



Genesis: Visual Augmentation Primitives for Multimodal Chain-of-Thought Reasoning

Fan Yang^{1,3,*}, Lifeng Chen^{4,*}, Yipu Wang^{2,3,*}, Han Wang^{3,*},
Feilong Chen^{7,◇,†}, Yidong Ma⁷, Zhiyang Chen⁵, Chaoyang Zhao^{1,†},
Hanqi Jiang⁴, Yibo Chen⁷, Xu Tang⁷, Yao Hu⁷,
Ming Tang¹, Xiaojie Jin⁴, Xiaolong Zheng^{2,3}, Jinqiao Wang^{1,3,6}

¹Foundation Model Research Center, Institute of Automation,

² State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation,

³University of Chinese Academy of Sciences, ⁴Beijing Jiaotong University

⁵Independent Researcher, ⁶Wuhan AI Research, ⁷Xiaohongshu Inc,

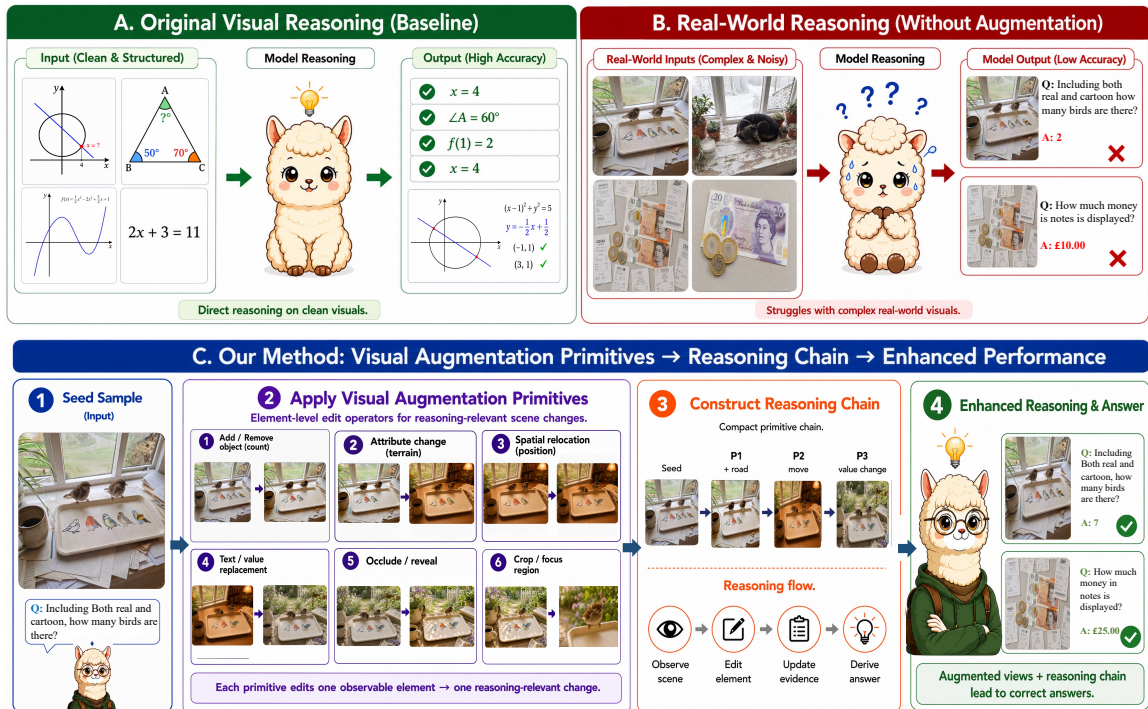


Figure 1: Motivation for Visual Augmentation Primitives. Existing structured STEM benchmarks are often visually clean, while real-world photographs require reasoning steps to be grounded in observable visual evidence. Our Visual Augmentation Primitives generate diverse real-world variants with coherent multimodal reasoning chains.

Abstract

Recent multimodal large language models achieve promising results on visual reasoning tasks, approaching human-level performance on existing STEM benchmarks. However, these capabilities degrade under complex real-world scenes. We identify that chain-of-thought reasoning remains largely confined to the linguistic space, rarely anchoring each step in visual evidence. The main cause is the absence of such data during training. In this

work, we introduce a training data augmentation pipeline, **Visual Augmentation Primitives (VAP)**, which leverages state-of-the-art vision generative models to synthesize diverse photorealistic training samples and corresponding reasoning text. VAP introduces some basic operators and applies them to an seed image and get a variant. It also constructs coherent multimodal reasoning chains throughout the generation process. Such training data force models to perceive visual content and perform logical reasoning according to them at every step. We further introduce multimodal chain-of-thought rubric-based reinforcement learning

*Equal contribution. ◇Project leader. †Corresponding author.

to enhance this. It rewards not only text reasoning but also visual reasoning by scoring intermediate reasoning steps for logical validity, completeness, and consistency. Moreover, we construct the UNREAL benchmark to systematically evaluate visual reasoning on hyperrealistic scenes across controllable difficulty levels. Experiments show that a 4B model built on this framework surpasses 8B baselines, with performance improving consistently as the model scales to 32B, validating the scaling law for visual reasoning. All models, data, and code will be publicly released.

