VisualThinker: R1-Zero's "Aha Moment" in Visual Reasoning on a 2B Non-SFT Model

Hengguang Zhou[*]
University of California, LA

Xirui Li[*]
University of California, LA

Ruochen Wang[†]
University of California, LA

Minhao Cheng
Pennsylvania State University

Tianyi Zhou
University of Maryland

Cho-Jui Hsieh
University of California, LA

TurningPoint

Abstract

The recent DeepSeek-R1 demonstrated how reinforcement learning with simple rule-based reward can enable autonomous development of complex reasoning in large language models, characterized by the "aha moment", in which the model manifest self-reflection and increased response length during training. However, attempts to extend this success to multimodal reasoning often failed to reproduce these key characteristics. In this study, we present the first successful replication of these emergent characteristics for multimodal reasoning on only a non-SFT 2B model. Starting with Qwen2-VL-2B and applying reinforcement learning directly on the SAT dataset, our model achieves **59.47%** accuracy on CVBench, outperforming the base model by approximately **~30%** and exceeding SFT setting by **~2%**. By further incorporating a small amount of cold-start data, we achieved **70.58%** accuracy on CVBench, a performance surpass GPT-4o-mini. In addition, we observed that applying RL to instruct models often leads to trivial and low-diversity reasoning trajectories and presents our insights and attempts to understand and mitigate this issue.

1 Introduction

Recently, DeepSeek R1 [5] has demonstrated how Reinforcement Learning (RL) with simple rule-based incentives can enable a Large Language Model (LLM) to build complex reasoning capabilities autonomously. A key finding from this work was the emergence of advanced reasoning patterns without explicit supervision—what the researchers termed the "aha moment," characterized by self-reflection and spontaneous increase in response length during training as the model learned to explore increasingly sophisticated problem-solving strategies.

Many researchers [3; 37; 24; 30] have attempted to extend this success to multimodal reasoning. However, most of these efforts have primarily struggled to reproduce the key characteristics exhibited by DeepSeek R1 mentioned above—specifically, the emergent "aha moment" and increased response length during reasoning. These implementations also failed to demonstrate the autonomous development of sophisticated reasoning strategies observed in DeepSeek R1's training.

[*] Equal contribution
[†] Main Advisor

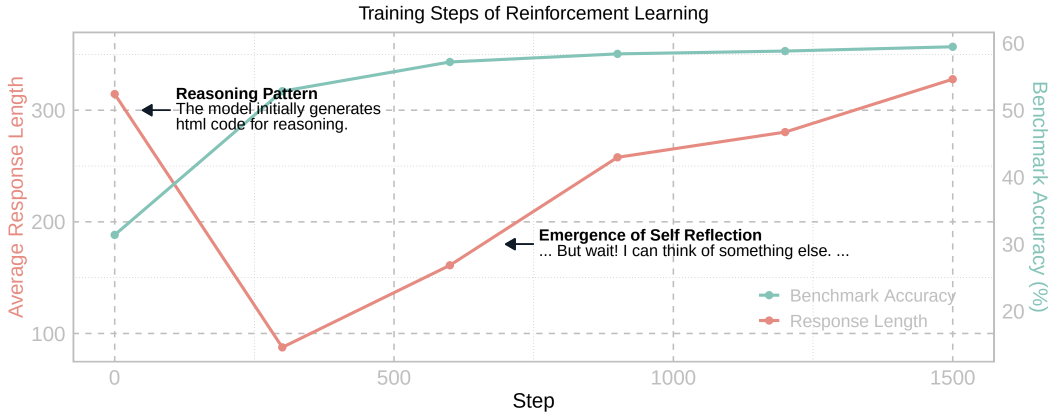


Figure 1: **The training dynamics of VisualThinker-R1-Zero on Qwen2-VL-2B base model**. Benchmark accuracy is measured on CV-Bench, and the average response length is calculated from rolling outs on SAT training samples. Initially, we observed a drop in length because the base model tended to generate HTML code. This behavior was quickly suppressed by RL, leading the model to adopt a more appropriate output format and a regular increase in response length. Afterwards, we observed a multimodal 'aha moment'—the emergence of self-reflection in models' response, as described in the DeepSeek-R1 paper, followed by a consistent positive correlation between response length and benchmark accuracy.

In this work, we present **the first successful replication** of these key characteristics for multimodal reasoning on **only a non-SFT 2B model**. With both "aha moment" and increased length (Figure 1), our approach demonstrates that direct application of reinforcement learning on non-sft model can sufficiently induce sophisticated reasoning capabilities even in smaller multimodal models without supervised fine-tuning. We start from the Qwen2-VL-2B [29] base model and directly perform reinforcement learning. Without any SFT, our model achieves **59.47% accuracy on CVBench** [27], outperforming the base model by approximately **~30%** and exceeding the SFT model by **~2%**. Inspired by [5], we further present **VisualThinker R1**, a derivative initialized with a small amount of reasoning trajectories curated from **VisualThinker R1-Zero** and existing visual reasoning trajectory dataset to ease RL cold-start instability while keeping the generation diversity typically lost in large-scale supervised fine-tuning, achieving **70.58% accuracy on CVBench**, comparable to GPT4o.

In addition, we share our insights and failed attempts in achieving R1-like visual reasoning using RL with instruct models. We observed that starting from a supervised fine-tuned model often failed to reproduce the observations and findings reported by DeepSeek-R1. Upon investigation on this issue, we found that (1) Despite improved performance, RL on instruct model leads to **trivial reasoning** rather than genuine problem-solving strategies, and (2) freezing visual components or naive length reward are ineffective at inducing deeper reasoning capabilities. Our study characterizes RL's behavior in visual-centric tasks and provide **the first clear evidence that emergent multimodal reasoning can arise under the right conditions**. We believe these findings open a promising direction for studying multimodal reasoning dynamics and lay the groundwork for future work seeking to induce and understand emergent reasoning in multimodal setting.

2 Related Work

2.1 Multimodal Reasoning

Researchers have demonstrated that LLMs can be post-trained to elicit enhanced reasoning abilities [16; 31]. With pre-trained visual encoders to understand visual content alongside textual data, multimodal LLM's reasoning abilities are investigated and enhancement typically requires sophisticated prompting designs [11] or large amounts of reasoning training data [26; 33]. The research community is increasingly interested in developing more natural methods to incentivize higher intelligence in models without relying on extensive supervised data or complex prompting techniques.

2.2 The “Aha Moments” in DeepSeek R1

A recent breakthrough study, DeepSeek R1 [5], demonstrated that reinforcement learning can incentivize a model’s reasoning abilities without any supervised reasoning data. Intriguingly, researchers [5] discovered an "aha moment" when directly applying RL with rule-based reward on mathematical datasets—a qualitative behavioral shift during RL training in which the model begins to autonomously exhibit more sophisticated reasoning behaviors such as revisiting earlier steps, expressing uncertainty, and performing self-correction. We summarize the key characteristic of DeepSeek R1 and compare them with our model and other multimodal replications in Table 1. Specifically, we highlight two emergent phenomena: the "aha moment" and increasing response length.

The "aha moment" refers to the model’s autonomous development of advanced problem-solving strategies during training, while the increasing response length indicates the model naturally learns to allocate more thinking time for reasoning tasks. It remains questionable whether existing multimodal replications [9; 3; 8] without these key characteristics can be considered truly valid implementations of the R1-Zero approach.

Following our first replication of multimodal "aha moment", several works attempt to elicit sophisticated reasoning patterns in VLMs through alternative approaches. LMM-R1 [19] introduced a two-stage training pipeline that transfers textual reasoning ability to multimodal reasoning ability. Vision-R1 [12] incorporated high-quality data as cold-start to overcome the challenges of trivial reasoning behavior. MM-EUREKA [15] further extends these explorations by scaling the cold-start dataset and model size. Other subsequent works explore similar directions towards visual reasoning by incorporating diverse data [34], explicitly enhancing perception [4], encouraging rethinking strategy [28] applying iteration of training [7].

