distribution training and validation sets, in most cases, this merely reflects the model repeating existing patterns rather than truly activating generalizable reasoning capabilities. To address this, Skywork-R1V3 introduces a unique metric, which utilizes the entropy of critical tokens to validation. Specifically, we focus on the output entropy values at critical positions where the model begins reasoning (such as when generating tokens like "Wait..." or "Alternatively..." to enter the reasoning process). Models with genuine reasoning capabilities typically exhibit high uncertainty at these positions, demonstrating characteristics of divergent thinking. In contrast, models that merely mimic reasoning

style and tone usually produce low-entropy, deterministic content at these key points.

Based on this insight, we can efficiently identify model checkpoints that have truly learned to reason during reinforcement training while filtering out those that simply "follow the script." As depicted in Figure 5, the level of entropy at these critical points shows a strong correlation with the model's actual reasoning performance on the validation set. This mechanism provides a novel and efficient method for model selection during reinforcement training, ensuring that the final chosen model weights have indeed acquired generalizable reasoning abilities.

5.2 The Connector Module Activation is Vital in RL

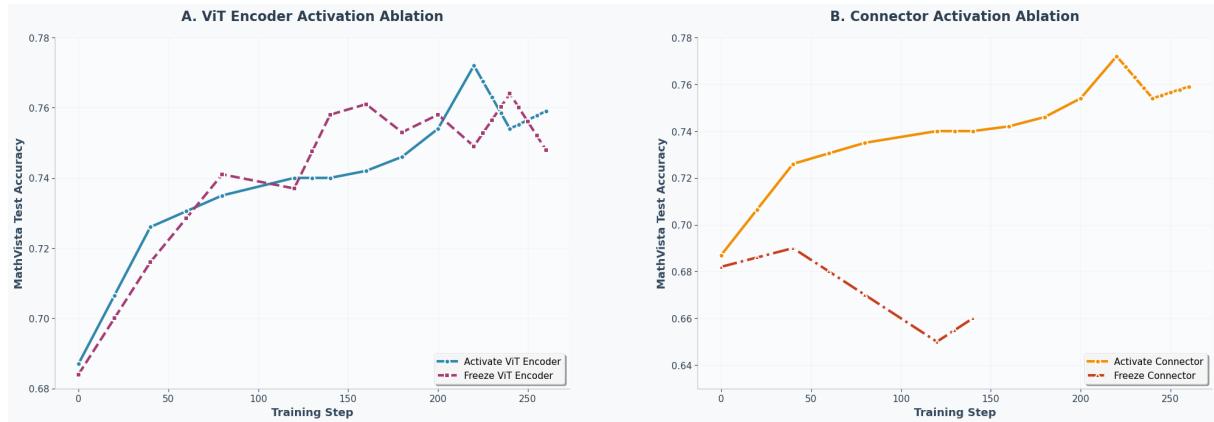


Figure 6: Ablation Studies of Module Activation Impact on MathVista Performance

Here, we investigate primarily the activation of the Skyword-R1V3 module in RL. The core difference between multimodal reasoning reinforcement learning and text-based reasoning reinforcement learning lies in the connector and the visual encoder. During reinforcement learning, we observed that whether the visual encoder is activated (with trainable parameters) has a significant impact on the final performance of R1V3.

Visual Encoder Activate vs. Freeze With the connector remaining trainable, whether or not the visual encoder parameters are updated, the reward curve remains largely consistent (depicted in Figure 6A). This indicates that freezing the visual encoder does not disrupt the normal progression of the training process. However, allowing the visual encoder to participate in the training continues to bring additional performance gains. Even if the visual encoder has been well-pretrained, finetuning it during the reinforcement learning phase for specific cross-modal reasoning tasks remains valuable, as it helps the model capture more task-relevant representational details.

Connector Activate vs. Freeze We found that the trainability of the connector is an absolute prerequisite for stable model learning. Once the connector is frozen or removed during the reinforcement learning phase, the model quickly experiences training failure: the reward curve drops sharply, the unstable gradient norm, and the model output degenerates into meaningless repetitive text, completely losing its reasoning ability (as shown in Figure 6B).

This experiment demonstrates that the connector, as the core bridge integrating visual and language modalities, plays a fundamental role in cross-modal reasoning optimization. It ensures that the representations of the two heterogeneous modalities can be effectively guided, interacted with, and aligned during training, collectively pointing to the latent shared reality in the Platonic Representation Hypothesis (Huh et al., 2024). In contrast, the activation of the visual encoder acts as a supplementary optimization on this stable bridge, enhancing the model's representation accuracy for specific task details, but its role is not as foundational and critical as that of the connector.

5.3 The Distribution Shift in Curriculum Learning Hinder Generalization

We attempted to introduce a staged curriculum learning strategy, aiming to construct a reinforcement training process that progresses from easy to difficult. Specifically, we began by using a medium-difficulty math problem set (such as regular problems covering K12 knowledge, defined as "Normal Problems") for initial training to obtain a model checkpoint at this stage. Then, using this checkpoint as the starting