1 Introduction

Recent multimodal large language models (MLLMs) achieve strong performance on visual reasoning tasks such as geometry, function analysis, and quantitative problem solving. Representative models such as GPT-5.5 (OpenAI, 2025b), Gemini 3.5 (Google DeepMind, 2025), Claude 4.7 (Anthropic, 2025), and Qwen 3.7 (Qwen Team, 2025) establish impressive results on complex problems in mathematics, physics, and chemistry that combine textual descriptions and visual evidence (Yang et al., 2026c; Lu et al., 2024; Yue et al., 2024; Lu et al., 2022). As illustrated in Figure 1(A), these models already achieve strong generalization across a range of visual reasoning tasks.

However, as shown in Figure 1(B), these capabilities degrade on complex real-world scenes, where those models frequently produce incoherent reasoning chains and misread the spatial relations between objects (Fu et al., 2024; Liu et al., 2023a). We attribute this degradation to two intertwined causes. First, chain-of-thought reasoning remains largely confined to the linguistic space, with intermediate steps rarely anchored in multimodal visual evidence, leading to hallucinated reasoning that compounds into incorrect final answers (Tong et al., 2024). Second, the scarcity of multimodal chain-of-thought training data prevents MLLMs from leveraging scaling laws to improve multi-hop visual reasoning capabilities on real-world scenes. This raises a fundamental research question: *Can we extend the emergent reasoning capabilities of MLLMs to complex real-world scenes?*

In this paper, we introduce **Visual Augmentation Primitives** (VAP), a set of controllable, generation-based scene editing primitives. VAP defines global scene analysis, key element identification, and edit instructions as atomic opera-

tions, each explicitly aligned with a corresponding step in the multimodal chain-of-thought. As illustrated in Figure 1(C), this design ensures that every step from holistic scene understanding to localized edit regions remains consistent with its corresponding reasoning step. Applying these primitives through instruction editing models across diverse visual contexts produces large-scale training data paired with multimodal chain-of-thought supervision, laying the foundation for scaling visual reasoning. Moreover, we further observe that existing RLVR optimizes only final-answer correctness and lacks supervision over intermediate reasoning, which easily leads to inconsistencies between the chain-of-thought and the final answer. We therefore propose **Genesis**, a multimodal reasoning language model that combines VAP with multimodal chain-of-thought rubric-based reinforcement learning. By extending rubric-based reward design (Lightman et al., 2024; Wang et al., 2024a) from the text-only domain to visual reasoning, the rubric rewards can score each intermediate reasoning step along three dimensions, namely logical consistency, reasoning completeness, and multi-step process coherence.

Meanwhile, to test whether models truly rely on visual evidence for reasoning, we construct the **UNREAL** benchmark. UNREAL deliberately designs scenes that contradict common sense, so that models cannot arrive at correct answers by relying solely on language priors but must anchor every reasoning step in the visual content actually presented in the image. The benchmark generates image and question pairs with controllable difficulty levels, enabling systematic evaluation of both hallucination robustness (Li et al., 2023b; Guan et al., 2024) and visual reasoning capability. Experiments validate the effectiveness of our approach. A 4B-parameter Genesis model surpasses 8B-parameter baselines on multiple tasks, and scaling to 32B achieves significant improvements across visual reasoning and VQA benchmarks.

- We introduce **Visual Augmentation Primitives**, which defines the visual elements of real-world images as editing primitives. Instruction-guided generative models apply these primitives to produce photorealistic image variants paired with multimodal reasoning chains, requiring models to unify visual perception with logical reasoning at every step.
- We introduce multimodal chain-of-thought rubric-based reinforcement learning, which

extends rubric-based reward design from the text-only domain to visual reasoning, scoring intermediate reasoning steps for logical validity, completeness, and consistency.

- We construct the **UNREAL** benchmark to systematically evaluate visual reasoning on scenes that contradict text priors across diverse visual contexts. A 4B Genesis model surpasses 8B baselines, and scaling to 32B achieves significant improvements across visual reasoning and VQA benchmarks. All models, data, and code will be publicly released.