Table 1: **Comparison between DeepSeek R1 and its multimodal replications**. VisualThinker-R1-Zero is the first model that replicates the two key chararistics of DeepSeek R1 on vision tasks - the emergent "aha moment" and increased response length during reasoning.

Feature	**DeepSeek R1-Zero [5]**	**VisualThinker R1 Zero (Ours)**	**R1-V [3]**	**R1-Multimodal-Journey [9]**	**open-r1-multimodal [8]**
Base Model	DeepSeek V3	Qwen2-VL-2B	Qwen2-VL-2B-Instruct	Qwen2-VL-2B-Instruct	Qwen2-VL-2B/7B-Instruct
Modality	Language	Vision + Language	Vision + Language	Vision + Language	Vision + Language
Aha Moment	Yes	Yes	No	Yes	No
Response Length Dynamics	↑	↑	↓	↓	↓
Supervised Fine-tuning?	No	No	Yes	Yes	Yes

3 VisualThinker

In this section, we first introduce the training recipe for VisualThinker-R1-Zero, where a small non-SFT base model exhibits the “aha moment” when we directly apply RL to the model. To mitigate cold-start instability in RL, we further present VisualThinker-R1, a model initialized with a small set of reasoning trajectories sampled from VisualThinker-R1-Zero and existing visual reasoning datasets.

Base Model Our method builds on Qwen-2-VL-2B [29] as the base model, applying GRPO [23] with a tailored chat template and prompting strategy to enhance its reasoning capabilities. We argue that applying GRPO to the base model is a more efficient and effective approach to replicate multimodal R1-style reasoning, as the base model preserves greater generation diversity. As shown in Section 5.1, the instruct variant, Qwen-2-VL-2B-Instruct, requires substantially more training resources and still fails to reproduce the “aha moment”, exhibiting several notable failure modes.

Training Recipe We train our models on **SAT** [21], a VQA dataset comprising 218k question-answer pairs synthesized using a photo-realistic physics engine to enhance spatial intelligence. Our training focuses on the static subset, which includes questions on relative spatial relationships, relative depth, and object counting. To let the base model explore various spatial reasoning for each question q in the dataset Q, we apply the following chat template:

Prompt Template

A conversation between User and Assistant. The user asks a question about the image, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer.\n User: {QUESTION} \n Assistant: Let me solve this step by step.\n <think>

For each question q in the dataset, the model generates a response o using this prompt template, and it is then optimized using the RL objective.

RL algorithm Existing attempts applying RL on top of fine-tuned visual models failed to replicate DeepSeek r1's key characteristics. In contrast, we witness **prolonged reasoning trajectory** and **"aha moment"** with an **overlooked approach that directly applies GRPO [23] to an non-SFT model.** Our findings suggest that this setting is the key to true "aha moment" in multimodal reasoning. Now we briefly review the GRPO algorithm we adopted for RL training.

To reduce the overhead of training an additional value function model as required by PPO [22], GRPO uses the average reward of sampled responses from the policy model as a baseline for computing the advantage. Specifically, given an input question q, we first sample a group of responses $\{o_1, o_2, \cdots, o_G\}$ and compute their corresponding rewards $\{r_1, r_2, \cdots, r_G\}$ with the reward model. The advantage is then computed as:

$$\hat{A}_{i,t} = \widetilde{r}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}. \tag{1}$$

The policy model is then optimized by maximizing the following KL objective:

$$\begin{split}\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(O|q)}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\Bigg\{\min\Bigg[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t},\\ \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\Bigg] - \beta\mathbb{D}_{KL}\left[\pi_\theta||\pi_{ref}\right]\Bigg\}\Bigg],\end{split} \tag{2}$$

where π_θ and π_{old} are the current and old policy, and ϵ and β are hyper-parameters introduced in PPO.

Reward Modeling Following DeepSeek-R1, our RL approach remains elegant, avoiding the use of reward models [18] or Monte Carlo tree search (MCTS)-like techniques [35]. Specifically, we employ a rule-based reward function that evaluates responses based on their format and correctness:

- If the response provides a final answer and is correct, the model receives an accuracy reward of +1.
- If the response encloses its thinking in `<think></think>` and the final answer in `<answer></answer>` tags, the model receives a format reward of +1.
- otherwise, the model receives 0 reward.

Initial experiments suggest that this reward function helps the policy model quickly converge towards generating responses in the desired format.

VisualThinker R1 Inspired by the design in DeepSeek R1 [5], we introduce **VisualThinker R1**, a variant initialized with a small number of curated reasoning trajectories to ease the cold start instability of RL training while preserving the generation diversity often lost in large-scale supervised fine-tuning. Specifically, we use 7K long-chain-of-thought samples drawn from VisualThinker R1-Zero's trajectories on the SAT dataset and the R1-Distilled Visual Reasoning Dataset adapted from R1-V [3]. We furthr include a limited set of examples challenging dynamic visual reasoning (e.g. egocentric movement, object movement) from SAT [20] during the reinforcement learning stage. Together, these components form our final training recipe.

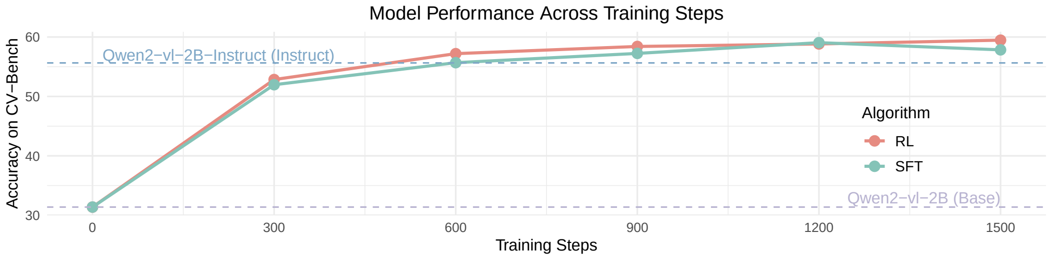


Figure 2: **Comparison between RL and SFT training.** Our method achieves a significant improvement over the base model and the instruction fine-tuned model. Specifically, Qwen2-VL-2B + R1 outperforms Qwen2-VL-2B (base model) by approximately ~30 and Qwen2-VL-2B-Instruct (instruction fine-tuned model) by ~5%, and Qwen2-VL-2B SFT (base model + SFT) by ~2%.

4 Experiments

In this study, we demonstrate that small non-sft model could solve vision-centric spatial reasoning task with our method, a type of benchmark often pose challenges to even larger models.

4.1 Evaluation

To evaluate the generalization of our method, we use three existing spatial benchmarks, CVBench [27], BLINK (spatial subsets) [10], and Visual Spatial Relations (VSR) [13]. These vision-centric benchmarks formulate natural language questions to assess models' abilities in spatial relationships, object counting, depth ordering, and relative distance. This setup allows us to systematically examine how well our recipe could improve spatial reasoning capabilities.

4.2 Highlighted Results and Findings

Table 2: **Accuracy (%) on vision-centric benchmarks**. The **bold** and underline values indicate the highest and second-highest scores among the open-source models.

Model	Count	Relation	Depth	Distance	CV Avg	RelDepth	SpatRel	BLINK Avg	VSR Avg
Proprietary Models									
GPT-4o [17]	70.43	80.46	81.50	80.00	77.59	72.58	85.31	79.40	80.27
GPT-4o-mini [17]	63.45	80.61	61.33	71.16	68.95	61.29	76.92	69.66	68.51
Open-Sourced Reasoning Models									
VLAA-Thinker-Qwen2VL-2B [2]	25.88	32.61	58.50	44.66	39.23	55.64	**65.73**	61.04	59.68
VLM-R1 (OVD) [3]	51.52	51.85	52.50	44.17	50.15	40.32	2.80	20.22	**80.77**
Baselines									
Qwen2-VL-2B [29]	54.69	22.46	0.16	31.66	31.38	13.70	0.69	6.74	0.00
Qwen2-VL-2B + SFT	60.02	68.92	55.00	45.83	57.84	58.06	47.55	52.43	35.80
Ours									
VisualThinker R1-Zero	59.64	66.76	54.16	56.66	59.47	50.80	55.94	53.18	62.32
VisualThinker R1	**63.32**	**79.08**	**75.00**	**66.50**	**70.58**	**61.29**	63.63	**62.54**	63.69

RL on base model autonomously developing longer responses correlated with improved performance on CV-Bench We compare our method with the non-sft pre-trained model, Qwen2-VL-2B as the baseline on CV-Bench. During training, we observed that the model autonomously generated increasingly longer responses, accompanied by improved performance, as shown in Figure 1. Specifically, VisualThinker-R1-Zero outperforms Qwen2-VL-2B (base model) by approximately ~30 and Qwen2-VL-2B-Instruct (instruction fine-tuned model) by ~5%.