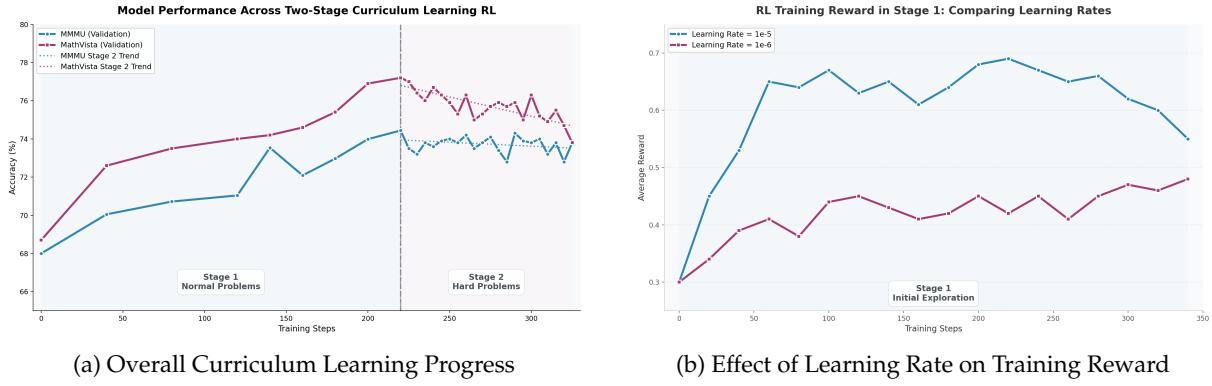


Figure 7: Comparison of Curriculum Learning Aspects. This figure analyzes key aspects of curriculum learning. **(a)** illustrates the significant drop in validation accuracy during Stage 2. **(b)** shows that a higher learning rate can lead to faster learning efficiency but also instability during Stage 1.

point, we switched to a high-difficulty problem set (such as competition questions, defined as "Hard Problems") to continue training, with the goal of achieving progressive enhancement in reasoning ability.

However, the experimental results (Figure 7) indicate that difficulty-based data switching failed to improve generalization, likely due to distribution shift. Specifically, after switching to the Hard Problems for training, although the model's adaptation to complex tasks improved (with stage 2 reward values continuously increasing), its performance on the unified evaluation set showed divergence: the accuracy on Normal problems significantly decreased, and the performance on some logic and physics subtasks fluctuated or stagnated. This phenomenon indicates that the complex skills, special patterns, or high-level strategies learned by the model on the Hard Problems, while helpful for tackling high-difficulty tasks, might conflict with the core reasoning paths relied upon by medium-difficulty problems, ultimately weakening its overall generalization ability.

5.4 Learning Rate Strategy of RL

As depicted in Figure 7b, our initial experiments demonstrate that the learning rate significantly impacts the efficiency of RL, as the higher. This hyperparameter dictates the magnitude of each weight update during optimization. We define a higher learning rate as $1e-5$ and a lower rate as $1e-6$. Although higher rates typically enable a faster initial reward gain in traditional ML (consistent with our Stage 1 results), they introduce substantial risks of instability later in training. Specifically, Stage 1 revealed that $1e-5$ accelerates early reward accumulation but can precipitate model collapse in subsequent phases, manifesting itself as reward diminishment and entropy collapse.

Conversely, Figure 8a and 8b show that in Stage 2, maintaining a learning rate of $1e-5$ after initial convergence led to a critical observation: the model's entropy quickly diverged from its stabilized trend and exhibited a sharp increase in the log metrics. This unexpected surge signifies policy destabilization – the model effectively "unlearns" its acquired behavior due to the overly aggressive updates. The large learning rate prevents sustained finetuning and jeopardizes the stability of the converged policy. Consequently, while $1e-5$ enhances initial learning speed, the significant risk of late-stage instability necessitates careful consideration for efficiency and final performance.

5.5 Additional Reinforcement Learning Trials

In our experimental setup, we also explored two approaches: Clip-Higher and Dynamic Sampling.

Dynamic Sampling In our experiments, dynamic sampling moderately impacts our baseline model. Specifically, it mitigates diminishing training progress, stabilizes the learning curve for extended training steps, yet ultimately fails to improve validation set performance.

Clip-Higher Clip-Higher was primarily intended to mitigate pattern convergence, yet it's not performing as expected. We're observing severe instability and anomalous gradient norms during training. We hypothesize this aligns with the covariance-driven entropy dynamics described by Cui et al. (2025): entropy change is governed by the negative covariance between action log-probabilities and advantage-driven logit updates. Specifically, when high-advantage tokens coincide with low initial probabilities (acting as exploration candidates), their gradient updates exhibit a large positive covariance. Clip-Higher,

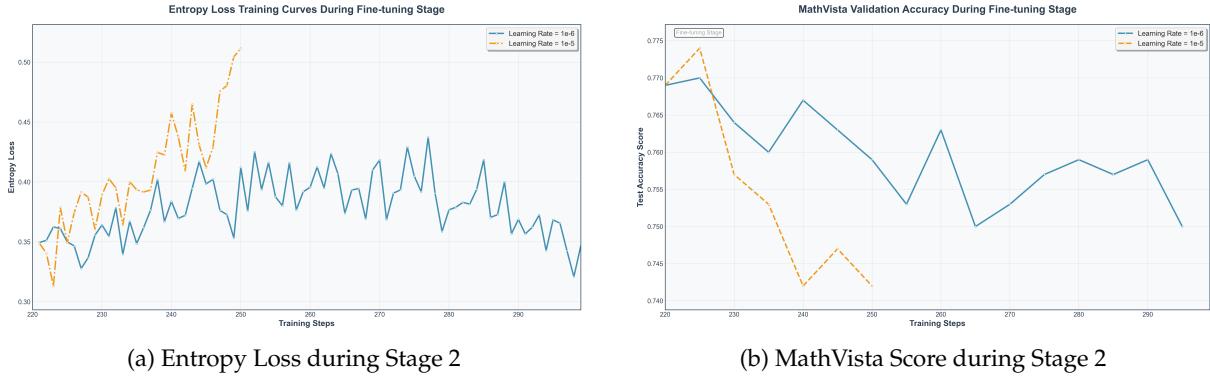


Figure 8: Impact of Learning Rates on RL Performance in Stage 2 (Hard Problems). This figure shows key metrics for different learning rate strategies on Hard Problems **(a)** A higher learning rate indicates the policy’s instability on exploration when tackling Hard Problems. **(b)** This indicates the model struggle to generalize from harder to easier problems.

by permitting larger probability increments, amplifies this effect. While this amplification can boost exploration, when combined with high learning rates, it risks over-amplifying noisy advantages and destabilizing the loss landscape. This, in turn, leads to severe loss explosions and ultimately, model collapse.

