



7-4-2025

Open Vision Reasoner: Transferring Linguistic Cognitive Behavior for Visual Reasoning

Yana Wei^{1*}, Liang Zhao^{2*†}, Jianjian Sun^{2*}, Kangheng Lin³, Jisheng Yin⁴, Jingcheng Hu⁵, Yinmin Zhang², En Yu⁶, Haoran Lv², Zejia Weng², Jia Wang², Chunrui Han², Yuang Peng⁵, Qi Han², Zheng Ge², Xiangyu Zhang², Daxin Jiang², Vishal M. Patel^{1†}

¹Johns Hopkins University ²StepFun ³BUPT ⁴UCAS ⁵THU ⁶HUST

Abstract

The remarkable reasoning capability of Large Language Models (LLMs) stems from cognitive behaviors that emerge when reinforcing against verifiable rewards. This work investigates how to transfer this principle to Multimodal LLMs (MLLMs) to unlock advanced visual reasoning. We introduce a two-stage paradigm built on Qwen2.5-VL-7B: a massive text-only cold-start fine-tuning, followed by multimodal reinforcement learning (RL) spanning nearly 1,000 steps—surpassing all prior open-source efforts in scale. This pioneering work reveals three fundamental insights: 1) Behavior transfer emerges surprisingly early in cold start due to linguistic mental imagery. 2) Cold start broadly memorizes visual behaviors, while RL critically discerns and scales up effective patterns. 3) Transfer strategically favors high-utility behaviors such as visual reflection. Our resulting model, Open-Vision-Reasoner (OVR), achieves state-of-the-art performance on a suite of reasoning benchmarks, including 95.3% on MATH500, 51.8% on MathVision and 54.6% on MathVerse. We release our model, data, and training dynamics to catalyze the development of more capable, behavior-aligned multimodal reasoners.

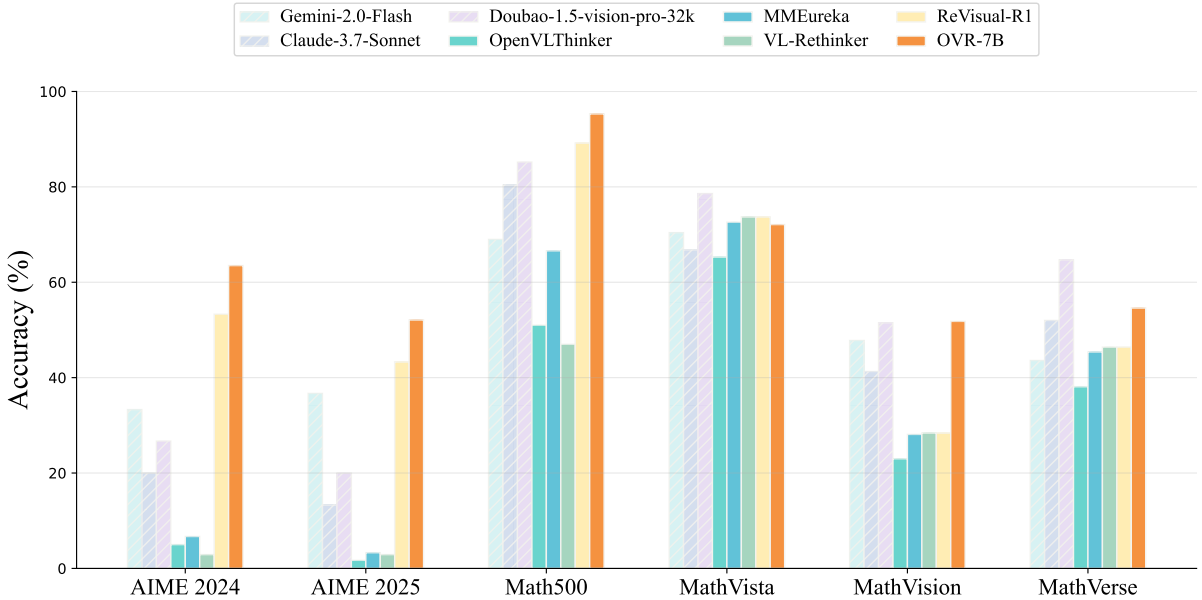


Figure 1: **Performance comparison** with state-of-the-art models on both textual (AIME 2024, AIME 2025 [1], MATH500 [2]) and multimodal (MathVista [3], MathVision [4], MathVerse [5]) math reasoning benchmarks. **Open Vision Reasoner (OVR)** demonstrates superior results among open-source models and performs competitively with commercial counterparts.

* Core contribution

† Corresponding authors: zhaoliang02@stepfun.com, vpatel36@jhu.edu

1. Introduction

"The eye sees only what the mind is prepared to comprehend." — Robertson Davies

Shifting Reinforcement Learning from Human Feedback (RLHF) [6] to Reinforcement Learning from Verifiable Reward (RLVR) [7, 8] has endowed LLMs [9, 7] with unexpectedly powerful reasoning across mathematics, code, and general problem-solving. At its core, verifiable reward—where correctness is determined by objective, often rule-based criteria—is inherently less susceptible to "reward hacking" [10, 11] than a learned reward model. This robustness proves instrumental in large-scale RL, enabling the internalization and activation of what recent studies [12, 13, 14] term **cognitive behaviors**—patterns like backtracking and subgoal decomposition that are empirically crucial for advanced reasoning.

The multimodal domain, inherently grounded in verifiable visual facts [15, 16], is uniquely suited for this paradigm. Yet, early multimodal RL efforts paradoxically adopted RLHF, relying on learned reward models to approximate objective correctness [17, 18, 19]. Inspired by the success of RLVR in language models, recent efforts have started exploring rule-based rewards in the multimodal setting. Perception-R1 [16] incorporates supervisions such as IoU and Euclidean distance to enhance the perceptual alignment of MLLMs, while works such as R1-OneVision [20] and VLAA-Thinking [21] construct behavior-rich visual reasoning trajectories through complex pipelines including iterative distillation and synthesizing. Recently, ReVisual-R1 [22] adopts a language-only cold start, relying on purely linguistic reasoning as a foundation without engaging with the visual modality.

Despite this encouraging progress, these approaches still leave a foundational question unanswered: **How can linguistic cognitive behaviors transfer to MLLMs for advanced visual reasoning?** To address this, we build upon the "RL with a cold start" paradigm [7] by conducting large-scale training on Qwen2.5-VL-7B [23], establishing it as a powerful testbed to systematically analyze how such behaviors emerge and scale in the multimodal domain.

To this end, we introduce a robust two-stage methodology designed to first instill linguistic cognitive patterns and then activate them for visual reasoning. Our process begins with a large-scale cold start, fine-tuning Qwen2.5-VL-7B on over 2 *million* examples to build a strong foundation. This is followed by a prolonged reinforcement learning phase under the *Open-Reasoner-Zero* [24] framework, leveraging over 0.3 *million* mixed-modality examples. To the best of our knowledge, this represents the largest open-source RL practice on this model. The resulting model, **Open-Vision-Reasoner** (OVR), validates our approach by achieving strong performance across both language and multimodal benchmarks. As shown in Fig. 1, it achieves **63.5%** on AIME2024 and **95.3%** on MATH500 for math reasoning, as well as **51.8%** on MathVision and **54.6%** on MathVerse for visual reasoning.

To further trace the transfer and evolution of cognitive patterns throughout training, we develop a in-depth **visual cognitive behavior analysis**. Three central insights are worth highlighting: (1) Behavior transfer emerges remarkably early in cold start, driven by linguistic patterns encoding *mental imagery* [25, 26] as illustrated in Fig. 11. (2) Cold start broadly *memorizes* diverse visual cognitive behaviors, while RL critically *discerns* and scales up effective patterns. (3) Transfer follows a *strategic* path, favoring behaviors with high utility such as visual reflection. These findings deepen the understanding on visual intelligence scaffolded by linguistic reasoning [27].

We further examine how this paradigm impacts a foundational capability of MLLMs—**visual perception**. While linguistic cold start introduces perceptual degradation, our study shows that

Table 1: **Visual Cognitive Behaviors and Linguistic Counterparts.** We define four key visual cognitive behaviors, providing formal definitions, illustrative examples, and their corresponding linguistic counterparts.

Visual Behavior	Example	Definition	Linguistic Counterpart
Visual Reflection	“Let me see the image again.”	The model explicitly revisits the image after identifying a potential mistake or inconsistency in its reasoning, indicating an effort to correct course.	Backtracking
Visual Divide-and-Conquer	“Let’s first look at the numbers on the left.”	The model breaks down a complex visual problem into sub-components or regions, each addressed sequentially to reach the final answer.	Subgoal Setting
Visual Verification	“I will now verify this against the image.”	The model confirms that its intermediate conclusions are visually grounded by cross-referencing with the image before proceeding.	Verification
Goal-driven Visual Tracing	“To get this answer, I need to find an object that...”	The model starts from a desired visual conclusion and reasons backwards to identify relevant image evidence that supports it.	Backward Chaining

multimodal RL can effectively *recover* this loss. However, we also observe the *limited scalability* of RL when focused solely on perceptual tasks, as reward signals increase without corresponding growth in reasoning complexity (e.g., token length). This limitation motivates a more deliberate integration of diverse, primitive visual cognitive behaviors. Such efforts represent a promising direction toward unlocking the potential of more advanced RL frameworks—multi-turn or even agentic RL built upon visual manipulation and imagination.

In summary, this paper advances the field through the following **three key contributions**:

- We construct a two-stage training pipeline consisting of a linguistic cold start followed by large-scale multimodal RL, enabling effective transfer of cognitive behaviors in MLLMs.
- Our *Open Vision Reasoner*, the largest open-source RL practice on Qwen2.5-VL-7B, achieves superior performance on both linguistic and multimodal reasoning benchmarks.
- We conduct an in-depth analysis of visual cognitive behaviors in OVR and provide valuable insights into their transfer and evolution across training stages.