RL demonstrates superior performance compared to SFT We further compare our method against the base model SFT on the same SAT dataset. Our experimental results demonstrate that applying reinforcement learning directly to the base model yields superior performance compared to SFT, with an improvement of approximately 2% as illustrated in Figure 2.

RL can help model improve general performance across vision-centric datasets Apart from CVBench, we tested our method on various spatial reasoning datasets, including BLINK and VSR, illustrated in Table 2. Our method demonstrated improved performance over the Qwen2-VL-2B (base) by ~30%, and the Qwen2-VL-2B SFT (base + SFT) by ~2% on CV-Bench. On the BLINK and VSR benchmark, our method also achieves around 27% advantage comparing against the model trained with SFT. This suggests that visual reasoning could significantly benefit from R1-Zero training, demonstrating more scalable training through RL's exploration of diverse reasoning.

Cold-start data incorporation further enhances performance We discovered that VisualThinker R1-Zero's performance could be further improved by approximately 11% through the strategic incorporation of a small amount of cold start reasoning trajectories and dynamic visual capability challenging samples. This optimization resulted in a total accuracy of 70.58% (VisualThinker R1), a level comparable to state-of-the-art proprietary models such as GPT-4o-mini.

Our training recipe outperforms existing methods As demonstrated in Table 2, our novel training recipe consistently achieves superior results across all evaluated vision-centric benchmarks compared to other approaches.

Multimodal Aha Moment A particularly intriguing phenomenon observed during the training of DeepSeek-R1-Zero is the occurrence of an "aha moment". This aha moment indicates DeepSeek-R1-Zero spontaneously builds a reasoning strategy, rethinking its initial approach for improved reasoning capability:

```
...
**Wait, wait. Wait. That's an aha moment I can flag here.**
Let's reevaluate this step-by-step to identify if
...
```

Following DeepSeek-R1, we extend this notion to the multimodal setting. Specifically, we define the **"multimodal aha moment"** we study in this work as:

> **Multimodal Aha moment** An emergent shift in multimodal large language model's reasoning behavior during RL training, where the model spontaneously begins exploring in more sophisticated, advanced strategy in visual reasoning.

Along with strong performance gain, we also observed our model exhibit very interesting "aha moment" aligned with the finding in DeepSeek R1. During training, the model spontaneously revisit its previous understanding and explore alternative options:

```
...
Therefore, dark brown wooden bed with white blanket is not above
the doorway.
**But wait! I can think of something else.**
...
```

We include more qualitative examples of this behavior in the Appendix F.

To quantitatively visualize the emergence of self-reflection behavior, we illustrate the mean and variance of frequency of reasoning-indicative keyword during training on identical samples across multiple runs in Figure 3. We observe reasoning-indicative keywords (e.g., "wait", "again") surge during Steps 600 – 900, marking the emergence of reflection and refinement strategies, and stabilize beyond Step 900, indicating consolidation of reasoning patterns.

5 Discussion

5.1 Investigating the Missing Reasoning Emergence in RL-trained Instruct Models

Starting from scratch with a non-SFT multimodal model, we have demonstrated that RL empowers **VisualThinker-R1-Zero** to attain robust reasoning capabilities without any supervised fine-tuning data. In addition to this replication, one might be inclined to apply RL directly to supervised fine-tuned models given its stronger instruction following capability. However, when attempting to replicate this success by applying RL to supervised fine-tuned models, we encountered a phenomenon: the absence of the "aha moment", where model develops emergent sophisticated reasoning behavior during training.

In following paragraphs, we share our investigations into this phenomenon, providing insights that may benefit future research. We first demonstrate how RL on instruct models leads to trivial reasoning

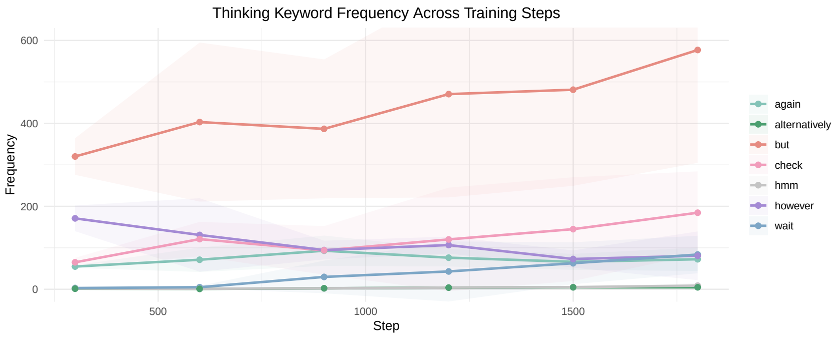


Figure 3: **Statistical analysis of reasoning-indicative keywords ("wait," "again", "but", "however", "so", "hmm", "check", "alternatively", and "mistake") across training steps in multi runs.** The results highlight the emergence of reflection and refinement behaviors after 600 steps, followed by stabilization of reasoning patterns beyond Step 900..

patterns rather than genuine reasoning capabilities (Figure 5). We then share our failed attempts to elicit sophisticated reasoning from instruct models, including **freezing the vision encoder during training** and **using a length-based reward to encourage longer responses**. Despite these efforts, we were unable to induce diverse exploration of advanced reasoning behavior purely through RL training on instruct models. In contrast, base models readily exhibit such diversity, suggesting that RL on base models might be a more effective option at this stage, and leaving the challenge of eliciting similar behavior in instruct models for future work.

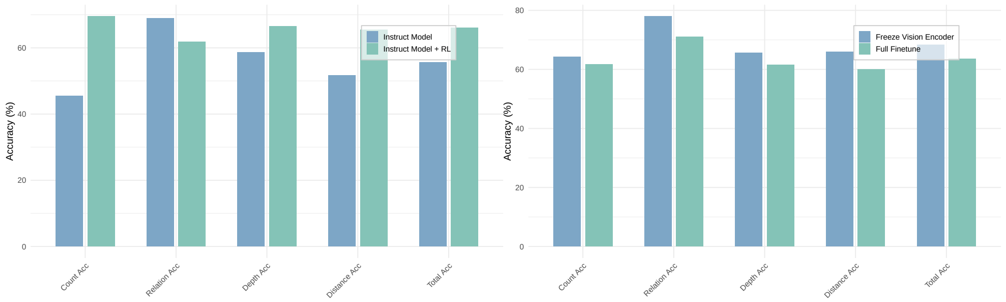


Figure 4: **Evaluation of reinforcement learning on instruct models.** (a) (left) Applying RL on instruct model indeed improves its performance. (b) (right) Freezing either the language or vision component further improves performance of RL on instruct models.

The Missing Reasoning Patterns Applying RL on instruct models degenerate into short and trivial reasoning rather than developing genuine reasoning capabilities (Figure 5) despite improvement on the model's performance (Figure 4 Left). The observed reasoning trajectory often follows a structure: a trivial and generic strategy within `<think></think>` tags followed by the answer between `<answer></answer>` tags. In this pattern, model bypasses the intended reasoning process and jumps directly to the final answer. While this may lead to immediate stronger training signal, it comes at the cost of exploring diverse and generalizable reasoning strategy.