5.6 Effectiveness of Connector-Only Tuning After RL Stage

After undergoing RL training on mathematical data, the model demonstrates improved reasoning ability. To further enhance its generalization, we apply Connector-Only Tuning with multi-domain knowledge data to re-balance the knowledge base and achieve better performance across a wide range of domains. Here, we conduct an ablation study to investigate the effect of tuning different components (ViT/LLM/Connector) in this stage.

As summarized in Table 2, tuning the LLM or the ViT components alongside the connector significantly degrades performance, with accuracy dropping to 73.2 and 69.7 respectively. This suggests that excessive parameter updates in these core components disrupt the reasoning patterns reinforced during RL. In contrast, Connector-Only Tuning, which leaves the model’s intrinsic language and vision representations untouched, effectively preserves the reasoning capability while incorporating multi-domain knowledge. Consequently, Connector-Only Tuning significantly improves the model’s performance on the MMMU benchmark, achieving an accuracy of 76.0. This phenomenon further demonstrates the fundamental role of the connector module in facilitating robust cross-modal alignment in multimodal reasoning models.

Table 2: Ablation study on tuning different components after RL stage (MMMU benchmark).

RL Stage (baseline)	Connector-Only Tuning	Connector + LLM Tuning	Connector + ViT + LLM Tuning
Accuracy (%)	74.5	76.0	73.2

6 Discussions

In the following, we conducted a series of carefully designed exploratory experiments to better understand the nature of reasoning in visual language models.

6.1 Exploring Whether Visual Language Models Develop Reasoning via Memorization or Generalization

In practice, we observe that some models are able to produce responses with a reasoning-like style and achieve relatively high scores on in-distribution validation sets. However, in most cases, such behavior merely reflects memorization of existing patterns rather than genuine activation of a reasoning ability that generalizes across domains. These models tend to generate outputs with a *reasoning style* without truly acquiring a *generalizable reasoning ability*. The following experiments illustrate this distinction.

We take the cold-start model as the baseline and compare two different training strategies with **mathematics data**: RL (implementation details refer to Section 3) and SFT (conducted with a global batch size of 64 and a learning rate of $1e-5$, scheduled using a cosine decay with a warmup ratio of 0.03) using reasoning-style data distilled from our Skywork-R1V3-38B. We evaluate both models on *in-domain* (MathVista) and *out-of-domain* (MMMU) benchmarks to examine their ability to generalize beyond the domain of training.

Table 3: Comparison of memorization and generalization between RL and SFT.

Method	In-Domain	Out-of-Domain	Reasoning Style
Baseline	71.3	68.1	-
RL	77.2	74.5	Yes
SFT	75.9	65.4	Yes

As shown in Table 3, both RL and SFT enhance the model’s in-domain mathematical reasoning ability and result in responses with reasoning-like style. However, only RL leads to improved out-of-domain performance (74.5 on MMMU), demonstrating that the reasoning ability acquired through RL is genuinely generalizable. In contrast, SFT fails to generalize and even harms the out-of-domain performance (dropping to 65.4), likely due to overfitting to biased knowledge introduced at the finetuning stage. These results highlight that RL effectively activates a true generalizable reasoning ability, while SFT tends to merely imprint a reasoning style on the outputs without substantial generalization.

6.2 Exploration on Slow thinking to Fast thinking in Inference

Slow Thinking We adapted the following chat template in slow-thinking mode. The structure follows a turn-based format:

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>human
[user input]<|im_end|>
<|im_start|>assistant
<think>[reasoning process]</think>
[summary response]<|im_end|>
```

NoThink Mode We adapted the following chat template to explicitly activate the fast thinking mode. The key distinction between this template and the slow thinking template lies in the inclusion of the `<think>` token after the `assistant` role token. Removing this token prompts the model to output an answer directly without engaging in intermediate reasoning. We denote the inference without `<think>` token as NoThink mode.

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>human
[user input]<|im_end|>
<|im_start|>assistant
[response]<|im_end|>
```

NoWait Mode Following Wang et al. (2025a), we implement another fast-thinking method by constraining the model’s generation during inference. Specifically, we prohibit certain deliberative tokens such as `Wait`, `Alternatively`, and `Hmm` via logit manipulation. This strategy suppresses the tendency toward exploratory reasoning and encourages immediate answer production, denoted as NoWait mode.

Table 4: Comparison of slow-thinking and fast-thinking inference modes on the MMMU benchmark.

	Slow Thinking	NoThink	NoWait
Avg Tokens	1272	263	1040
Accuracy (%)	76.0	68.9	74.1

Table 4 presents a comparison of slow-thinking and fast-thinking inference modes on the MMMU benchmark. The results reveal a clear trade-off between response length and reasoning performance.