2. Cognitive Behavior Preliminaries

Recent studies have highlighted that the emergence of robust reasoning in LLMs is closely tied to the acquisition of certain *cognitive behaviors* [13]. These behaviors reflect structured internal reasoning patterns akin to human problem-solving. Key examples include: (1) Backtracking — revising a previously chosen strategy upon identifying inconsistencies (e.g., “This approach won’t work because...”), (2) Verification — checking intermediate steps or partial results, (3) Subgoal Setting — breaking down complex problems into manageable components (e.g., “First, we need to...”), and (4) Backward Chaining — reasoning from the desired outcome back to required inputs (e.g., “To get 75, we need a number divisible by...”). The four patterns form a kind of textual inner monologue that emerges naturally in language-based reasoning.

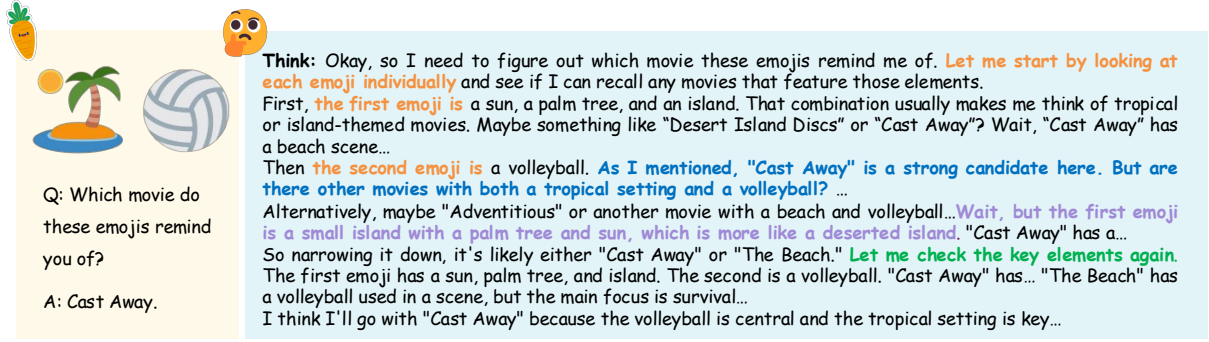


Figure 2: **Multiple Cognitive Behaviors in a Single Response.** This case shows triggered visual-specific cognitive behaviors like **visual divide-and-conquer**, **reflection**, **goal-driven visual tracing**, along with the linguistic behavior **backtracking**.

Based on this, we investigate the *transfer* of cognitive behaviors from language to vision. We define the visual extensions of the aforementioned behaviors—visual reflection, divide-and-conquer, visual verification, and goal-driven visual tracing. Their formal definitions, examples, and corresponding linguistic counterparts are provided in Table 1, while Fig. 2 presents a multimodal example encompassing both linguistic and visual cognitive behaviors. In the next section, we present a simple yet effective MLLM training pipeline comprising a linguistic cold start followed by multimodal RL, and systematically analyze the transfer and scaling of these visual cognitive behaviors in Section 5.2.

3. Open-Vision-Reasoner

In this section, we introduce **Open-Vision-Reasoner (OVR)**, a strong multimodal reasoning model build from Qwen2.5-VL-7B [28], from perspectives of training pipeline (Section 3.1), RL algorithm (Section 3.2) and data construction (Section 3.3).

3.1. Training Pipeline

To facilitate efficient cognitive development and cross-modal generalization, we employ the popular "RL with a cold start" paradigm [7] with two sequential training stages:

- **Stage 1: Linguistic Cold Start.** The LLM module is supervised fine-tuned on language-only reasoning datasets distilled from DeepSeek-R1 [7], establishing core cognitive behaviors such as backtracking and subgoal decomposition within a purely linguistic setting.
- **Stage 2: Multimodal RL.** We apply reinforcement learning with Open-Reasoner-Zero [24] setting on both text and multimodal tasks using verifiable match rewards. This promotes reasoning generalization and aligns previously learned cognitive patterns with visual contexts, enabling effective cross-modal transfer.

3.2. RL Algorithm

For the RL stage of our training pipeline, we adopt a lightweight Proximal Policy Optimization (PPO) [29] with Generalized Advantage Estimation (GAE) [30], following the policy and reward design used in *Open-Reasoner-Zero* [24]. We detail the RL for multimodal tasks below:

Proximal Policy Optimization For each input, consisting of an image I and a textual prompt q , the policy network π_θ generates n responses $\{o_1, \dots, o_n\}$. Each response o_i is a trajectory $\tau_i = (s_0^{(i)}, a_0^{(i)}, \dots, s_{T_i-1}^{(i)}, a_{T_i-1}^{(i)})$ of length T_i . The state $s_t^{(i)}$ includes q (and potentially encoded I features) and previously generated tokens; $a_t^{(i)}$ is the token generated at step t . A reward $r_t^{(i)}$ is computed at each timestep t of trajectory τ_i .

We use GAE to balance bias and variance in advantage estimation. The advantage $\hat{A}_t^{(i)}$ for state-action pair $(s_t^{(i)}, a_t^{(i)})$ in trajectory τ_i is:

$$\hat{A}_t = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \delta_{t+l}, \quad \text{where } \delta_{t'} = r_{t'} + \gamma V_\phi(s_{t'+1}) - V_\phi(s_{t'}). \quad (1)$$

V_ϕ is the value function, γ, λ are discount and GAE factors, and $V_\phi(s_T) = 0$ for terminal states. π_θ is updated by maximizing $\mathcal{J}_{\text{PPO}}(\theta)$ using experiences (s_t, a_t, \hat{A}_t) sampled under an older policy π_{old} :

$$\mathcal{J}_{\text{PPO}}(\theta) = \hat{\mathbb{E}}_{\pi_{\text{old}}} \left[\min(\rho_t(\theta) \hat{A}_t, \text{clip}(\rho_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t) \right]. \quad (2)$$

Here, $\rho_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}$ and ϵ is a clipping parameter (e.g., 0.2). $\hat{\mathbb{E}}_{\pi_{\text{old}}}$ denotes the empirical average over samples from π_{old} . V_ϕ is trained by minimizing $\mathcal{J}_{\text{value}}(\phi)$ on samples from π_{old} , using the empirical discounted returns $R_t = \sum_{k=0}^{T-t-1} \gamma^k r_{t+k}$:

$$\mathcal{J}_{\text{value}}(\phi) = \hat{\mathbb{E}}_{\pi_{\text{old}}} \left[(V_\phi(s_t) - R_t)^2 \right]. \quad (3)$$

Reward Function. We adopt the minimalist rule-based reward design, which evaluates only the correctness of model outputs while ignoring formatting or stylistic preferences. Specifically, we extract the predicted answer encapsulated within `\boxed{\}` in the model’s output and compare it against the reference answer. A binary reward is assigned—1 for exact matches, and 0 otherwise—enabling a clear, scalable and unhackable reward signal for reinforcement learning.

3.3. Dataset Construction

To support cognitive transfer, we carefully curate datasets specifically tailored to each training stage, encompassing both language-only and multimodal domains.

Data Collection. We firstly broadly collect prompt-answer pairs to develop both language and multimodal reasoning skills across mathematical, scientific, and logical domains. For language-only scenarios, we utilize public benchmarks including AIME (up to 2023), MATH [2], Numina-Math [31], Tulu3 MATH [32], and OpenR1-Math-220k [33], and other open-source datasets. We also synthesize general logical problems via programmatic generation to further enrich reasoning diversity. Multimodal scenarios incorporate datasets covering geometry problem solving (Geometry3k [34], GeoQA [35], Geos [36]), visual discrimination (IconQA [37], Pixmo [38], ChartQA [39]), visual puzzles (PuzzleVQA [40], AlgoPuzzleVQA [41]), STEM (TQA [42], ScienceQA [43], K12 from [44]) and multimodal math (AtomThink [45], in-house curated math).

Data Curation. To refine data quality, we employ a multi-step curation process. *Firstly*, we employ a pre-trained model to automatically filter out samples with high training loss, which typically indicate noise or excessive complexity. *Secondly*, rule-based and model-assisted methods then identify and remove undesirable patterns [31]. *Thirdly*, we apply reweighting

Table 2: **Comparison on Language Reasoning and General Benchmarks.** Best results are **bold** and the second-best are underlined for *open-source models*. [†] indicates metrics reproduced by ourselves.

Model	AIME 2024	AIME 2025	MATH500	GPQA Diamond	MMLU	MMLU-Pro
<i>Open-source Models</i>						
Qwen2.5-7B [28]	6.7 [†]	6.7 [†]	77.6 [†]	32.8 [†]	72.6 [†]	57.5 [†]
Qwen2.5-VL-7B [28]	6.7 [†]	6.7 [†]	67.4 [†]	31.8 [†]	69.6 [†]	51.7 [†]
Open-Reasoner-Zero-7B [24]	17.9	15.6	81.4	36.6	-	-
DeepSeek-R1-Distill-Qwen-7B [7]	55.5	39.2 [†]	92.8	49.1	-	-
QwQ-32B-Preview [46]	50.0	33.5	90.6	<u>54.5</u>	-	-
Skywork-R1V-38B [47]	72.0	-	94.0	61.6	-	-
ReVisual-R1 [22]	53.3	<u>43.3</u>	89.2	47.5	-	-
<i>Close-source Models</i>						
Gemini-2.0-Flash [48]	33.4	36.7	69.0	35.4	-	-
OpenAI-o1-mini [9]	63.6	-	90.0	60.0	85.2	80.3
Claude 3.7 Sonnet [49]	20.0	13.3	80.4	61.1	-	80.0
Doubao-1.5-vision-pro-32k [50]	26.7	20.0	85.2	56.1	-	-
OVR-7B	<u>63.5</u>	52.1	95.3	49.8	77.2	67.9

to balance coverage, down-weighting overrepresented categories while emphasizing rare but valuable instances. To the end, we distill responses from DeepSeek-R1 [7] to construct approximately 2 million cold-start data. To ensure the unhackability and stability during RL, we further exclude problems incompatible with our reward functions (e.g., proof-style questions) and apply difficulty-based heuristic filtering, removing both overly trivial and infeasible samples to ensure well-calibrated learning. This leaves around 300k multimodal RL data. Further details refer to the appendix.