Trivial Reasoning Trajectory

Question: Which object is closer to the camera taking this photo, the box (highlighted by a red box) or the desk (highlighted by a blue box)?
Response: <think>To determine which object is closer to the camera, I will measure the distance from the camera to each object. </think> <answer>box </answer>

Figure 5: **Example response of applying RL to supervised fine-tuned model.**

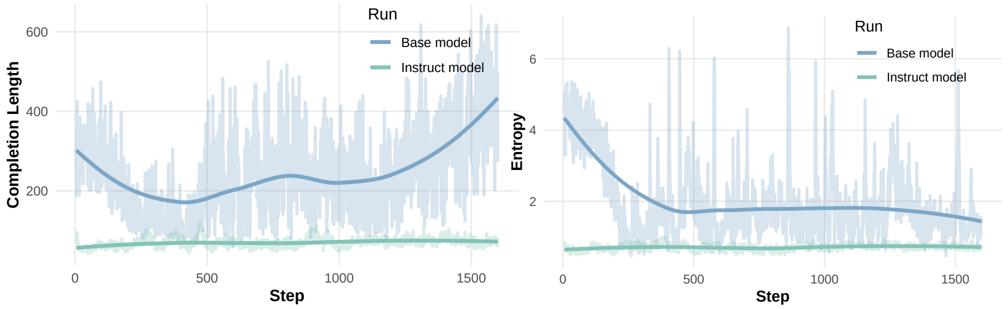


Figure 6: **RL training dynamics and example output. Left:** Completion length across training steps for RL on base and instruct models. **Right:** Entropy over tokens across training steps for RL on base and instruct models.

To better understand and quantify the training dynamics of instruct model, we measure the response length and entropy during training. As shown in Figure 6, the consistently higher entropy and response length of base models throughout training points to exploration of more diverse and sophisticated reasoning behaviors. In contrast, instruct models rapidly converge into trivial reasoning marked by extremely low response length and entropy, and fails to explore meaningfully throughout the training.

This finding aligns West et al. [32]'s observation that base models exhibit greater creativity than aligned models, providing a quantitative mechanism explaining why starting from base model enables more sophisticated reasoning capabilities while instruct models might remain stuck in local optima.

While instruct models may seem appealing for RL due to their stronger capabilities, we showed that it is not trivial to elicit this exploration behavior on instruct model by sharing a few of our failed attempts below.

Can Freezing Vision Encoder Elicit "Aha Moment"? Given the improved performance alongside the degenerating response diversity characteristic of RL training on instruct model, we hypothesize that the improvement could be attributed to improved visual representation. We thus investigated whether we could encourage reasoning exploration by freezing the visual components during training. Shown in Figure 4 and Figure 7, freezing vision components slightly improved performance, initially alleviates the decline in response length, yet it eventually leads to the same trivial reasoning patterns observed in full fine-tuning.

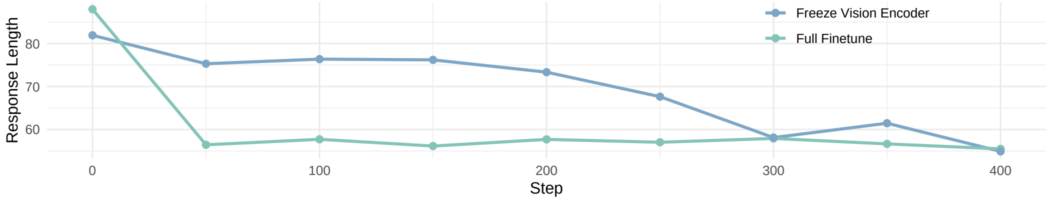


Figure 7: **Response length across training steps for different fine-tuning settings during RL.** The x-axis represents training steps, while the y-axis shows the response length. Models with different fine-tuning configurations are compared: Freeze Vision Encoder (blue), and Full Finetune (green). Despite our hypothesis, freezing the vision encoder did not force the model to develop sophisticated reasoning strategies, as evidenced by the consistently short responses length.

Failed Attempts to Induce "Aha Moment" through Length Rewards One key characteristic of DeepSeek R1's success is its increased response length during training. Inspired by this, we investigated whether we could encourage more sophisticated reasoning patterns by directly rewarding longer responses. Thus, we implemented a length-based reward mechanism: in addition to the vanilla accuracy and format rewards described in Section 3, we added an auxiliary reward of $+0.001$ for each additional token generated. However, as shown in Figure 8, this approach failed to produce the desired emergence of sophisticated reasoning. Instead, it led to reward hacking behaviors, with models generating extremely long yet meaningless and repetitive content as illustrated in Figure 9.

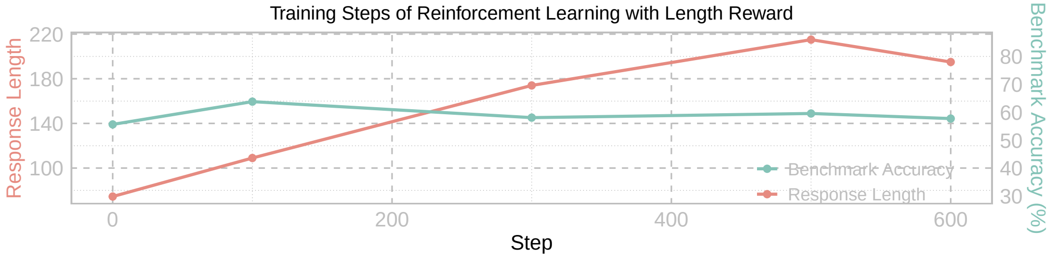


Figure 8: **Model performance during training with length-rewarded RL on instruction finetuned models.** Despite implementing length rewards to encourage longer responses, model performance shows no improvement in reasoning quality as response length increases. While average response length (red line) grows substantially, benchmark accuracy (green line) remains relatively stable, suggesting that longer responses do not necessarily translate to enhanced reasoning capabilities.

Length Reward Hacking Example

Question: Considering the relative positions of the giraffe (annotated by the red box) and the person in the image provided, where is the giraffe (annotated by the red box) located with respect to the person?
Model Response: <think> - The giraffe (annotated by the red box) is located in the foreground of the image. - The person is in the background, standing behind the giraffe. - The giraffe is not below the person. - The person is not in front of the giraffe. - ... [Repeats until it reaches max_new_tokens limit] ... </think>

Figure 9: **Example of reward hacking**. Model hacks reward by repeating trivial response.

5.2 Ablation Study on VisualThinker R1

5.2.1 Ablation on Cold Start

As detailed in Section 4, we stabilized the cold start training stage and unlock greater potential of our base model training with a small amount of reasoning trajectories data. To isolate the effect of cold-start trajectories, we conduct ablation studies comparing models trained solely on these trajectories versus model trained with the full recipe (i.e., VisualThinker R1 = Cold Start + GRPO). Results in Table 3 indicates that the performance gains are still primarily driven by the RL training.

Table 3: **Ablation study on the effect of cold start trajectories**. We can observe that the full training recipe is significantly higher performance comparing to model trained only on cold start data. This indicates that the performance gains are still primarily driven by the RL training.

Training Method	Total Acc (%)	Count Acc (%)	Relation Acc (%)	Depth Acc (%)	Distance Acc (%)
Cold Start Only	48.48	62.44	29.08	50.17	49.50
Full Recipe	70.58	63.32	79.08	75.00	66.50

5.2.2 Ablation on Training Temperature

To understand the impact of temperature on RL training, we conduct an ablation study on varying training temperature. Overall, we observed that training with low temperature (e.g., 0.25) exhibits excessive determinism, limiting exploration and ultimately constraining the model's maximum achievable accuracy. Additionally, all settings with temperature below 1.0 showed a consistent pattern of response length shrinking in later training stages, suggestion a collapse in exploration and failure to discover more structured reasoning strategy. Thus, we adopted a training temperature of 1.0, in which we observe more well-structured reasoning.

Table 4: **Ablation study on the effect of temperature**. Training with temperature 1.0 produces model with more well-structured reasoning behavior while achieve competitive quantitative performance.