The *Slow Thinking* mode achieves the highest accuracy (76.0%), benefiting from its explicit step-by-step reasoning process, albeit at the cost of generating significantly longer outputs (1272 tokens on average). In contrast, *NoThink*, which removes the reasoning phase entirely, produces extremely concise responses (263 tokens) but suffers a substantial drop in accuracy (68.9%). *NoWait*, which enforces a more direct answering style by banning deliberative tokens during inference, strikes a better balance: it reduces the average output length by approximately 18% compared to *Slow Thinking* while retaining a relatively high level of accuracy (74.1%).

6.3 The Effectiveness of Thinking Budget

We further investigate how the allocated *thinking budget*, defined as the maximum token allowance during generation, affects the model’s reasoning performance. The results, summarized in Figure 9, reveal a test-time scaling law: reasoning accuracy on the MMMU benchmark improves steadily with increased token budgets.

As shown in the figure, the accuracy on the MMMU benchmark increases steadily with a larger thinking budget. Specifically, allocating only 1,024 tokens yields an accuracy of 61.6%, which already lags behind the *no-thinking* mode baseline (68.9%). Increasing the budget to 2K tokens closes this gap, reaching 67.7%. Further increasing the budget to 4K and 8K tokens leads to significant gains, with accuracies of 72.3% and 75.2%, respectively. When the budget is extended to 16K tokens, the performance saturates at 76.0%, achieving the highest observed accuracy. These results demonstrate a clear trade-off between computational cost and reasoning effectiveness. Allowing a sufficient token budget is critical to fully leverage the model’s step-by-step reasoning capabilities.

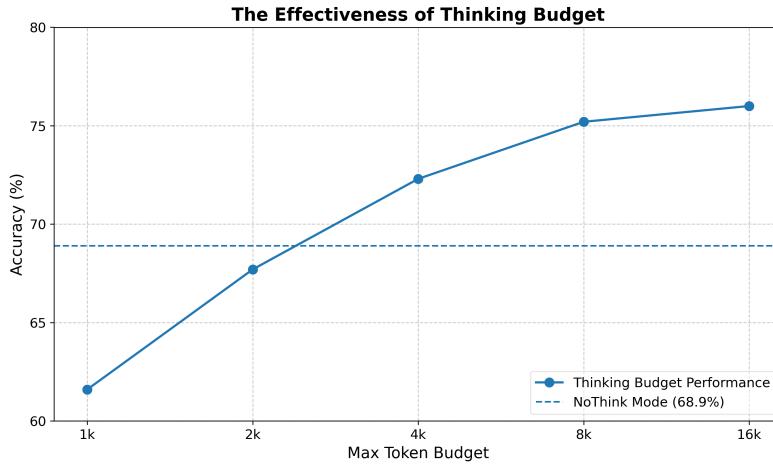


Figure 9: Model performance with different thinking budgets.

6.4 Hallucination in Skywork-R1V3’s Chain-of-Thought Impairs Reasoning Performance

Table 5: Impact of “I can’t see the image” on MMMU Performance

Subset	Number of Samples	Accuracy (%)
All	900	76.0
Without “I can’t see the image”	853	77.3
With “I can’t see the image”	47	51.1

We observe an intriguing phenomenon in Skywork-R1V3: hallucinated statements like “can’t see the image” frequently appear within chain-of-thought (CoT) rationales but are consistently absent from the final summary responses. This inherited behaviour—first noted in Skywork-R1V2, which reveals that despite explicitly stating an inability to perceive images within its reasoning trace, the model’s subsequent rationale demonstrates clear understanding of the visual content. This paradox represents one of the first documented challenges in transferring text-based reasoning capabilities to vision-language settings.

Critically, we found that the presence of this “can’t see the image” hallucination within the CoT significantly impedes model performance in downstream evaluations, as shown in Table 5. Our detailed analysis suggests this effect stems from the hallucination disrupting the coherence of the reasoning

process. While the phenomenon may also relate to the entropy of crucial reasoning tokens, we leave further investigation of this mechanism as an open question for the community.

6.5 Analysis on Entropy Token in Visual Reasoning Task

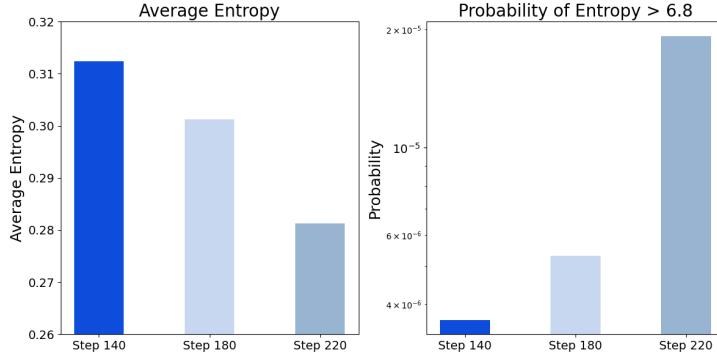


Figure 10: Entropy distribution at different GRPO steps

We observed an intriguing phenomenon during the RL stage when the reward steadily increased: while the average entropy of the generated tokens (i.e., the mean entropy of the vocabulary probability distribution at each forward step) progressively decreased, the probability of encountering high-entropy samples, although small, actually increased.