4. Experiments

In this section, we first elaborate our implementation of *Open-Vision-Reasoner* (OVR). Then, we present superior performance across textual benchmarks (Section 4.2) and multimodal scenerios (Section 4.3).

4.1. Implementation Details

Our model is based on Qwen2.5-VL-7B [23] and employs a two-stage training strategy. In the first stage of cold start, we independently fine-tune the LLM module for 5 epochs with a batch size of 640, a sequence length of 64k, and a learning rate of 2×10^{-4} leveraging the default Qwen2.5 configuration [28]. During the subsequent stage of reinforcement learning, following Open-Reasoner-Zero [24], we utilize PPO and configure GAE with $\gamma = 1$ and $\lambda = 1$ to fully capture long-term dependencies crucial for reasoning tasks, enabling stable training. This RL phase proceeds for 900 iterations, during which we adopt a curriculum for the sequence length: it begins at 24k for the first 300 iterations, increases to 32k through iteration 700, and expands to 48k thereafter, with our latest models continuously undergoing this refinement process. We adhere to strict on-policy updates for the policy model and undertake multiple optimization steps for the critic model. Please note that our final model is an **uniform average of several representative intermediate checkpoints**, ensuring balanced and robust performance across various benchmarks. Additional details can be found in the appendix.

Table 3: **Evaluation Results on Visual Reasoning Benchmarks.** Best results are **bold** and the second-best are underlined for *open-source models*. [†] Indicates results reproduced by ourselves.

Model	MathVista	MathVision	MathVerse vision-only	DynaMath worst	WeMath		LogicVista	MMMU-Pro	CharXiv	
					strict	loose			reas.	desc.
<i>SFT Methods</i>										
LLaVA-OneVision-7B [51]	62.6	17.6	17.6	9.0	17.7	-	32.0	24.1	23.6	48.7
InternLM-XComposer2.5 [52]	64.0	17.8	16.2	8.2	14.1	-	34.7	-	-	-
InternVL3-8B [53]	70.5	28.6	33.9	23.0	37.5	-	43.6	-	37.6	73.6
InternVL2.5-8B [54]	64.5	17.0	22.8	9.4	23.5	-	36.0	34.3	32.9	<u>68.6</u>
InternVL2-8B [55]	58.3	20.0	20.4	9.2	20.2	-	33.6	29.0	-	-
Qwen2-VL-7B [56]	61.6	19.2	25.4	11.0	22.3	-	33.3	30.5	34.6	58.0
Qwen2.5-VL-7B [28]	69.2 [†]	25.5 [†]	41.1	21.8	31.2 [†]	53.1 [†]	47.9	-	<u>36.4[†]</u>	67.3 [†]
QvQ-72B-Preview [57]	70.3	34.9	48.2	30.7	39.0	-	58.2	-	-	-
Kimi-VL-16B [58]	66.0	21.8	34.1	18.0	32.3	-	42.7	-	-	-
<i>Close-source Models</i>										
Gemini-2.0-Flash [48]	70.4	47.8	43.6	42.1	47.4	-	52.3	-	-	-
OpenAI-GPT-4o [59]	59.9	31.1	40.6	34.5	42.9	-	64.4	-	-	-
Claude 3.7 Sonnet [49]	66.8	41.3	52.0	39.7	58.2	-	49.3	-	-	-
GPT-4o mini [60]	55.1 [†]	27.3 [†]	30.0 [†]	31.6 [†]	31.4 [†]	48.8 [†]	41.4 [†]	37.6 [†]	34.10 [†]	74.92 [†]
doubao-1.5-vision-pro-32k [50]	78.6	51.5	64.7	44.9	64.2	-	65.7	-	-	-
<i>RL-based Methods</i>										
VLAA-Thinker-Qwen2-7B [21]	59.6	19.8	33.9	15.2	30.5	-	36.0	-	-	-
VLAA-Thinker-Qwen2.5-7B [21]	68.0	26.4	48.2	22.4	41.5	-	48.5	-	-	-
R1-Onevision-7B [20]	64.1	29.9	40.0	-	-	<u>61.8</u>	-	-	-	-
OpenVLThinker-7B [61]	65.3	23.0	38.1	16.8	35.2	-	44.5	-	-	-
MM-Eureka-Qwen-7B [44]	72.6	28.1	45.4	23.0	21.8	-	46.3	-	-	-
MMR1-Math-v0 [62]	69.8	30.7	42.8	17.4	31.9	-	46.8	-	-	-
ThinkLite-7B-VL [63]	71.6	24.6	42.9	16.5	41.8	-	42.7	-	-	-
R1-VL-7B [64]	63.5	24.7	40.0	-	-	-	-	-	-	-
X-REASONER [65]	69.0	29.6	-	-	-	-	-	<u>43.0</u>	-	-
VL-Rethinker-7B [66]	73.7	28.4	46.4	17.8	36.3	-	42.7	<u>41.7</u>	-	-
ReVisual-R1 [22]	73.1	48.8	53.6	27.5	42.0	-	52.3	-	-	-
WeThink [67]	<u>70.9</u>	<u>27.2</u>	<u>44.7</u>	24.4	48.0	-	53.0	-	-	-
Skywork-R1V-38B [47]	60.6	42.1	40.4	-	34.1	-	50.6	-	-	-
OVR-7B	72.1	51.8	54.6	33.5	44.6	64.8	54.8	50.2	44.5	73.6

4.2. Enhanced Language Reasoning and General Capabilities

Our model is initially evaluated on a variety of language benchmarks, which cover mathematical reasoning and general problem-solving skills. Specifically, we include *AIME 2024*, *AIME 2025* [1], *MATH500* [2], *GPQA Diamond* [68], *MMLU* [69], and *MMLU-Pro* [70]. We compare *Open-Vision-Reasoner* (OVR) with strong LLM baselines, including *Qwen2.5-7B* [71], *DeepSeek-R1-Distill-Qwen-7B* [7] and *Open-Reasoner-Zero-7B* [24].

our model demonstrates exceptional reasoning capabilities. On the challenging AIME 2024 and 2025 benchmarks, it dramatically **surpasses other 7B open-source models by an average of over 10%**, achieving performance comparable to leading 32B models. This superiority extends to general reasoning tasks, with significant gains of **+4.6%** on MMLU and **+10.4%** on MMLU-Pro over parameter-matched competitors. These results highlight the effectiveness of our curated, high-quality cold-start training data.

4.3. Superior Visual Reasoning Abilities

To evaluate whether the introduced cognitive behavior transfer leads to cross-modal benefits, we further assess the model on a suite of multimodal reasoning benchmarks. These tasks involve image-grounded mathematical reasoning, general multimodal reasoning, and chart understanding. Specifically, we include *MathVista* [3], *MathVision* [4], *MathVerse* [5], *DynaMath* [72], *WeMath* [73], *LogicVista* [74], *MMMU-Pro* [75], and *CharXiv* [76] for evaluation. We compare our model against strong MLLM baselines, including SFT-based methods such as *LLaVA-OneVision* [77] and *Qwen2.5-VL* [23], as well as recent rule-based RL methods like *Open-*

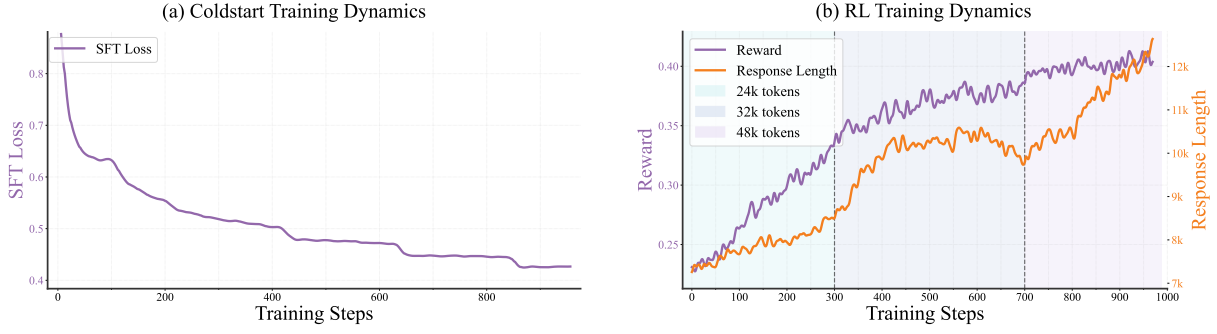


Figure 3: **Training Dynamics.** (a) The cold-start stage shows a step-wise loss decrease. (b) In the RL stage, reward (purple, left axis) and average response length (orange, right axis) grow steadily, with sharp surges after each sequence length expansion.

VLThinker [61], *MM-Eureka* [44] and *ReVisual-R1* [22].