Temperature	Total Acc (%)	Count Acc (%)	Relation Acc (%)	Depth Acc (%)	Distance Acc (%)
0.25	57.16	59.39	64.77	55.50	47.67
0.50	61.06	56.59	71.53	59.00	57.66
0.75	59.51	62.69	67.53	53.50	52.66
1.00	59.47	59.64	66.76	54.16	56.66

6 CONCLUSION

This paper presents VisualThinker R1, the first successful multimodal replication of DeepSeek R1's emergent reasoning characteristics. By applying reinforcement learning directly to a non-fine-tuned Qwen2-VL-2B model, we observed both the "aha moment" and increased response length during training, key indicators of autonomous reasoning development. Empirically, our approach achieved 59.47% accuracy on CVBench, outperforming both base and instruction-tuned models without any supervised fine-tuning. We further incorporate a small amount of cold-start data and achieve 70.58% accuracy on CVBench, a performance surpass GPT-4o-mini. We also share our insights and failed attempts in achieving R1-like reasoning using RL with instruct model: applying RL to supervised fine-tuned models leads to trivial reasoning trajectories rather than genuine problem-solving strategies. These findings explore RL training in multimodal reasoning, potentially offering a more scalable approach to developing AI systems with robust visual reasoning capabilities.

7 Reproducibility Statement

We have taken several steps to ensure the reproducibility of our results. A detailed description of experimental details, including all models, packages used, and hyperparameter settings are outlined in B. Furthermore, we provide our source code in the supplementary materials.

References

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

[2] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models, 2025. URL `https://arxiv.org/abs/2504.11468`.

[3] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. `https://github.com/Deep-Agent/R1-V`, 2025. Accessed: 2025-02-02.

[4] Zhangquan Chen, Ruihui Zhao, Chuwei Luo, Mingze Sun, Xinlei Yu, Yangyang Kang, and Ruqi Huang. Sifthinker: Spatially-aware image focus for visual reasoning. *arXiv preprint arXiv:2508.06259*, 2025.

[5] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL `https://arxiv.org/abs/2501.12948`.

[6] Huilin Deng, Ding Zou, Rui Ma, Hongchen Luo, Yang Cao, and Yu Kang. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. *arXiv preprint arXiv:2503.07065*, 2025.

[7] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*, 2025.

[8] EvolvingLMMs-Lab. open-r1-multimodal. https://github.com/EvolvingLMMs-Lab/open-r1-multimodal, 2025. Accessed: March 6, 2025.

[9] FanqingM. R1-multimodal-journey. https://github.com/FanqingM/R1-Multimodal-Journey, 2025. Accessed: March 6, 2025.

[10] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive, 2024. URL https://arxiv.org/abs/2404.12390.

[11] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2025.

[12] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2025. URL https://arxiv.org/abs/2503.06749.

[13] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning, 2023. URL https://arxiv.org/abs/2205.00363.

[14] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

[15] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning, 2025. URL https://arxiv.org/abs/2503.07365.

[16] OpenAI. Learning to reason with llms, 2024. URL https://openai.com/index/learning-to-reason-with-llms/.

[17] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi

Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

[18] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

[19] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl, 2025. URL https://arxiv.org/abs/2503.07536.

[20] Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. Sat: Spatial

aptitude training for multimodal language models, 2024. URL https://arxiv.org/abs/2412.07755.

[21] Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, Kuo-Hao Zeng, et al. Sat: Spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

[22] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

[23] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[24] Haozhan Shen, Zilun Zhang, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model. https://github.com/om-ai-lab/VLM-R1, 2025. Accessed: 2025-02-15.

[25] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*, 2024.

[26] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025.

[27] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2025.

[28] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025.

[29] 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, 2024. URL https://arxiv.org/abs/2409.12191.

[30] Xiaodong Wang and Peixi Peng. Open-r1-video. https://github.com/Wang-Xiaodong1899/Open-R1-Video, 2025.

[31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

[32] Peter West and Christopher Potts. Base models beat aligned models at randomness and creativity. *arXiv preprint arXiv:2505.00047*, 2025.

[33] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

[34] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

[35] Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search, 2024. URL https://arxiv.org/abs/2406.03816.

[36] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.

[37] Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng andDongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. `https://github.com/hiyouga/EasyR1`, 2025.

[38] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Contents

A Limitations and Future Works

While our study demonstrates the emergence of complex reasoning behaviors in a non-SFT 2B multimodal model, several limitations should be noted.

Intensive Resource Demand One limitation of our approach is its high demand for computational resources. The adopted GRPO training approach in our work is highly resource-intensive, requiring substantial GPU memory and compute time. This is largely due to the need to generate and evaluate long CoT rollouts during training. The reasoning model often produces lengthier responses than non-reasoning models at inference time, resulting in increased decoding cost. These resource demands may pose challenges to the scalability of the approach.

Domain and Task Our current exploration focuses on benchmarks that test spatial and visual reasoning capabilities but do not cover more advanced visual understanding. Specifically, the datasets we use, such as SAT and CVBench, do not require the model to interpret complex visual scenes, or focus on fine-grained visual details. Understanding how rl-based visual reasoning model generalizes to more complex visual settings remains an important direction for future exploration.

Modality Limitations on Reasoning Steps Our exploration purely focuses on models reason in language. However, in cases such as describing relative spatial layouts, geometric transformations, or occlusion relationships, where text-only reasoning could be cumbersome or ambiguous, reasoning on more modalities (sketches, spatial diagrams, or visual annotations) are potentially more effective. Expanding beyond text-only reasoning remains a promising direction.

Generalization to Other Base Models Remains Unverified A stronger demonstration of our conclusion's generalizability would be to apply RL directly to other base models. Unfortunately, there are very few open-source small multimodal non-sft models available, and the ones that exist are very limited in capability. While we attempted to train other open-source base models (e.g., SmolVLM [14], Paligemma2 [25], and InternVL3 [38]), we found that many of these models struggled to produce even the minimally structured outputs needed for reward supervision, particularly under the constraints of non-SFT initialization and small model size. This appears to stem not from a limitation of the RL method itself, but rather from the models' insufficient capacity to generate structured responses that can be meaningfully guided by a reward signal.

Despite these limitations, our results offer a strong proof of concept: with appropriate reward structures and training strategies, a 2B non-SFT multimodal model can develop emergent reasoning behaviors previously thought to require much larger or instruction-tuned models. While further investigation is needed, we see no clear evidence of fundamental barriers that would prevent extending this paradigm to larger models, more complex tasks, or richer multimodal interactions in the future.

B Experiments Settings

B.1 Implementation Details

All experiments are conducted using four NVIDIA H100 GPUs (80GB each), setting the batch size to 1 per device. The model is trained for $1,500$ steps with a learning rate of 1×10^{-6} and a temperature of 1.0. We found that allowing long responses is essential for observing increasing response length during training, so we set the maximum response length to 700. During GRPO optimization, we sample 8 responses per step and apply a KL coefficient of 0.04.

B.2 Evaluation Settings

For models whose benchmark results are not publicly available, we conduct our own evaluation. We use the official evaluation code provided by the model authors if it is available. Otherwise, we evaluate the models under both the reasoning and no-reasoning settings using the corresponding prompt templates demonstrated in Figure 10. Reasoning models are evaluated with reasoning prompt template, while non-reasoning models are evaluated with non-reasoning prompt template.

Table 5: **Hyper-parameters of VisualThinker R1 GRPO training.**

Setting	**Value**
Batch Size per Device	1
Gradient Accumulation Steps	2
Training Steps	1500
Learning Rate	1×10^{-6}
Temperature	1.0
Maximum Response Length	700
Number of Responses per GRPO Step	8
KL Coefficient	0.04

Evaluation Prompt

reasoning prompt: {Question} Output the thinking process in <think> </think> and final answer in <answer> </answer> tags.
non-reasoning prompt: {Question}

Figure 10: The evaluation prompt template

B.3 Specification of Models for Comparison

Table 6 lists the specific versions of the proprietary models used in our comparison. For GPT-4o and GPT-4o-mini, we report results based on the versions released on `2024-08-06` and `2024-07-18`, respectively.