Figure 10 (left) illustrates the average entropy at different reinforcement training stages. It can be clearly seen that as training steps progress from 140 to 220, the average entropy consistently declines. However, Figure 10 (right) indicates that the probability of generating tokens with high entropy (defined here as entropy greater than 6.8) notably increases at later training stages. This counterintuitive observation suggests that although the model generally becomes more deterministic in its predictions through reinforcement learning, it also occasionally exhibits greater uncertainty, possibly reflecting increased exploration or exposure to challenging samples that require more deliberate reasoning (Wang et al., 2025c).

6.6 The Limitations of Self-Incentivized Reasoning for Instructed Models in RL Training

Test time scaling law (Jaech et al., 2024) strongly suggests that extended reasoning traces ("thinking before answering") enhance model performance on complex tasks. However, the field lacks robust visual reasoning models capable of sustaining such processes. Two dominant approaches exist for encouraging longer reasoning traces during training: (1) self-incentivized methods where models generate verification steps through instructional prompts (e.g., "The reasoning process MUST BE enclosed within tags. The final answer MUST BE put in \boxed{.}"), and (2) cold-start data priming prior to RL optimization.

Our experiments reveal a critical limitation of the self-incentive approach. As shown in Figure 11, models fail to develop genuinely long Chains of Thought (CoT) from scratch, exhibiting neither sustained "aha moments" nor length expansion proportional to reward growth before collapsing in later RL stages. This stands in stark contrast to cold-start initialization, which enables progressive reasoning trace elongation and significantly boosts both training and evaluation rewards. These findings align with recent work like DeepSeek-R1 and suggest that instruction-tuned models may suppress critical reasoning tokens (e.g., "Wait," "Alternatively") during finetuning—precisely those tokens that diversify reasoning paths and increase solution probability. The emergence of "aha moments" appears fundamentally linked to these suppressed tokens. Notably, RL-Zero approaches (Hu et al., 2025) demonstrate success comparable to cold-start methods, indicating that the specific RL initialization point is less critical than ensuring models can access and amplify critical tokens during their thinking process.

7 Conclusion

Skywork-R1V3 represents a groundbreaking advance in open-source vision-reasoning models (VRMs), establishing a new multimodal reasoning benchmark through its reinforcement learning (RL)-based training paradigm. As the first predominantly RL-enhanced open-source VLM, it achieves exceptional visual reasoning capabilities without external pretraining data while demonstrating remarkable generalization through cross-domain transfer—particularly from mathematical to other subject domains.

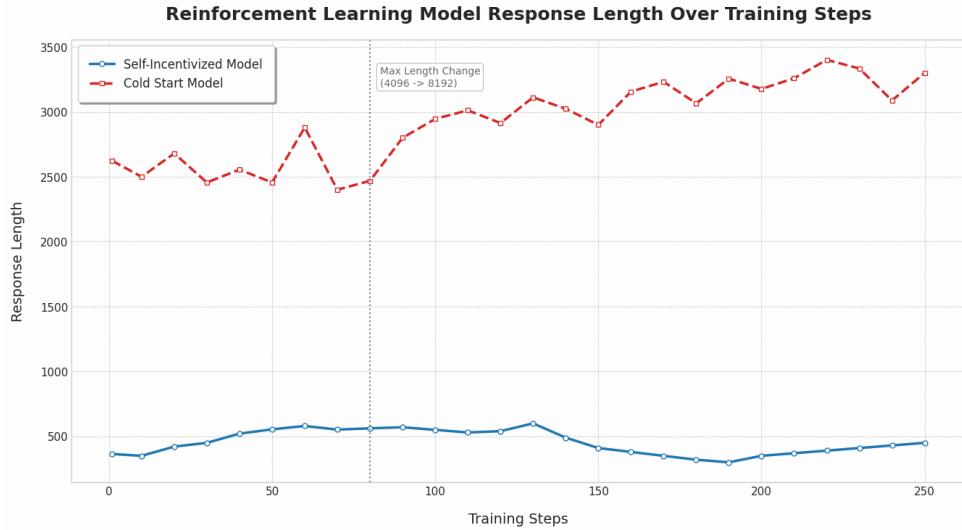


Figure 11: Response length during training: A comparison of two base models

Our work showcases RL’s transformative potential for refining vision-language alignment and provides crucial insights into cross-modal reasoning transfer. By seamlessly integrating textual and visual understanding, Skywork-R1V3 establishes the foundation for scalable, generalizable multimodal AI systems. This advancement marks a major milestone in open-source VLM development, offering a robust framework for RL-driven multimodal intelligence.

8 Limitations and Future Work

Limitations The scope of this report is intentionally narrow, and as such, we do not explore the specifics of vision encoder architectures or large language model (LLM) pre-training. Consequently, enhancing general visual capabilities was not a primary objective of this work. Looking ahead, we identify meticulous data curation as a critical direction for future research. We believe a more rigorously curated dataset will be instrumental in developing robust and capable models. Additionally, further improvements are needed in Video Understanding and Agent Training.

Another limitation we must acknowledge is the hallucination issue inherent in the reasoning contents of R1V2. We observed that this can hinder efficient reasoning and occasionally lead to erratic outputs. While this version of the reasoning model does not negate the benefits of reinforcement learning (RL), we plan to address this issue in future work.

Integrating Tool Use into Agentic Vision-Language Models with Reinforcement Learning The recent emergence of models such as OpenAI o3 and o4-mini highlights the growing potential of VLMs to perform tool-augmented reasoning. Such models extend beyond single-turn inference of Skywork-R1V3 by incorporating operations like code execution, web browsing, and image manipulation (e.g., cropping and zooming) to process and iteratively interact with visual inputs. Building upon this progress, a pivotal direction for future research lies in equipping VLMs with the ability to invoke external tools to address complex, multi-step multimodal tasks. In particular, end-to-end reinforcement learning approaches—such as Reinforcement Learning with Verifiable rewards (RLVR) (Mroueh, 2025)—offer a promising paradigm for allowing models to acquire tool-use skills through environmental interaction. However, their development faces significant challenges due to system optimization complexity and prohibitive inference computational costs, presenting a critical research gap that warrants further investigation.