As shown in Table 3, our model sets a new benchmark for 7B models in visual reasoning. It is **the first model based on Qwen2.5-VL-7B [23] post-training to surpass the 50% mark on MathVision**, while also achieving top 7B performance on DynaMath and MathVerse. This overall strong performance is further highlighted by its significant margins over previous SOTA methods on MMMU-Pro (+7.2%). These results indicate that reasoning capabilities acquired through language training can effectively transfer to multimodal tasks, leading to measurable improvements in visual reasoning.

5. Discussion

5.1. Analysis of Training Dynamics

In this section, we present a comprehensive overview of the training dynamics as illustrated in Fig. 3, and provide a detailed analysis of how text and multi-modal reasoning metrics evolve throughout the process as shown in Fig. 4.

During the initial cold-start phase (Fig. 3 (a)), the model’s loss rapidly descends to below 0.5. Subsequently, across multiple training epochs, the loss exhibits a step-wise, gradual decrease. In parallel, we observe a corresponding surge in performance across all benchmarks (Fig. 4), which first ascend sharply before transitioning to a phase of slower, more incremental improvement toward their peak. A noteworthy observation is that the *aggressive* training strategy detailed in Section 4.1—employing a large batch size in concert with a high learning rate—proves to be essential. This approach is critical for breaking the model’s inherent constraints, thereby successfully imbuing it with new cognitive paradigms and **sculpting a more favorable landscape for reinforcement learning**. It is a prerequisite that enables our model, which originates from an instruction-tuned base, to ultimately achieve text performance that is comparable to, or even surpasses models initialized from base [28] or math-specific checkpoints [78].

Furthermore, Fig. 3 (b) reveals how the model’s reward and average token length in the RL phase steadily advance from an initial 7k to exceed 12k. Owing to the stability of the training configuration inherited from *Open-Reasoner-Zero* [24], OVR is successfully trained on a diverse corpus of over 20 multi-modal and language-only datasets without encountering any training collapse or performance degradation. Critically, whenever the token length begins to plateau or even decline, we strategically switch to a longer context length, which invariably catalyzes the

next wave of rapid reward growth. Fig. 4 captures the coincident yet unsurprising convergent growth trajectory shared by all eight reasoning benchmarks, spanning both text and multi-modal domains, as they progressively ascend towards their zenith amidst fluctuations.

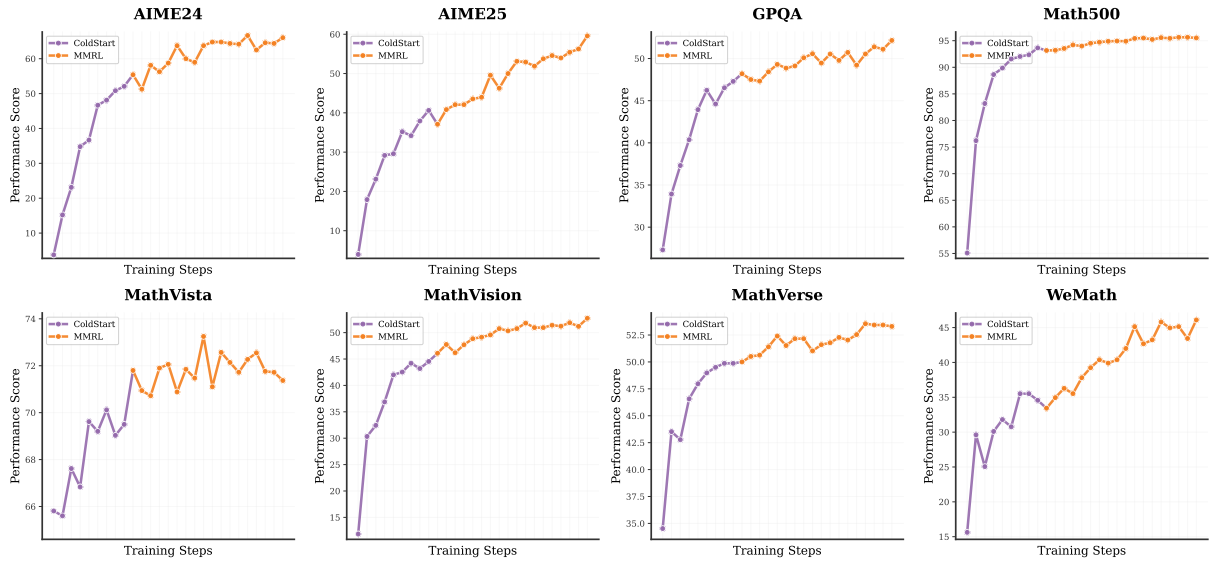


Figure 4: **Performance Evolution on Reasoning Benchmarks.** OVR demonstrates sustained and convergent growth across both linguistic and multi-modal benchmarks throughout the cold start (left) and RL (right) training phases.

5.2. Multimodal Cognitive Behavior Analysis

Recent studies have highlighted the emergence of cognitive behaviors in MLLMs during visual reasoning tasks—phenomena often dubbed “visual aha moments” [79, 21]. In this work, we move beyond plain observations and systematically investigate how these behaviors are transferred from their linguistic counterparts. Our analysis centers on the four pivotal visual cognitive behaviors introduced in Section 2 which are drawn from foundational research on cognitive patterns [13]. To quantify this process, we employ GPT-4o [60] to analyze the emergence of each behavior within the inference traces of our OVR model.

Visual behaviors emerge remarkably early from cold start. Following Section 5.1, we tracked the dynamics of visual reflection, a significant behavior mentioned in previous studies [80, 81], throughout OVR’s training. As depicted in Fig. 5, this vision-specific behavior emerges in significant quantities from the *very beginning* of the cold-start phase and fluctuates throughout subsequent training steps. Strikingly, we observed that even in linguistic problems, DeepSeek-R1’s responses [7] frequently exhibited signs of *mental imagery* [25, 26] as shown in Fig. 11(a). The model appeared to construct internal visualizations to support mathematical reasoning, often articulated through phrases such as “let me visualize...” or “let me see the image.” Once this linguistic scaffolding was introduced into our MLLM, these *mental images* were rapidly grounded in actual visual input, enabling their rapid and effective generalization within OVR.

Cold-start learns broadly, large-scale RL discerns critically. We further investigate how cognitive behaviors scale during large-scale RL. As shown in Fig. 5(a), after an initial, rapid instillation

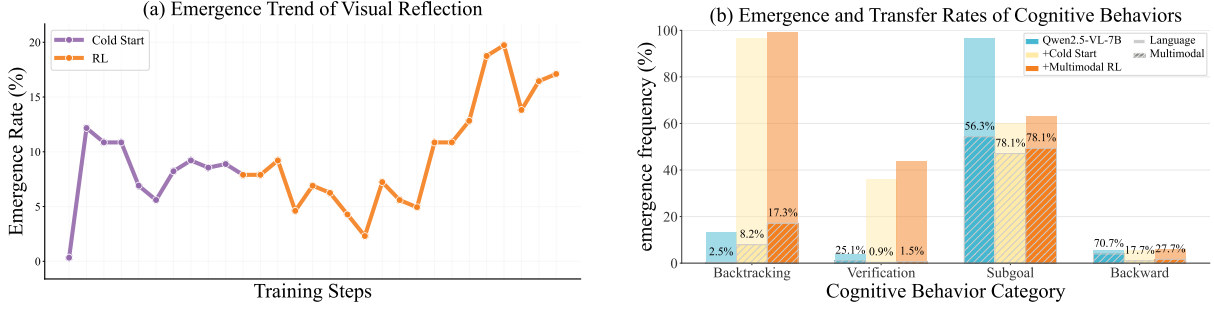


Figure 5: **Multimodal Cognitive Behavior Analysis.** (a) Emergence of visual reflection across the cold start and RL training steps. (b) Emergence and transfer rates of four visual cognitive behaviors across base models and training stages. Numerical values denote the language-to-vision transfer rates for each behavior.

of patterns during the aggressive cold-start phase, their prevalence is *first suppressed then amplified to unprecedented levels* during multimodal RL. This counter-intuitive dynamic suggests a clear division of labor: the cold-start phase learns broadly, indiscriminately memorizing all available patterns. In contrast, RL discerns critically, acting as a strategic filter for the crucial tokens [82] and scaling up pivotal behaviors. This process of RL—**discarding the dross to select the essence**—is significant for achieving superior generalization.

Visual transfer of cognitive behaviors is strategic. To analyze the transition from linguistic to visual cognition, we track the emergence and transfer rates (detailed in Appendix C.1) of four core cognitive behaviors across both language and vision modalities. As shown in Fig. 5(b), the emergence of backtracking and verification steadily increases across training stages, underscoring their growing importance. Among these, the transfer rate of backtracking shows consistent growth—from 2.5% to 17.3%—while verification exhibits near-zero transfer throughout both the cold-start and RL phases. This indicates that transfer is a *strategic* process, for which we posit two potential explanations: (1) Backtracking transfers more readily due to DeepSeek-R1’s [7] inherent “mental imagination” capabilities, while verification, lacking a direct linguistic precursor, is more difficult for the MLLM to internalize. (2) Mirroring how humans naturally and instinctively process visual information [80], backtracking is a more *fundamental* component of complex visual reasoning, making its amplification a higher priority during the strategic RL phase. We will investigate these hypotheses in greater depth in our future work.

5.3. Beyond Behavior: Visual Perception Analysis and Future Work

Beyond behavioral dynamics, we extend our discussion to a essential capability of MLLMs: **visual perception** under the cold start plus large-scale RL paradigm. In particular, we investigate two key areas of interest—*perceptual hallucination* and *scaling properties*—through a dedicated study on our OVR model.