Table 6: **The versions of proprietary model used for comparison**

Model	Version
GPT-4o	gpt-4o-2024-08-06
GPT-4o-mini	gpt-4o-mini-2024-07-18

B.4 Computation Resource

All experiments are conducted using four NVIDIA H100 GPUs (80GB each), setting the batch size to 1 per device. The model is trained for $1,500$ steps and takes roughly 30 hours to complete.

C Correlation Between Emergent "Aha Moment" Dynamics and Performance Improvements

This section provides extended analysis illustrating the link between the emergence of reflective reasoning behaviors (the "aha moment") and performance improvements during reinforcement learning. Our experiments suggest a strong and highly consistent correlation pattern across multiple runs.

Across multiple RL runs of VisualThinker-7B, we consistently observe the same three-phase response-length dynamics, which aligns closely with changes in model accuracy.

Phase 1 — Response-Length Drop (Format Compliance Phase): First, at the start of training, response length decreases as the model focuses on optimizing format reward. During this stage, accuracy remains stable or increases slightly.

Phase 2 — Rapid Response-Length Growth (Emergence Phase / "Aha Moment"):Second, a rapid emergence phase follows, characterized by a sharp increase in response length and the appearance of reflective behaviors (e.g., "wait", "let me reconsider"), as also supported by the keyword statistics shown in Table 12.

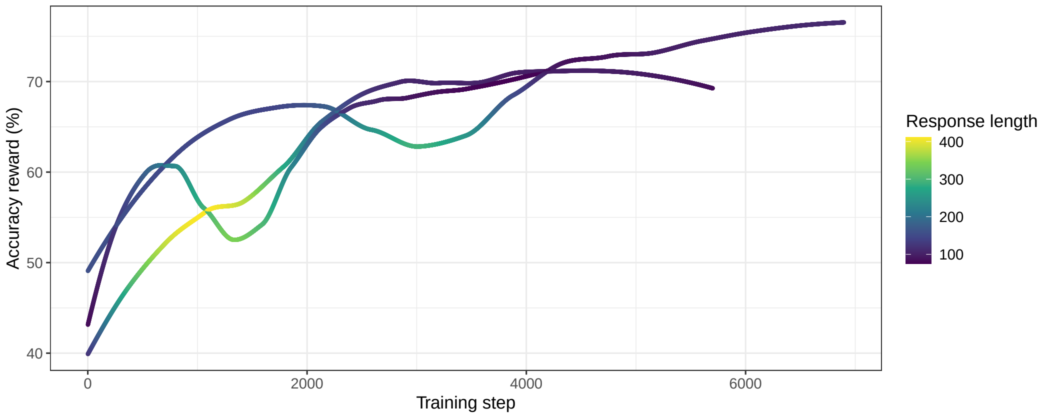


Figure 11: Across all runs, accuracy improvements consistently follow a shared sequence with the accuracy rise always aligning with the end of the emergence phase—strongly indicating that reflective-behavior emergence is tightly correlated with the development of more effective reasoning strategies under RL.

Accuracy during this stage becomes worse, consistent with increased exploratory behavior.

Phase 3 — Stabilization After the Emergence: Third, once this exploratory phase concludes, the model converges to a more efficient reasoning strategy, and response length stabilizes or decreases.**Importantly, it is during this post-emergence stabilization phase that accuracy consistently rises significantly across all runs.**

Across all runs, we observe a consistent ordering of events: (1) response-length drop, (2) emergence phase, (3) stabilization, and (4) accuracy improvement. The timing of the accuracy increase aligns closely with the conclusion of the emergence phase in every run. This pattern persist consistently across all runs, providing a strong empirical evidence that the emergence of reflective behavior is tightly correlated with the acquisition of more effective reasoning strategies under RL.

D Statistical Robustness and Multi-Seed Evaluation

This section provides additional statistical analysis to ensure that the empirical trends reported in the main paper are stable, reproducible, and not dependent on a single random seed. All experiments have been re-run with three or more randomized seeds, and updated figures and tables now report mean values together with variability measures.

D.1 Clarification of Reported Statistics

All plots now display the mean across multiple runs, with shaded regions representing the standard error of the mean. Tables throughout the paper have been updated to report mean ± SEM wherever applicable. These additions remove ambiguity regarding whether values represent mean or median statistics and ensure that all main results are aggregated across multiple seeds rather than relying on a single initialization.

D.2 Multi-Seed Accuracy Results

Table 7 summarizes the multi-run performance on downstream benchmarks. The improvements remain consistent across seeds, confirming the stability of the observed gains.

D.3 Multi-Seed Reflective-Reasoning Keyword Dynamics

To verify that the reflective-reasoning indicators are not artifacts of a particular seed, we aggregate keyword statistics across multiple runs. Table 8 reflects the updated multi-seed values corresponding to Figure 3 in the main paper.

Table 7: Multi-run evaluation on downstream benchmarks.

Benchmark	CVBench	BLINK
VisualThinker-R1-Zero (Ours)	60.24 ± 3.40	55.99 ± 3.06

Table 8: Multi-run reflective-reasoning keyword dynamics across training steps.

Keyword	300	600	900	1200	1500	1800
again	54.80 ± 3.37	71.20 ± 29.43	93.00 ± 36.65	76.20 ± 21.27	66.20 ± 19.32	72.40 ± 28.27
wait	3.00 ± 1.26	5.00 ± 3.41	29.80 ± 39.34	43.00 ± 72.50	62.80 ± 50.47	83.80 ± 45.23
but	320.20 ± 44.24	403.20 ± 191.77	386.80 ± 167.49	470.60 ± 249.14	481.20 ± 231.87	577.00 ± 272.22
however	171.00 ± 30.93	131.00 ± 88.56	94.60 ± 22.52	106.60 ± 20.44	73.00 ± 20.75	80.80 ± 58.66
hmm	1.80 ± 1.60	1.20 ± 1.17	2.40 ± 3.32	4.00 ± 4.77	5.20 ± 2.40	8.80 ± 5.46
alternatively	1.40 ± 0.49	1.00 ± 1.10	2.40 ± 3.83	4.00 ± 3.03	4.80 ± 3.54	4.80 ± 4.45
check	65.00 ± 10.92	121.00 ± 41.51	93.80 ± 58.63	120.20 ± 124.83	145.00 ± 125.12	184.60 ± 99.69

These aggregated results confirm that the qualitative trends highlighted in the main paper—response-length dynamics, emergent self-reflective phrasing, and downstream benchmark gains—hold consistently across different seeds, demonstrating that the findings are not attributable to a single lucky initialization.

E Extended Evaluation on Larger and Diverse Non-SFT Base Models

To assess the generality of the proposed RLVR training recipe, we extend our study beyond the 2B non-SFT base model used in the main paper. In particular, we evaluate two additional non-SFT base models: (1) **Qwen2-VL-7B**, a larger model from the same family, and (2) **InternVL-2.5-1B**, a smaller model from a different architecture family.

To ensure fair comparisons, we evaluate only the *non-SFT base variants* of these stronger model families, avoiding confounding factors introduced by large-scale instruction tuning pipelines. This isolates the contribution of our post-training recipe rather than the influence of proprietary or disproportionately large SFT data.

Across both models, we consistently observe:

1. **Accuracy gains** on CVBench,
2. **First-drop–then-rise response-length dynamics** (the characteristic "aha moment" trajectory),
3. **Reflection-indicative keyword patterns** following the same drop–rise–drop trend.

These findings demonstrate that the emergent behaviors induced by RLVR are reproducible, scalable, and architecture-agnostic.

Table 9: CVBench accuracy improvements across additional non-SFT base models.

Model	Base Acc.	+ RLVR Recipe
Qwen2-VL-7B	66.22 ± 0.00	78.02 ± 0.25
InternVL-2.5-1B	41.88 ± 0.00	50.88 ± 0.38

Table 10: Response-length statistics for Qwen2-VL-7B across training steps.