Towards to Unified Vision-Language Models While this work focuses on enhancing reasoning capabilities through targeted training strategies, an important future direction is the unification of visual understanding and generation within a single VLM. A unified architecture, jointly trained on understanding and generation tasks, has the potential to leverage complementary signals, enhance multimodal alignment, and improve generalization across diverse scenarios (Zhang et al., 2025). Pursuing this direction may provide valuable insights into the design of VLMs capable of seamlessly transitioning between comprehension and generation, thereby expanding their utility in real-world applications that demand both modalities.

Advancing Physics Reasoning and Embodied Agents for Real-World Comprehension While Skywork-R1V3 demonstrates strong multimodal reasoning capabilities, particularly in STEM domains like mathematics and physics, further progress requires deepening its understanding of physical dynamics within embodied contexts. Enhancing the model’s physics reasoning will establish foundations for advanced embodied intelligence. We specifically aim to integrate such models into agentic systems capable of perceiving and interacting with simulated or physical environments. This necessitates not only high-fidelity vision-language alignment but also temporal planning and adaptive learning through interaction. Online reinforcement learning provides a promising pathway for agents to acquire complex reasoning and tool-use abilities via environmental feedback loops. Such experiential learning yields physical grounding that significantly improves real-world applicability. Furthermore, future agents must process multimodal sensory inputs in real-time and operate within physical constraints to ensure generalization and robustness. These advances will ultimately enable autonomous systems that understand and reason about the physical world with human-like competence and reliability.

9 Contributions

Core Contributors: Wei Shen*, Jiangbo Pei, Yi Peng, Xuchen Song[†], Yang Liu

Contributors: Jian Peng, Haofeng Sun, Yunzhuo Hao, Peiyu Wang, Jianhao Zhang, Yahui Zhou

* wei.shen@kunlun-inc.com [†]Corresponding author

References

- Anthropic. Claude-3.7, 2024. URL <https://www.anthropic.com/news/clause-3-7-sonnet>.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.
- Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models, 2025. URL <https://arxiv.org/abs/2505.22617>.
- GOOGLE DEEPMIND. Gemini 2.5: Our most intelligent ai modeld. 2025. URL <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking>.
- DeepSeek-AI. Deepseek-v3 technical report, 2024. URL <https://arxiv.org/abs/2412.19437>.
- DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multimodality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models, 2024. URL <https://arxiv.org/abs/2310.14566>.
- Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, et al. Seed1.5-vl technical report, 2025. URL <https://arxiv.org/abs/2505.07062>.
- Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025.
- Jingcheng Hu, Yinmin Zhang, Qi Han, Dixin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Openreasoner-zero: An open source approach to scaling up reinforcement learning on the base model, 2025. URL <https://arxiv.org/abs/2503.24290>.
- Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024. URL <https://arxiv.org/abs/2405.07987>.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Hel-
yar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- Ai Jian, Weijie Qiu, Xiaokun Wang, Peiyu Wang, Yunzhuo Hao, Jiangbo Pei, Yichen Wei, Yi Peng, and Xuchen Song. Csvqa: A chinese multimodal benchmark for evaluating stem reasoning capabilities of vlms, 2025. URL <https://arxiv.org/abs/2505.24120>.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL <https://arxiv.org/abs/2304.08485>.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.
- 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.
- Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhui Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

- Youssef Mroueh. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification, 2025. URL <https://arxiv.org/abs/2503.06639>.
- Dang Nguyen, Jian Chen, Yu Wang, et al. Gui agents: A survey, 2024. URL <https://arxiv.org/abs/2412.13501>.
- OpenAI. Gpt-4o system card, 2024. URL <https://openai.com/index/hello-gpt-4o/>.
- Yi Peng, Peiyu Wang, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Li Ge, Rongxian Zhuang, Xuchen Song, Yang Liu, and Yahui Zhou. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought, 2025. URL <https://arxiv.org/abs/2504.05599>.
- Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, MiaoXuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. We-math: Does your large multimodal model achieve human-like mathematical reasoning?, 2024. URL <https://arxiv.org/abs/2407.01284>.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL <https://arxiv.org/abs/1707.06347>.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL <https://arxiv.org/abs/2402.03300>.
- Hui Shen, Taiqiang Wu, Qi Han, Yunta Hsieh, Jizhou Wang, Yuyue Zhang, Yuxin Cheng, Zijian Hao, Yuansheng Ni, Xin Wang, Zhongwei Wan, Kai Zhang, Wendong Xu, Jing Xiong, Ping Luo, Wenhu Chen, Chaofan Tao, Zhuoqing Mao, and Ngai Wong. Phyx: Does your model have the "wits" for physical reasoning?, 2025. URL <https://arxiv.org/abs/2505.15929>.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.
- Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, et al. Mimo-vl technical report, 2025. URL <https://arxiv.org/abs/2506.03569>.
- Qwen Team. Qvq: To see the world with wisdom. <https://qwenlm.github.io/blog/qvq-72b-preview/>, 2024.
- Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL <https://qwenlm.github.io/blog/qwq-32b/>.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL <https://arxiv.org/abs/2302.13971>.
- Beichen Wang, Juexiao Zhang, Shuwen Dong, Irving Fang, and Chen Feng. Vlm see, robot do: Human demo video to robot action plan via vision language model, 2024a. URL <https://arxiv.org/abs/2410.08792>.
- Chenlong Wang, Yuanning Feng, Dongping Chen, Zhaoyang Chu, Ranjay Krishna, and Tianyi Zhou. Wait, we don't need to "wait"! removing thinking tokens improves reasoning efficiency, 2025a. URL <https://arxiv.org/abs/2506.08343>.
- Peiyu Wang, Yichen Wei, Yi Peng, Xiaokun Wang, Weijie Qiu, Wei Shen, Tianyidan Xie, Jiangbo Pei, Jianhao Zhang, Yunzhuo Hao, Xuchen Song, Yang Liu, and Yahui Zhou. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning, 2025b. URL <https://arxiv.org/abs/2504.16656>.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.
- Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning, 2025c. URL <https://arxiv.org/abs/2506.01939>.