Cold start impairs perception, while RL enhances. We evaluated both stages of OVR, along with the base model Qwen2.5-VL-7B, on a comprehensive set of multimodal benchmarks targeting visual perception and recognition (MMBench [83], BLINK [84], MMStar [85], HallusionBench [86], POPE [87], RealWorldQA [88], PhyX [89], MME [90], MMVet [91]). As shown in Table 4, performance steadily improves across tasks such as MMBench and PhyX, under-

Table 4: Model Performance on Perception-centric Benchmarks.

Model	MMBench		BLINK	MMStar	HallusionBench	POPE	RealWorldQA	PhyX	MME	MMVet
	en	cn								
Qwen2.5-VL	85.3 [†]	84.6 [†]	53.7 [†]	62.1 [†]	49.0 [†]	86.1 [†]	69.3 [†]	37.5 [†]	1659.7 [†]	63.9 [†]
+Cold Start	86.1	82.1	51.5	62.4	55.0	82.5	63.1	47.7	1549.8	61.8
+Multimodal RL	86.6	84.2	54.1	62.7	53.6	83.2	65.5	50.0	1559.1	63.6

scoring the effectiveness of our training paradigm. The cold-start model shows declines on several tasks, notably increased *hallucinations* [92, 93], likely due to token distribution shifts from large-scale linguistic data [87]. However, the regained performance on benchmarks such as MMBench and BLINK demonstrate that long-term multimodal RL can effectively mitigate these issues by **discerning perceptual capabilities** that are critically for multimodal tasks. Looking ahead, degradation from cold start can be mitigated either by incorporating the linguistic data into model pretraining [94, 95], or by introducing more multimodal supervision during the cold start to establish a stronger visual foundation.

The current unscalability of RL for perception policy. Throughout the multimodal RL, we observed a strong correlation between the reward and the average response length in Fig. 3, a finding consistent with prior practices [7, 24]. This reinforces response length as an effective reward proxy, indicative of a *scaling property* tied to reasoning depth and computational resources. However, when focusing on specific discriminative perceptual tasks like OCR and counting, we observe a clear divergence. As shown in Fig. 6, while the reward can be effectively increased, the average response length remains largely stagnant.

This unscalable training dynamic on such challenging tasks hints at a more fundamental issue: *the absence of certain core visual cognitive behaviors*. Addressing this *fundamental capability gap* is paramount for achieving robust multimodal scaling. Emerging research offers promising avenues, such as multi-turn RL with agentic *tool-use* (e.g., OpenAI-o3 [96]) and the integration of intrinsic imagining through *mental images* [97, 98]. These approaches hold the potential to bridge current limitations and unlock more scalable multimodal reasoning.

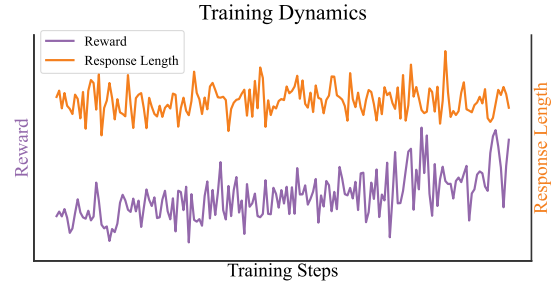


Figure 6: Training Dynamics on perception tasks including OCR and counting.

6. Related Work

Recent breakthroughs like OpenAI’s o1 [9] have highlighted the power of RL in unlocking and scaling reasoning capabilities [99, 100, 101] within LLMs. DeepSeek-R1-Zero [7] showed that reasoning capabilities can emerge purely through large-scale RL, leading to complex behaviors like self-verification and reflection. Open-source efforts like Open-Reasoner-Zero [24] further demonstrates that even minimalist RL approaches, such as vanilla PPO [29] with GAE [30] and simple rule-based rewards, can drive scaling in response length and benchmark performance on open-source models [28, 78].

MLLMs [102, 56, 23] have rapidly progressed from basic image captioning [103, 104] to more challenging reasoning tasks [105, 106, 107, 108]. Early efforts primarily relied on supervised

fine-tuning with Chain-of-Thought (CoT) datasets [109], while some explored explicit reflection [80] and self-correction [110] mechanisms to emulate human-like reasoning patterns. More recently, methods such as PerPO [18] and MDPO [17] adopt RL-based post-training approaches like DPO [111], where alignment is learned from paired positive/negative responses. These approaches generally follow the RL from Human Feedback (RLHF)[6, 12] or RL from AI Feedback (RLAIF) [112] paradigms, where signals from learned reward models or preference labels are utilized for optimization.

Inspired by the success of RLVR [7] in language models, MLLM research has shifted toward rule-based RL like GRPO [113] into the multimodal domain. This has led to two major lines of efforts: (1) designing task-specific reward objectives [16, 21, 33], and (2) constructing multimodal “thinking” datasets that embed cognitive behaviors within CoT sequences [114, 115, 61, 44]. Additionally, recent powerful MLLMs adopt a language-only cold start [22, 94], using verbal reasoning sequences as a foundation for subsequent multimodal learning. These approaches encourages human-like behaviors [13] or so-called “visual aha moments” in the model responses.

Despite these advances in MLLM, a fine-grained understanding of the underlying reasoning mechanisms remains less explored. In contrast, recent study [13] centered on LLMs posit that effective reasoning is causally linked to the model’s acquisition of certain *cognitive behaviors*, such as verification, backtracking, subgoal setting, and backward chaining. The test-time studies have observed that invoking these patterns improve performance [116]. Entropy-based analysis further reveals that regions associated with cognitive tokens are critical for diverse and high-quality reasoning [82]. The multimodal work like Long-Perceptual-Thoughts [117] attempts to explicitly instill such patterns by synthesizing long-form multimodal CoT data.

7. Conclusion

In this paper, we propose a two-stage training framework to investigate cognitive behavior in the multimodal domain. By combining a linguistic cold start followed by a large-scale multimodal RL, our approach enables effective cross-modal transfer and scaling of cognitive patterns. Our model *Open Vision Reasoner*, the largest open-source RL practice built on Qwen2.5-VL-7B, demonstrates strong performance across both linguistic and perceptual benchmarks. Beyond performance, we provide a systematic analysis of visual cognitive behaviors, revealing how they emerge and evolve through different training stages. We hope our findings inspire future research on cognitively aligned multimodal agents and open up new possibilities for scaling vision-language reasoning through behavior-centered learning.

References

- [1] Mislav Balunović, Jasper Dekoninck, and Martin Vechev Ivo Petrov, Nikola Jovanović. Matharena: Evaluating llms on uncontaminated math competitions, February 2025.
- [2] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*, 2021.
- [3] 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.
- [4] Ke Wang, Juntao Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.
- [5] 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.
- [6] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [7] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [8] Kimi Team. Kimi k1.5: Scaling reinforcement learning with llms. 2025.
- [9] OpenAI. Learning to reason with llms. <https://openai.com/index/learning-to-reason-with-llms/>, 2025.
- [10] Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krashenninikov, and David Krueger. Defining and characterizing reward hacking, 2025.
- [11] En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, et al. Unhackable temporal rewarding for scalable video mllms. *arXiv preprint arXiv:2502.12081*, 2025.
- [12] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024.
- [13] Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

- [14] Rosie Zhao, Alexandru Meterez, Sham Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. Echo chamber: RL post-training amplifies behaviors learned in pretraining. *arXiv preprint arXiv:2504.07912*, 2025.
- [15] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.
- [16] En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025.
- [17] Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*, 2024.
- [18] Zining Zhu, Liang Zhao, Kangheng Lin, Jinze Yang, En Yu, Chenglong Liu, Haoran Wei, Jianjian Sun, Zheng Ge, and Xiangyu Zhang. Perpo: Perceptual preference optimization via discriminative rewarding. *arXiv preprint arXiv:2502.04371*, 2025.
- [19] Ke Zhu, Liang Zhao, Zheng Ge, and Xiangyu Zhang. Self-supervised visual preference alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 291–300, 2024.
- [20] 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.
- [21] 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.
- [22] Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Wei-jie Wang, Xiaoye Qu, and Yu Cheng. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning, 2025.
- [23] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [24] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.
- [25] Philip N Johnson-Laird. Mental models in cognitive science. *Cognitive science*, 4(1):71–115, 1980.
- [26] Marilyn Ford. Mental models: towards a cognitive science of language, inference, and consciousness, 1985.
- [27] Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Virgo: A preliminary exploration on reproducing o1-like mllm, 2025.

- [28] Binyuan Hui, Jian Yang, Zeyu Cui, Jiayi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.
- [29] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [30] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.
- [31] Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. NuminaMath. [<https://huggingface.co/AI-MO/NuminaMath-CoT>] (https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.
- [32] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025.
- [33] Loubna Ben Allal, Lewis Tunstall, Anton Lozhkov, Elie Bakouch, Guilherme Penedo, and Gabriel Martín Blázquez Hynek Kydlicek. Open r1: Evaluating llms on uncontaminated math competitions, February 2025.
- [34] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021.
- [35] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*, 2021.
- [36] Liangyu Xu, Yingxiu Zhao, Jingyun Wang, Yingyao Wang, Bu Pi, Chen Wang, Mingliang Zhang, Jihao Gu, Xiang Li, Xiaoyong Zhu, et al. Geosense: Evaluating identification and application of geometric principles in multimodal reasoning. *arXiv preprint arXiv:2504.12597*, 2025.
- [37] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021.
- [38] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.
- [39] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

- [40] Yew Ken Chia, Vernon Toh Yan Han, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. *arXiv preprint arXiv:2403.13315*, 2024.
- [41] Deepanway Ghosal, Vernon Toh Yan Han, Chia Yew Ken, and Soujanya Poria. Are language models puzzle prodigies? algorithmic puzzles unveil serious challenges in multimodal reasoning. *arXiv preprint arXiv:2403.03864*, 2024.
- [42] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007, 2017.
- [43] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.
- [44] F Meng, L Du, Z Liu, Z Zhou, Q Lu, D Fu, B Shi, W Wang, J He, K Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.
- [45] Kun Xiang, Zhili Liu, Zihao Jiang, Yunshuang Nie, Runhui Huang, Haoxiang Fan, Hanhui Li, Weiran Huang, Yihan Zeng, Jianhua Han, et al. Atomthink: A slow thinking framework for multimodal mathematical reasoning. *arXiv preprint arXiv:2411.11930*, 2024.
- [46] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning. URL: <https://qwenlm.github.io/blog/qwq-32b>, 2025.
- [47] Yi Peng, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Li Ge, et al. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. *arXiv preprint arXiv:2504.05599*, 2025.
- [48] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [49] Anthropic. Claude. <https://www.anthropic.com/index/introducing-claude>, 2023.
- [50] Doubao AI. Doubao 1.5 vision pro 32k, 2025. Accessed: [2025-07-03].
- [51] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [52] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024.
- [53] 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.