Steps	0	300	600	900	1200	1500
Length	169.25	110.46	209.57	139.01	334.91	188.91

Table 11: Response-length statistics for InternVL-2.5-1B across training steps.

Steps	0	1200	2400	3600	4800	6000
Length	95.13	91.33	102.09	108.75	123.71	135.43

Table 12: Keyword statistics for Qwen2-VL-7B.

Keyword / Interval	300	600	900	1200	1500
again	44	53	61	88	86
wait	0	12	63	75	65
but	215	423	526	597	524
however	116	106	88	97	97
hmm	0	1	9	7	6
alternatively	0	3	3	4	4
check	295	288	296	309	295

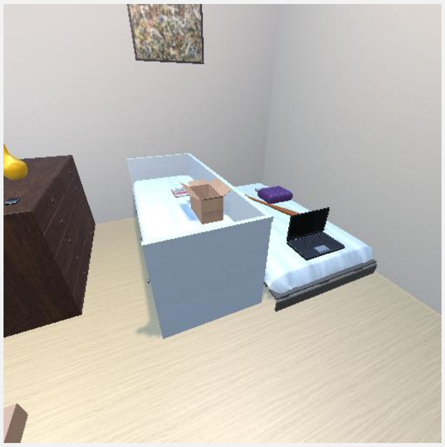
Table 13: Keyword statistics for InternVL-2.5-1B.

Keyword / Interval	1200	2400	3600	4800	6000
again	67	185	205	189	62
wait	31	53	106	227	3
but	3583	6197	5787	5100	3246
however	1719	3213	3188	2597	1477
hmm	7	9	23	32	0
check	245	460	453	419	302
alternatively	21	36	41	49	3

F Visual Examples

Below we demonstrate qualitative examples of reasoning during training and evaluation.

Example of multi-step reasoning during training

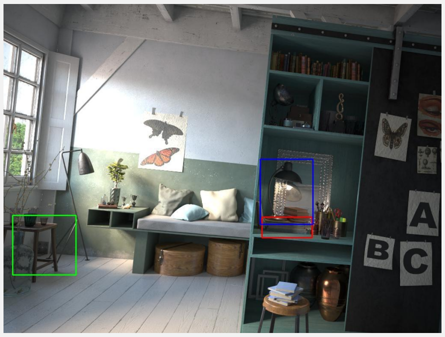


Question: How many Dressers are visible in the scene? Choose between the following options: 3, 0, 4, 1 or 2

Ground truth: 1

Response: Step 1: Identify the dresser in the room In the image, I can see a dresser on the left side. Step 2: Count the number of dressers The dresser is the only piece of furniture with the shape of a dresser in the image.
Conclusion: There is 1 dresser in the image. So, the answer is 1 dresser.</think> <answer>1</answer>

Example of thinking about alternatives on CVBench

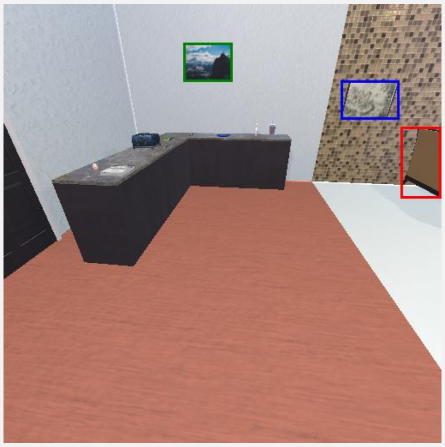


Question: Estimate the real-world distances between objects in this image. Which object is closer to the books (highlighted by a red box), the lamp (highlighted by a blue box) or the chair (highlighted by a green box)? (A) lamp (B) chair

Ground Truth: (A) lamp

Response: First, I will look at the positions of the objects highlighted by the boxes. The lamp is on the right side of the image, while the chair is on the left. The books are located in the middle of the image, so they are closer to the center than either the lamp or the chair.
Next, I will consider the size of the objects. The lamp is relatively small and placed on top of the bookshelf, while the chair is larger and placed on the floor. The books are also relatively small and placed on the stool.
Finally, I will take into account the perspective of the image. The camera is positioned at a higher angle, so the objects appear smaller and farther away than they actually are. However, the books are still closer to the center of the image than either the lamp or the chair.
Based on these observations, I conclude that the books are closer to the lamp than to the chair.</think>
So, the answer is (A) lamp. <answer>lamp</answer>,

Example of multi-step inconsistent reasoning during training



Question: Considering the relative positions, where is framed monochromatic wall decor painting (highlighted by a blue box) with respect to sofa (highlighted by a red box)? Choose between the following options: above or below

Ground truth: above

Response: The user asks for the position of the framed monochromatic wall decor painting with respect to the sofa. To answer this question, I need to know the location of both the sofa and the framed monochromatic wall decor painting. However, the location of the two objects is not provided. </think> <answer> above </answer> <reason> Since the location of both the sofa and the framed monochromatic wall decor painting is not provided, I cannot provide a detailed reasoning. However, I can provide an answer based on the information given: </reason> <answer> above </answer>

G Additional Analysis of Reflection-Indicative Keywords Across Baselines

Keyword Statistics for Baseline VLMs. Figure 3 in the main paper visualizes how reflection-indicative tokens (e.g., “wait,” “again”) dynamically emerge throughout RL training. Since this signal is collected *only during training rollouts*, equivalent mid-training curves cannot be constructed for public baseline models where intermediate checkpoints are not available. To provide a fair comparison, we evaluate each baseline model on the *same SAT prompts used in our RL rollouts* and report the keyword frequencies aggregated over fixed step intervals. Although these static statistics cannot reproduce the temporal emergence pattern seen in Figure 3, they reveal the extent to which different models naturally produce reflection-like reasoning behaviors. Among the four evaluated baselines, only **OpenVLThinker-3B** exhibits reflection keyword frequencies comparable to our model. The remaining models show minimal presence of such markers, suggesting limited spontaneous reflective reasoning. Below we present the full quantitative results (Tables 14–17), as well as the corresponding plots in Figures 12–15.

H Use of Large Language Models

We used Claude (Anthropic)[1] and ChatGPT (OpenAI)[2] as general-purpose editing tools for minor polishing of the manuscript. Specifically, the LLMs were used to:

[1] https://claude.ai/

[2] https://openai.com/

Table 14: Keywords statistics of **R1-VL 2B** [36].

Interval	again	wait	but	however	hmm	alternatively	check
0–299	1	0	6	1	0	0	22
300–599	4	0	5	0	0	0	30
600–899	3	0	8	0	0	0	19
900–1199	2	0	5	1	0	0	26
1200–1499	3	0	12	0	0	0	23

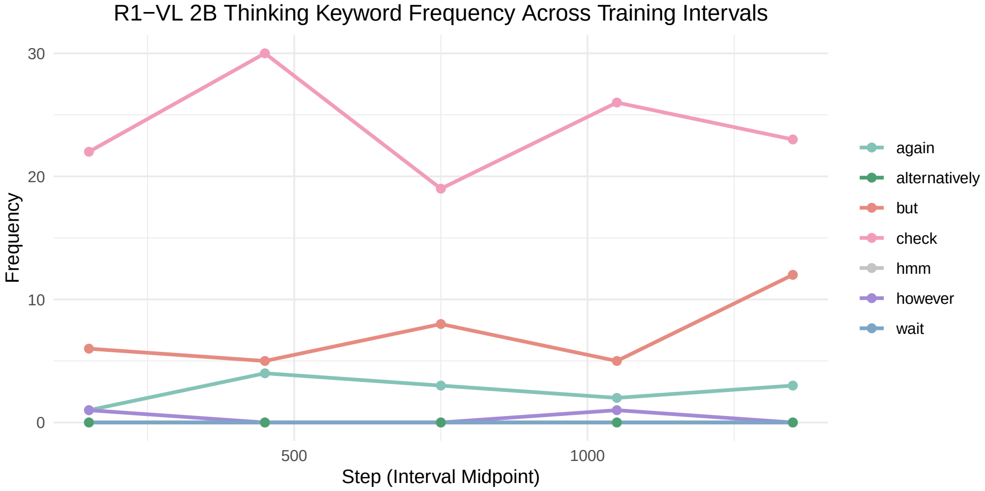


Figure 12: **Statistical analysis of reasoning-indicative keywords ("wait," "again," "but," "however," "so," "hmm," "check," "alternatively," "mistake") across evaluation intervals for baseline model R1-VL-2B [36].** The distribution highlights the degree to which reflection and refinement behaviors appear without RL fine-tuning.