Xiaokun Wang, Peiyu Wang, Jiangbo Pei, Wei Shen, Yi Peng, Yunzhuo Hao, Weijie Qiu, Ai Jian, Tianyidan Xie, Xuchen Song, Yang Liu, and Yahui Zhou. Skywork-vl reward: An effective reward model for multimodal understanding and reasoning, 2025d. URL <https://arxiv.org/abs/2505.07263>.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen, Yu-Jie Yuan, Jianhua Han, Hang Xu, Hanhui Li, Mrinmaya Sachan, and Xiaodan Liang. Seephys: Does seeing help thinking? – benchmarking vision-based physics reasoning, 2025. URL <https://arxiv.org/abs/2505.19099>.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts, 2024. URL <https://arxiv.org/abs/2407.04973>.

Weiyi Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models, 2025. URL <https://arxiv.org/abs/2504.15279>.

Yi Xu, Yuxin Hu, Zaiwei Zhang, Gregory P. Meyer, Siva Karthik Mustikovela, Siddhartha Srinivasa, Eric M. Wolff, and Xin Huang. Vlm-ad: End-to-end autonomous driving through vision-language model supervision, 2024. URL <https://arxiv.org/abs/2412.14446>.

Qiyi Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL <https://arxiv.org/abs/2503.14476>.

Jiakang Yuan, Tianshuo Peng, Yilei Jiang, Yiting Lu, Renrui Zhang, Kaituo Feng, Chaoyou Fu, Tao Chen, Lei Bai, Bo Zhang, and Xiangyu Yue. Mme-reasoning: A comprehensive benchmark for logical reasoning in mllms, 2025. URL <https://arxiv.org/abs/2505.21327>.

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*, pp. 9556–9567, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, 2025. URL <https://arxiv.org/abs/2409.02813>.

Jihai Zhang, Tianle Li, Linjie Li, Zhengyuan Yang, and Yu Cheng. Are unified vision-language models necessary: Generalization across understanding and generation, 2025. URL <https://arxiv.org/abs/2505.23043>.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?, 2024. URL <https://arxiv.org/abs/2403.14624>.

Jun Zhou, Jiahao Li, Zunnan Xu, Hanhui Li, Yiji Cheng, Fa-Ting Hong, Qin Lin, Qinglin Lu, and Xiaodan Liang. Firedit: Fine-grained instruction-based image editing via region-aware vision language model, 2025. URL <https://arxiv.org/abs/2503.19839>.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xinguang Wei, Hongjie Zhang, Haomin Wang, Weiyu Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhui Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL <https://arxiv.org/abs/2504.10479>.

Appendix

A Qualitative Examples

When she sees a folder named "models" in your personal laptop and all it contains some file ending with .pkl, .pt or .pth



Figure 12: example 1

Prompt

请解释这个 MEME 的含义

Model Response

<think> 好的，我现在需要分析用户提供的这个 meme。首先看一下图片和文字内容。图片上有两个人物：前面是一位女士坐在床边，低头看着手机，表情有些困惑或不悦。后面是一个男人站在洗手间门口，似乎正在换衣服或者整理衣物。文字部分写的是：“当她在你的个人笔记本电脑里看到一个名为“models”的文件夹，里面包含一些以 .pkl、.pt 或 .pth 结尾的文件。”

首先，我需要理解各个元素的含义。文件夹名 “models” 在编程中通常用来存放机器学习模型的文件。而 .pkl 是 Python Pickle 文件的扩展名，用于保存任意 Python 对象；.pt 和 .pth 则是 PyTorch 框架中常用的模型保存格式。所以这里的 “models” 文件夹实际上是指机器学习相关的模型文件。

接下来，为什么这会引发女士的不快呢？可能是因为 “models” 这个词在日常生活中也可能指代模特或与之相关的内容。女士可能误解了文件夹的用途，以为里面存放的是与模特相关的东西（比如照片或个人信息），从而产生怀疑或调侃的情绪。而实际上，这些文件是技术性的，和机器学习模型有关。

另外，图片中男士的动作可能暗示他正在逃避或回避女士的质问，进一步加强了这种误解带来的幽默感。这种反差（技术术语与日常生活中的误解）是这个 meme 的核心笑点。

需要确认是否存在其他可能的解释，但结合常见的编程术语和日常生活中的双关语，上述分析应该是正确的。最后，确保回答清晰易懂，解释技术和幽默之间的联系。</think>
这个 meme 的幽默点在于技术术语与日常生活中的误解之间的反差，具体解析如下：