- [54] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.
- [55] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.
- [56] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [57] Qwen Team. Qvq: To see the world with wisdom, 2024.
- [58] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.
- [59] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [60] GPT-4o. Hello gpt-4o, 2024.
- [61] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Open-vlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*, 2025.
- [62] 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. *arXiv preprint arXiv:2503.07536*, 2025.
- [63] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025.
- [64] Jingyi Zhang, Jiaying 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.
- [65] Qianchu Liu, Sheng Zhang, Guanghui Qin, Timothy Ossowski, Yu Gu, Ying Jin, Sid Kiblawi, Sam Preston, Mu Wei, Paul Vozila, et al. X-reasoner: Towards generalizable reasoning across modalities and domains. *arXiv preprint arXiv:2505.03981*, 2025.
- [66] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025.
- [67] Jie Yang, Feipeng Ma, Zitian Wang, Dacheng Yin, Kang Rong, Fengyun Rao, and Ruimao Zhang. Wethink: Toward general-purpose vision-language reasoning via reinforcement learning, 2025.

- [68] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.
- [69] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.
- [70] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems, NeurIPS 2024*, 2024.
- [71] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [72] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dyna-math: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models, 2024.
- [73] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multi-modal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.
- [74] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts, 2024.
- [75] 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, 2024.
- [76] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024.
- [77] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [78] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.
- [79] 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.
- [80] Yana Wei, Liang Zhao, Kangheng Lin, En Yu, Yuang Peng, Runpei Dong, Jianjian Sun, Haoran Wei, Zheng Ge, Xiangyu Zhang, et al. Perception in reflection. *arXiv preprint arXiv:2504.07165*, 2025.

- [81] Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. Training language models to self-correct via reinforcement learning, 2024.
- [82] Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective, 2025.
- [83] 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*, pages 216–233. Springer, 2024.
- [84] 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. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.
- [85] 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.
- [86] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.
- [87] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.
- [88] X.AI. Grok-2 beta release. <https://x.ai/blog/grok-2>, 2024. Accessed on: 2024-07-02.
- [89] Hui Shen, Taiqiang Wu, Qi Han, Yunta Hsieh, Jizhou Wang, Yuyue Zhang, Yuxin Cheng, Zijian Hao, Yuansheng Ni, Xin Wang, et al. Phyx: Does your model have the "wits" for physical reasoning? *arXiv preprint arXiv:2505.15929*, 2025.
- [90] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.
- [91] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.
- [92] Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. Are reasoning models more prone to hallucination? *arXiv preprint arXiv:2505.23646*, 2025.
- [93] Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *arXiv preprint arXiv:2505.21523*, 2025.

- [94] Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zihan Jiang, Zhixian Zheng, Zhichao Song, Zhenbo Luo, Yue Yu, Yudong Wang, Yuanyuan Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xinzhe Xu, Xingchen Song, Xing Zhang, Xing Yong, Xin Zhang, Xiangwei Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Weiwei Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shihua Yu, Shaohui Liu, Shande Wang, Rui Ma, Qiantong Wang, Peng Wang, Nuo Chen, Menghang Zhu, Kangyang Zhou, Kang Zhou, Kai Fang, Jun Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongshen Xu, Heng Qu, Haochen Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong Ma, Chang Liu, Can Cai, and Bingquan Xia. Mimo-vl technical report, 2025.
- [95] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.
- [96] OpenAI. Introducing openai o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>, 2025.
- [97] Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025.
- [98] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. *arXiv preprint arXiv:2506.21458*, 2025.
- [99] Mislav Balunovic, Jasper Dekoninck, Ivo Petrov, Nikola Jovanovic, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions, february 2025. URL <https://matharena.ai>.
- [100] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- [101] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.
- [102] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- [103] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [104] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, pages 408–424. Springer, 2024.
- [105] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

- [106] En Yu, Liang Zhao, Yana Wei, Jinrong Yang, Dongming Wu, Lingyu Kong, Haoran Wei, Tiancai Wang, Zheng Ge, Xiangyu Zhang, et al. Merlin: Empowering multimodal llms with foresight minds. In *European Conference on Computer Vision*, pages 425–443. Springer, 2024.
- [107] Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, et al. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. *arXiv preprint arXiv:2307.09474*, 2023.
- [108] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.
- [109] 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.
- [110] Jiayi He, Hehai Lin, Qingyun Wang, Yi Fung, and Heng Ji. Self-correction is more than refinement: A learning framework for visual and language reasoning tasks. *arXiv preprint arXiv:2410.04055*, 2024.
- [111] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.
- [112] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback, 2024.
- [113] 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.
- [114] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.
- [115] 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. *arXiv preprint arXiv:2503.06749*, 2025.
- [116] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025.
- [117] Yuan-Hong Liao, Sven Elflein, Liu He, Laura Leal-Taixé, Yejin Choi, Sanja Fidler, and David Acuna. Longperceptualthoughts: Distilling system-2 reasoning for system-1 perception. *arXiv preprint arXiv:2504.15362*, 2025.

Appendix

The appendix includes extended details on data curation (Appendix A), implementation (Appendix B), cognitive behavior evaluation (Appendix C), and additional case studies (Appendix D).

A. Cold-Start Data Curation Details

As mentioned in Section 3.3, a critical component of our initial policy development is the curation of the cold-start SFT dataset. This stage serves as the foundation for subsequent learning, particularly in shaping the model’s ability to exhibit structured reasoning and cognitive behavior. To this end, we adopt a multi-stage curation pipeline consisting of data collection, filtering, cleaning, and strategic reweighting.

Data Acquisition. We begin by assembling a broad set of prompt-response pairs that span diverse reasoning domains. These include math, science, and general logical reasoning tasks such as puzzles, deductive tasks, and constraint satisfaction problems. Our sources include a mix of public datasets as illustrated in Section 3.3.

Automated Filtering. To improve the signal-to-noise ratio, we filter the collected data using a lightweight pretrained LLM as a proxy for quality estimation. Each instance is passed through this model, and those with abnormally high training loss are flagged as noisy, ambiguous, or misaligned. We further apply rule-based and model-assisted pattern detectors to identify and eliminate undesirable data characteristics.

Difficulty Stratification. We explicitly incorporate samples from AMC, AIME, Olympiads, and AoPS forums to ensure difficulty levels. We then stratify the collected samples based on their source and inherent problem complexity to balance coverage across easy, intermediate, and challenging reasoning scenarios.

Reweighting and Balance. To address imbalances across domains and formats, we apply a reweighting scheme based on coverage and reasoning relevance. Over-represented formats are down-weighted, while rare but cognitively rich categories are given higher sampling probabilities. This ensures a more uniform distribution of reasoning challenges and minimizes overfitting to dominant patterns.

B. More Implementation Details

Model and Optimization Setup Our model is based on the Qwen2.5-VL [23]. During RL, both the policy and critic networks are initialized from the cold-start model. The value head is initialized from a uniform distribution $\mathcal{U}(-\sqrt{5}, \sqrt{5})$ without bias. The policy and critic networks do not share weights during training. We use the AdamW optimizer with $\beta = [0.9, 0.95]$ and no weight decay. Learning rates are set to 1×10^{-6} for the policy and 5×10^{-6} for the critic. We use constant learning rates with a linear warm-up of 50 steps, and employ sample packing for improved throughput. No KL regularization or entropy bonus is used, demonstrating that vanilla PPO remains stable under our setup.

PPO Training Dynamics Each PPO update is based on 512 unique prompts, each generating 16 sampled responses using temperature and top-p sampling (both set to 1.0). To ensure training stability, we enforce strict on-policy updates for the policy: each prompt generation corresponds to a single optimization step. In contrast, the critic performs 4 optimization steps per PPO update. We apply batch-level advantage normalization to stabilize training further.

C. Details for Cognitive Behavior Evaluation

In this section, we detail definitions of metrics in cognitive behavior analysis (Section 5.2) and show the prompt for evaluation.

C.1. Behavior Transfer Rate

To quantify how well language-acquired cognitive behaviors generalize to the visual modality, we define the Behavior Transfer Rate (BTR) for each behavior type introduced in Fig. 5. BTR is calculated as the ratio between the emergence rate of visual behaviors and that of their linguistic counterparts. Formally, we compute the Cognitive Behavior Emergence Rate in the visual modality (CBR_v) and in the language modality (CBR_l), both evaluated on the multimodal benchmark *MathVision (mini)*. The **BTR** is then defined as:

$$BTR = \frac{CBR_v}{CBR_l}$$

This metric reflects the cross-modal transfer efficiency of cognitive behaviors, with higher values indicating stronger behavioral generalization from language to vision.