- Improve grammar and sentence structure
- Enhance clarity of expression in selected passages
- Check for consistency and flow in expression

The LLMs did not contribute to research ideation, experimental design, analysis, or the generation of scientific content. All scientific claims, methodology, and core arguments are the authors' original work. The authors take full responsibility for all content in this submission.

Table 15: Keywords statistics of **VLAA-Thinker-Qwen2VL-2B** [1].

Interval	again	wait	but	however	hmm	alternatively	check
0–299	0	0	9	3	0	0	86
300–599	0	0	9	0	0	0	80
600–899	0	0	7	2	0	0	73
900–1199	0	0	5	1	0	0	84
1200–1499	0	0	2	0	0	0	86

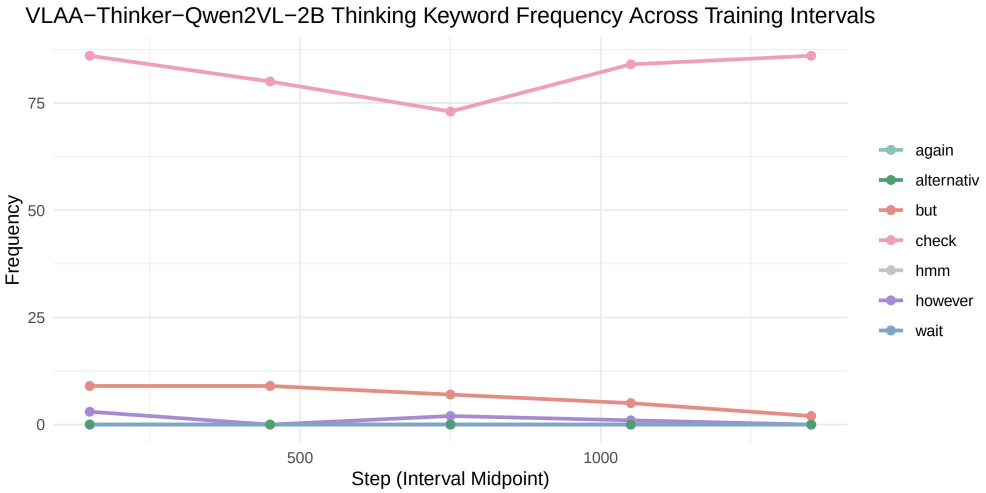


Figure 13: **Statistical analysis of reasoning-indicative keywords (“wait,” “again,” “but,” “however,” “so,” “hmm,” “check,” “alternatively,” “mistake”) across evaluation intervals for baseline model VLAA-Thinker-Qwen2VL-2B [1].** The distribution highlights the degree to which reflection and refinement behaviors appear without RL fine-tuning.

Table 16: Keywords statistics of **3B-Curr-ReFT** [6].

Interval	again	wait	but	however	hmm	alternatively	check
0–299	0	0	0	0	0	0	0
300–599	0	0	0	0	0	0	0
600–899	0	0	0	0	0	0	0
900–1199	0	0	1	0	0	0	0
1200–1499	0	0	2	0	0	0	0

Table 17: Keywords statistics of **OpenVLThinker-3B** [7].

Interval	again	wait	but	however	hmm	alternatively	check
0–299	27	1	264	12	35	0	27
300–599	24	0	255	8	23	0	24
600–899	21	1	335	11	28	0	34
900–1199	26	1	320	14	32	0	18
1200–1499	25	2	298	10	28	0	24

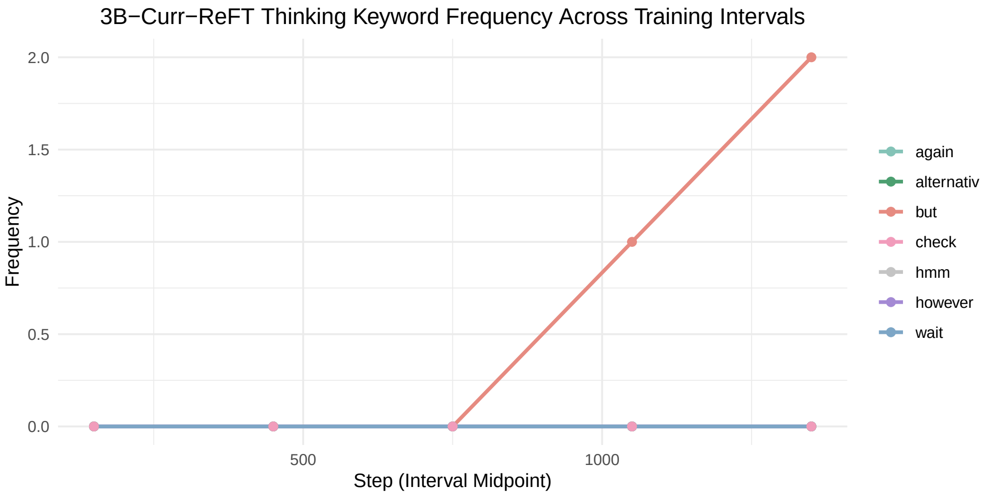


Figure 14: **Statistical analysis of reasoning-indicative keywords (“wait,” “again,” “but,” “however,” “so,” “hmm,” “check,” “alternatively,” “mistake”) across evaluation intervals for baseline model 3B-Curr-ReFT [6].** The distribution highlights the degree to which reflection and refinement behaviors appear without RL fine-tuning.

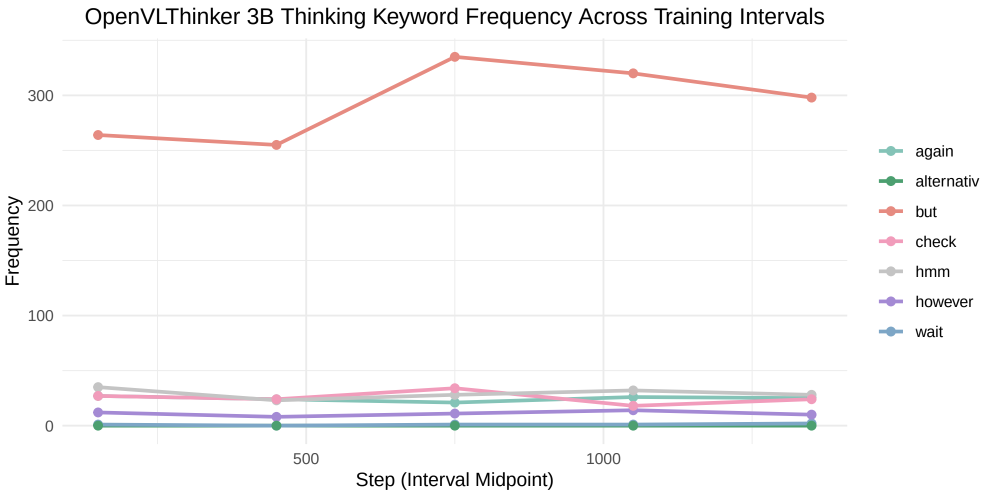


Figure 15: **Statistical analysis of reasoning-indicative keywords (“wait,” “again,” “but,” “however,” “so,” “hmm,” “check,” “alternatively,” “mistake”) across evaluation intervals for baseline model OpenVLThinker-3B [7].** The distribution highlights the degree to which reflection and refinement behaviors appear without RL fine-tuning.