- 技术背景：
文件夹名 “models” 及扩展名 .pkl、.pt、.pth，都是机器学习中的常见术语：
- .pkl 是 Python 用于保存对象的格式。
- .pt 和 .pth 是 PyTorch 框架保存模型的格式。
这里的 “models” 指代的是机器学习模型文件，而非字面意义的“模特”或“模型”（如人工制品）。
- 日常生活中的误解：
图片中的女性看到这些文件后，可能误以为 “models” 文件夹与现实中展示身材的“模特”相关（例如私密照片或不忠的证据），从而产生怀疑或调侃的情绪。而实际上这些文件只是技术文档，这种误会形成了笑点。
- 画面的补充：
男性站在浴室门口、似乎在整理衣物的动作，强化了女性对“另一段关系”的联想，让情境更显滑稽。

总结：这个 meme 利用了 “models” 一词在编程与日常用语中的双重含义，通过技术细节与生活场景的对比，制造了幽默效果。

Prompt

我买一千瓶这个饮料，中奖的钱的期望是多少？

Model Response

<think>...</think>

根据瓶身标签信息，需要计算各奖项的中奖金额期望值。具体如下：

各奖项及其中奖概率：

- 666 元红包: 0.000008% (即 0.0000008)
- 66 元红包: 0.005% (即 0.00005)
- 2 元红包: 0.5% (即 0.005)
- 1 元红包: 1.0% (即 0.01)
- 0.5 元红包: 16.5% (即 0.165)

各奖项期望值：

- 666 元: 0.0005328 元/瓶
- 66 元: 0.0033 元/瓶
- 2 元: 0.01 元/瓶
- 1 元: 0.01 元/瓶
- 0.5 元: 0.0825 元/瓶

汇总：

$$0.0005328 + 0.0033 + 0.01 + 0.01 + 0.0825 = 0.1063 \text{ 元/瓶}$$

1000 瓶:

$$1000 \times 0.1063 = 106.3 \text{ 元}$$



Figure 13: example 2

Prompt

The circuit in <image 1> consists of a general inverter connected to a load. Suppose the range of logical 1 is defined to be 4 to 5 V. Determine the minimum R_L in order that the output be 1 when the input voltage v_1 is zero.

Model Response

<think>...</think>

To determine the minimum R_L so that V_{out} stays between 4 V and 5 V when $v_1 = 0$ V, note that the transistor is off, and the divider between R_C and R_L sets V_{out} :

$$\frac{V_{cc}R_L}{R_C + R_L} \geq 4 \text{ V}$$

with $V_{cc} = 5 \text{ V}$ and $R_C = 5 \text{ k}\Omega$:

$$\frac{5R_L}{5k + R_L} \geq 4$$

cross-multiplying:

$$5R_L \geq 4(5k + R_L) \Rightarrow R_L \geq 20 \text{ k}\Omega$$

Final Answer:

20 kΩ

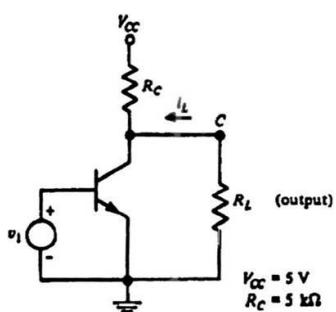


Figure 14: example 3

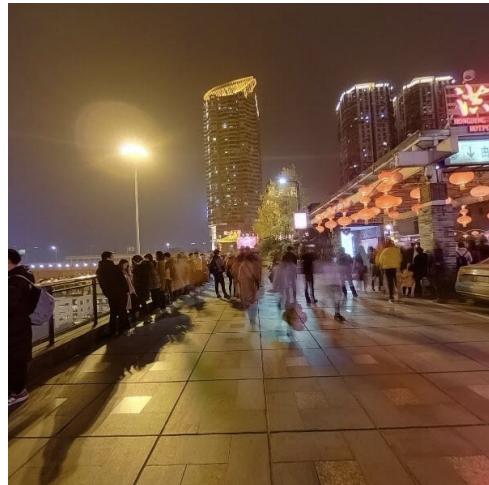


Figure 15: example 4

Prompt

You are participating in a geo-location challenge. Based on the provided image:

1. Carefully analyze the clues in the image (including architectural style, signage, vegetation, terrain, etc.);
2. Reason step by step to determine which continent, country, and city this location is most likely in, and explain your reasoning;
3. Based on your analysis, estimate the approximate latitude and longitude.

Please reason carefully and provide supporting evidence. Your final answer **must** include the following five lines of information, each on a separate line:

continent: [continent name]
country: [country name]
city: [city name]
lat: [latitude in decimal form]
lng: [longitude in decimal form]

Model Response

```
<think>...</think>
continent: Asia
country: China
city: Chongqing
lat: 29.5667
lng: 106.5500
```

Step-by-Step Explanation:

- **Architectural Clues:** The tall building with a golden crown-like structure and the presence of red lanterns suggest a Chinese urban setting, indicative of modern Chinese architecture often found in cities like Chongqing.
- **Signage Analysis:** The visible "WELCOME" sign with Chinese characters, along with the bustling waterfront promenade, aligns with Chongqing's famous riverside areas such as Nanbin Road.
- **Cultural Context:** Red lanterns and the crowded nighttime scene are characteristic of Chongqing's vibrant urban culture and popularity as a tourist destination.
- **Geographical Indicators:** The hilly terrain and riverfront setting are hallmark features of Chongqing's topography.

Thus, combining these observations, the location is most likely in Chongqing, China.