C.2. Evaluation Prompt

We design prompts for the LLM-based evaluation. Fig. 7 showcases the prompt template for the cognitive behavior *Backtracking* as an example.

Prompt example:

Problem Statement: "<QUESTION>"

Model's Reasoning: "<REASONING CONTENT>"

Evaluate the textual reasoning for Backtracking/Revision. This is when the model, in its text, realizes a chosen path, calculation, or assumption is incorrect or unproductive, and explicitly states it's trying a different approach, redoing calculations with different numbers/operations, or significantly altering its line of textual reasoning.

Example: "That calculation led to a dead end. Let's try subtracting Y from X instead." or "My previous assumption about the sequence was wrong. Starting over with..."

Count distinct instances of such explicit textual backtracking or major revision. Provide the count in <count>X</count> tags. If none, use <count>0</count>.

Figure 7: Prompt design for evaluating *Backtracking*.

D. More Case Studies

We present examples of OVR’s reasoning process during complex multimodal tasks. In Fig. 10, the model begins by perceiving the image and applying *visual divide-and-conquer* to interpret each emoji individually. It then engages in *subgoal setting*, breaking down the task to explore multiple candidate answers based on each emoji. Throughout the process, the model exhibits *backtracking*, revising earlier hypotheses by *cross-verifying* them with elements in the image. The reasoning process concludes with a final, well-justified prediction.



Question: Based on the paper, which one takes longer to train, RL phase or reward model?

Ground Truth: RL phase takes longer.

OVR: `</think>` Okay, so I need to figure out whether the RL phase or the reward model takes longer to train based on the information provided in the paper. Let me go through the relevant parts step by step.

First, the paper mentions the Seed1.5-VL model. It talks about a post-training framework using hybrid reinforcement learning with RLHF (Reinforcement Learning with Human Feedback) and RLVF (Reward Learning with Verifier Feedback). The RL phase is mentioned as requiring 24k GPU hours, while the reward model inference uses 60k GPU hours. Wait, but that's for the reward model during inference, right? Because it says "reward model inference use replicas each configured with tensor parallelism." So during training, maybe the reward model is trained for longer?

Wait, let me check again. The paper says: "The RL phase of Seed1.5-VL costs 60k GPU hours. The reward model is trained using the same framework as Seed1.5-VL pretraining phase, requiring 24k GPU hours." Hmm, that seems conflicting. Wait, no: "pretraining phase" is separate. Let me parse this sentence again.

Wait, the sentence says: "The RL phase of Seed1.5-VL costs 60k GPU hours. The reward model is trained using the same framework as Seed1.5-VL pretraining phase, requiring 24k GPU hours." So the RL phase takes 60k hours, and the reward model training (which is part of the post-training framework) uses 24k hours? That would mean the RL phase is longer than the reward model training. But maybe I'm misinterpreting.

Alternatively, perhaps there's a typo or confusion in the description. **Let me look again at the exact wording:**

"We use the same training framework as Seed1.5-VL pretraining phase for optimization techniques as in the pretraining phase for efficient actor and critic updates, and vLLM [81] for efficient progressive generation of rollouts. Specifically, the RL and RLVF training employs 3-D parallelism [61, 108, 126, 141]; rollout generation and reward/reference model inference use replicas each configured with tensor parallelism [7]. The RL phase of Seed1.5-VL costs 60k GPU hours. The reward model is trained using the same framework as Seed1.5-VL pretraining phase, requiring 24k GPU hours."

Wait, so here it says that during RL training (which includes both RL and RLVF?), it takes 60k hours. Then separately, the reward model is trained using the same framework as pretraining (which might be different) but takes 24k hours. But if pretraining took longer than post-training phases? Or maybe there's a mix-up here...

But wait, **let me check again:**

Original text from user input:

"We use ... optimization techniques as in the pretraining phase for efficient actor and critic updates ... The RL phase of Seed1.5-VL costs 60k GPU hours ... The reward model is trained using ... same framework as Seed1.5-VL pretraining phase ... requiring 24k GPU hours."

So during post-training phases:

- RL Phase: 60k
- Reward Model Training: 24k

Thus, RL Phase is longer than Reward Model Training.

Therefore, answer would be that the RL Phase takes longer to train than the Reward Model. `</think>`

Based on the provided information from the paper:

- **RL Phase:** Requires ****60k GPU hours****.
- **Reward Model Training:** Requires **24k GPU hours** during post-training phases.

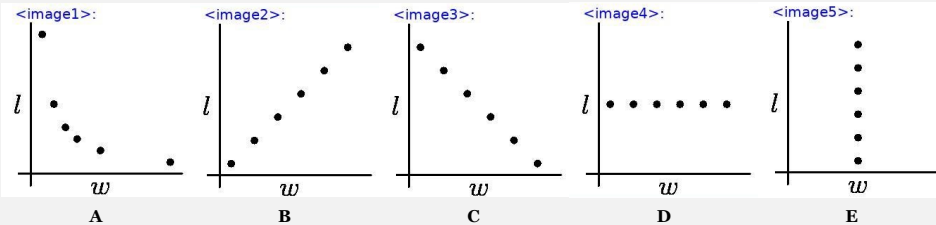
Conclusion: The **RL Phase** takes **longer** to train compared to **Reward Model Training**(60k vs. 24k GPU hours).

Answer: RL Phase

Figure 8: Case Study on OCR-based Document Understanding. Red highlights some of the cognitive behaviors.

Question: Jorge's teacher asks him to plot all the ordered pairs (w, l) of positive integers for which w is the width and l is the length of a rectangle with area 12. What should his graph look like?

Ground-Truth: A



OCR: Okay, so Jorge needs to plot all the ordered pairs (w, l) where w is the width and l is the length of a rectangle with area 12. Both w and l are positive integers. Let me think about how to approach this.

First, the area of a rectangle is calculated by multiplying its width and length. So, we have the equation $w * l = 12$. Both w and l have to be positive integers because they are dimensions of a rectangle, and you can't have a fraction of a unit in this context, right? The problem says positive integers, so we're dealing with whole numbers.

Alright, so I need to find all pairs of positive integers (w, l) such that when multiplied together, they equal 12. Let me list out the factors of 12. The factors of 12 are 1, 2, 3, 4, 6, and 12. But since both w and l have to be positive integers, each factor pair would be $(1,12)$, $(2,6)$, $(3,4)$, and then their reverses. **Wait, but since w is width and l is length, does the order matter?** In the ordered pair (w, l) , does it matter if we switch w and l ? For example, is $(2,6)$ different from $(6,2)$? Because if the rectangle has width 2 and length 6, that's the same as a rectangle with width 6 and length 2, just rotated. But the ordered pair (w, l) would be different. However, in terms of the graph, since we're plotting (w, l) , the points would be distinct even if they are just swaps. But wait, the area is the same, but the coordinates are different. Hmm...

Therefore, we need to consider all possible ordered pairs (w, l) where w and l are positive integers, and $w * l = 12$. Let's list all such pairs.

Starting with $w = 1$: $1 * 12 = 12$, so $(1,12)$

$w = 2$: $2 * 6 = 12$, so $(2,6)$

$w = 3$: $3 * 4 = 12$, so $(3,4)$

$w = 4$: $4 * 3 = 12$, so $(4,3)$

$w = 6$: $6 * 2 = 12$, so $(6,2)$

$w = 12$: $12 * 1 = 12$, so $(12,1)$

So the ordered pairs are $(1,12)$, $(2,6)$, $(3,4)$, $(4,3)$, $(6,2)$, $(12,1)$. Each of these is a distinct ordered pair.

So let's visualize this. The x-axis is w (width) from 1 to 12, and the y-axis is l (length) from 1 to 12. Each of these points should be plotted.

Looking at the options given:

(A) image1: shows points that seem to go from $(1,12)$ down to $(12,1)$, but maybe in a different arrangement.

Wait, but **let me check the images.**

(image1): a scatter plot with points arranged from left to right along the w -axis increasing, and l decreasing. So like $(1,12)$, $(2,6)$, $(3,4)$, $(4,3)$, $(6,2)$, $(12,1)$. But maybe not all points are there.

(image2): points increasing in both directions? Not sure.

(image3): Points decreasing in l as w increases, but maybe with gaps?

(image4): Points all along the w -axis, so l is constant? That can't be right because l would have to be 12 for $w=1$, 6 for $w=2$, etc.

(image5): Points all along the l -axis, so w is constant? Similarly, that's not possible.

Wait, but the correct plot should have points at $(1,12)$, $(2,6)$, $(3,4)$, $(4,3)$, $(6,2)$, $(12,1)$. So when plotted on a graph with w on the x-axis and l on the y-axis, these points would form a diagonal line from $(1,12)$ to $(12,1)$, but only at those specific integer points...

Let's check the options again.

Assuming image1 has all six points in decreasing order, image2 might have some other arrangement, image3 maybe have some other arrangement, image4 all on w -axis, image5 all on l -axis.

But since the problem says "positive integers", and there are exactly six pairs, the correct graph should have six points. So image4 and image5 can be eliminated because they have multiple points on a single axis, which is not possible since for each w , l is determined (except when $w=12$, $l=1$ and vice versa). So image4 is points along w -axis ($l=1$?), but that's not correct. Similarly, image5 would have points along l -axis ($w=1$?), which is also not correct.

So between image1, image2, image3.

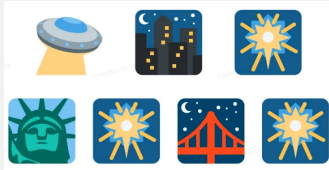
If image1 has points $(1,12)$, $(2,6)$, $(3,4)$, $(4,3)$, $(6,2)$, $(12,1)$, then image1 is correct. If image2 has points in a different order, maybe increasing, but since w is on the x-axis and l on y-axis, increasing w would correspond to decreasing l , so the points should form a decreasing trend...

Therefore, the correct answer is image1, which is option A. </think>

The problem requires plotting all ordered pairs (w, l) where w and l are positive integers such that the area ($w * l = 12$). The factor pairs of 12 are $(1, 12)$, $(2, 6)$, $(3, 4)$, $(4, 3)$, $(6, 2)$, and $(12, 1)$. These pairs form six distinct points on a graph with w on the x-axis and l on the y-axis. The correct graph should show all six points arranged in a decreasing trend from $(1, 12)$ to $(12, 1)$. Among the options, **image1** (Option A) correctly displays all six points in this pattern.

Answer: A

Figure 9: Case Study on Math Reasoning. Red highlights some of the cognitive behaviors.



Question: Which movie do these emojis remind you of?

Ground Truth: Independence Day

OVR: <think> Okay, let's try to figure out which movie these emojis remind me of. First, I'll list out all the emojis again to make sure I have them right. **The first row** has a UFO, a city at night with buildings, and a sun. **The second row** includes the Statue of Liberty, a sun again, a bridge (probably the Golden Gate), and another sun-like emoji.

Hmm, let's break them down one by one. The UFO is a classic alien spacecraft, so that might hint at a movie involving extraterrestrial life. The city at night could be a setting for a crime story or maybe a thriller. The sun emojis might relate to a story that involves weather changes, maybe a sun that's either a hero or a villain. The Statue of Liberty is a specific symbol, often representing freedom or the United States. The bridge, especially if it's the Golden Gate, is a famous landmark in San Francisco. Putting this together, maybe a movie that has a UFO, involves a city at night, and elements related to the sun or weather changes. Also, the Statue of Liberty and the Golden Gate Bridge are both in the US, so perhaps the movie is set in America. The sun emojis might indicate a weather theme, like a storm or a heat wave. Wait, but the sun could also be a recurring element in a story, like in "Twister" where weather is a big part, but that's a disaster movie. Alternatively, "The Wizard of Oz" has a big tornado and a sun, but that's more of a classic fairy tale.

Wait, the Statue of Liberty is a big clue. There's a movie called "Night at the Museum" where the statue comes to life, but I don't think that's it. Another possibility is "Independence Day" which features a massive alien spaceship attacking Washington D.C. The city at night could be the capital, and the UFO matches. The sun emojis might be part of the movie's imagery, like the aliens using sunlight or something. Also, "Independence Day" has a lot of action with the aliens and the U.S. government. The Statue of Liberty is in New York, so maybe the attack happens there? Wait, in the movie, the attack starts in the Midwest and then moves towards Washington D.C., but maybe other parts of the U.S. are involved.

The Golden Gate Bridge is a significant location in the movie "San Francisco 1971" which involves a killer, but that's a different setting. Alternatively, "The Blob" is a horror movie where a giant blob invades a town, but again, not sure. Wait, the sun emojis could be related to the movie "The Day After" which is about a nuclear attack, **but the sun might not fit.**

Another angle: the combination of a UFO, a city at night, and the sun. Maybe "The Fifth Element" has a flying saucer and a futuristic city, but the sun isn't a main element. The Statue of Liberty is prominent in "The Happening" where a virus takes over, but that's more about a virus.

Wait, the sun emojis might be part of the movie's title or a key element. For example, "Sunset" or "Sunrise" but those are too generic. Alternatively, "The Solar Storm" but that's not a movie I know.

Let me think of movies that have all these elements. "Invasion of the Body Snatchers" has alien-like plants taking over people, but the emojis don't quite match. "Alien" has a spaceship but not a city at night. "E.T. the Extra-Terrestrial" has a flying saucer and a boy, but not a city at night.

Wait, maybe "Transformers" has a sunken city in the movies, like in "Transformers: Age of Extinction" there's a sinking city, but the sun emojis might be part of the visual effects. The Statue of Liberty isn't a main part of that movie.

Alternatively, "The Matrix" has digital cities and machines, but not UFOs.

Wait, "The Blob" has a giant blob that engulfs a town, but again, not sure.

Wait, the sun emojis could be related to the movie "Twister" which involves severe thunderstorms and a giant tornado. The sun might be part of the storm system, but the UFO and city at night don't fit.

Wait, "Independence Day" is a big one. The movie is about an alien spaceship attacking Earth, specifically the U.S. The attack starts in the Midwest and moves towards the East Coast. The city at night could be a reference to the attack on Washington D.C. The UFO is the main alien ship. The sun emojis might be part of the movie's imagery, maybe the aliens' heat rays or something. The Statue of Liberty is a U.S. symbol, so if the attack is targeting the U.S., that makes sense. The Golden Gate Bridge could be part of the attack's path.

Let me check the plot of "Independence Day" again. The movie features a huge alien mothership that lands in the Midwest, and the U.S. military tries to destroy it. The attack is depicted with the aliens using heat-ray weapons, which might look like sunbursts. The city at night could be a reference to the attack on Washington D.C., which is a major city. The Statue of Liberty is a symbol of the U.S., so it's part of the national pride theme. The golden gate bridge is a key location in San Francisco, which might be on the attack's path.

Yes, that seems to fit. All the emojis relate to the movie "Independence Day." The UFO is the alien mothership, the city at night could be Washington D.C., the sun emojis might represent the aliens' heat rays, the Statue of Liberty is a U.S. symbol, and the Golden Gate Bridge is part of the attack's movement towards the East Coast.

Another possibility is "The Day After" which is about nuclear war, but the sun emojis don't fit. "2012" is about a comet, but again, not matching.

So, putting it all together, the emojis are likely from "Independence Day." </think>

The emojis collectively remind me of the movie "**Independence Day**" (1996). Here's how each emoji aligns with the film:

- 1.UFO:** Represents the massive alien mothership that attacks Earth.
- 2.City at Night:** Refers to the attack on Washington, D.C., a major U.S. city.
- 3.Sun Emojis:** Symbolize the aliens' heat-ray weapons, which resemble bright, radiant bursts.
- 4.Statue of Liberty:** A U.S. icon, emphasizing the movie's focus on defending American soil.
- 5.Golden Gate Bridge:** A key geographic reference for the attack's progression along the East Coast.

The combination of alien threats, U.S. landmarks, and dramatic imagery matches the movie's plot about an extraterrestrial invasion.

Answer: Independence Day

Figure 10: **Case Study on Emoji Quiz.** Red highlights some of the cognitive behaviors. The model first performs *visual decomposition* to interpret each emoji individually. It then engages in *subgoal setting* to explore multiple candidate answers, accompanied by *backtracking* to revise earlier hypotheses. The reasoning concludes with a final, justified prediction.

Question: The number of points equidistant from a circle and two parallel tangents to the circle is:
 (A) 0. (B) 2. (C) 3. (D) 4. (E) infinitely many.
 (20th American High School Mathematics Examination, 1969)

DeepSeek-R1: *<think>* \n\nThe problem is from the 20th American High School Mathematics Examination in 1969. It states: \n\n\"The number of points equidistant from a circle and two parallel tangents to the circle is\" with options ranging from 0 to infinitely many.\n\nFirst, I need to understand what is being asked. We have a circle and two parallel tangents to this circle. I need to find points that are equidistant from the circle and from these two tangents. The phrasing is a bit ambiguous. Let me read it again.\n\n\"It says: equidistant from a circle and two parallel tangents.\" I think it means equidistant from the circle itself and from the pair of tangents. But distance to a circle usually means the shortest distance to the circumference, and distance to a line is the perpendicular distance.\n\nBut it says \"from a circle and two parallel tangents.\" It might mean equidistant from the circle and from each of the tangents, but that doesn't make sense because the tangents are lines, and the circle is a curve.\n\nPerhaps it means that the point is equidistant from the circle and from the two tangents considered together, but that seems messy.\n\nAnother interpretation: maybe it means that the distance to the circle is equal to the distance to each tangent, but that would probably be overconstrained.\n\nI recall that in some contexts, \"equidistant from a curve and a line\" means that the shortest distance to the curve equals the shortest distance to the line.\n\nBut here it's a circle and two lines.\n\nPerhaps the problem intends for the distance to the circle to be equal to the distance to the pair of tangents, but since the tangents are parallel, the distance to the pair might be the distance to the closer one or something.\n\nThat doesn't sound right.\n\n**Let me visualize this.** Let me sketch a circle. Assume it's a circle with center O. Draw two parallel tangents to this circle. Since they are parallel, they must be on opposite sides of the circle. For example, if the circle is centered at the origin, and the tangents are horizontal, say $y = r$ and $y = -r$, but if the circle has radius r , and center at $(0,0)$, then the tangents would be $y = r$ and $y = -r$? No.\n\nIf center is at $(0,0)$, radius r , then the tangent lines are $x = \pm r$ for vertical tangents, but for horizontal, if I want parallel tangents, say both horizontal. So, if I have a circle $x^2 + y^2 = r^2$, the horizontal tangents are $y = r$ and $y = -r$. Is that correct? When $y = r$, $x^2 + r^2 = r^2$ implies $x = 0$, so yes, the point $(0,r)$ is on the circle, and the tangent is horizontal. Similarly for $y = -r$. So yes, $y = r$ and $y = -r$ are the two parallel tangents.\n\nBut in this case, they are symmetric.\n\nNow, I need points P such that the distance from P to the circle is equal to the distance from P to each of the tangents? But that might not be what it means.\n\nThe problem says \"equidistant from a circle and two parallel tangents\"

...

Figure 11: **Case Study on DeepSeek-R1.** Red highlights the *mental imagery* mentioned in Section 5.2.