2 Related Work

2.1 Visual Reasoning Benchmarks and Their Limitations

A wide range of benchmarks evaluate visual reasoning in multimodal models. Broad assessments such as SEED-Bench (Li et al., 2023a) and MM-Bench (Liu et al., 2024) cover general capabilities, while specialized suites including MathVista (Lu et al., 2024), MMMU (Yue et al., 2024), and ScienceQA (Lu et al., 2022) target STEM problem solving. Most rely on multiple-choice formats or templated questions and struggle to probe genuine visual understanding beyond their authored templates. Diagnostic suites such as BLINK (Fu et al., 2024) and HallusionBench (Guan et al., 2024) further reveal that frontier MLLMs still falter on basic perception and confidently hallucinate on real-world images (Tong et al., 2024). Our UNREAL benchmark complements this line by stress-testing reasoning on hyperrealistic generated scenes with controllable difficulty, enabling systematic evaluation across scalable difficulty levels.

2.2 Chain-of-Thought Supervision and Reasoning Consistency

Chain-of-thought (CoT) prompting (Wei et al., 2022) has become central to multi-step reasoning, with multimodal variants such as Multimodal-CoT (Zhang et al., 2023) and Visual CoT (Shao et al., 2024a; Yang et al., 2026a) extending it to vision-language inputs. A persistent limitation is that reasoning chains in MLLMs often drift toward language priors rather than observable visual evidence, and this drift compounds as chains grow longer (Tong et al., 2024). Process-level supervision approaches such as Let’s Verify

Step by Step (Lightman et al., 2024) and Math-Shepherd (Wang et al., 2024a) address the same problem in the text domain by rewarding intermediate steps, but such process reward models have rarely been applied to visual reasoning at scale. Genesis brings this direction into the visual domain by pairing every synthetic scene with reasoning chains grounded in visual evidence and by training under a rubric-based reinforcement objective that scores intermediate visual and logical steps.

2.3 Scalable Data Synthesis for Multimodal Training

Synthetic data generation has become a standard response to annotation bottlenecks in vision-language research. LLM-driven approaches such as LLaVA-Instruct (Liu et al., 2023b), ShareGPT4V (Chen et al., 2024b), and the All-Seeing project (Wang et al., 2024b) bootstrap instruction-following data from existing image-caption pairs, while ALLaVA (Chen et al., 2024a) further scales this strategy with GPT-4V supervision. These pipelines remain anchored to the source images they start from, limiting the diversity and visual complexity of the resulting data. Recent advances in instruction-guided image generation (Podell et al., 2024; Black Forest Labs, 2024; OpenAI, 2025a; Yang et al., 2026b) make it feasible to synthesize entirely novel scenes with controllable attributes, opening a fundamentally different scaling regime. Genesis builds on this opportunity through two complementary mechanisms. Logical answer expansion and scene variance generation turn a single seed sample into a family of training examples sharing one coherent reasoning chain. The resulting data is further coupled with rubric-based reinforcement learning that scores intermediate reasoning steps for process-level supervision.

3 Method

In this section, we present Visual Augmentation Primitives and its two-stage data pipeline (Section 3.1), multimodal chain-of-thought rubric-based reinforcement learning (Section 3.2), and the UNREAL benchmark (Section 3.3).

3.1 Visual Augmentation Primitives

Current training data does not require each reasoning step to depend on visual observations, leaving complex visual reasoning capability inadequately trained. We introduce **Visual Augmentation Primitives** (VAP), a set of controllable scene editing



Figure 2: Two-stage data construction pipeline. **Stage 1 (Surgical Answer-Changing Editing)**: the Analyst extracts the global reasoning focus and partitions visual attributes into frozen and free dimensions, the Director defines an edit primitive, the Generator executes it, and the Verifier validates the output. **Stage 2 (Axis Diversification & CoT Annotation)**: the Planner selects free dimensions to diversify, the Generator produces visual variants, a four-step CoT is constructed for each, and the final dataset $\mathcal{D}_{\text{syn}}$ is assembled.

primitives. Instruction-guided generative models apply these primitives to modify real-world images, producing diverse photorealistic variants where each correct answer is tied to observable changes in the scene.

The VAP-based pipeline operates in two stages coordinated by five functional modules, with the **Analyst** and **Planner** instantiated by Gemini (Google DeepMind, 2025), the **Director** by Qwen (Qwen Team, 2025), the **Generator** by GPT-Image (OpenAI, 2025a), and a rule-based **Verifier** for quality control (Figure 2).

Surgical Answer-Changing Editing (Stage 1). For each input triplet (I, q, a) , the **Analyst** examines the image to extract the global reasoning focus, including the task type, counting scope, target entities, and reasoning requirement. It then partitions the visual attributes (lighting, viewpoint, object color, background, etc.; full list in Appendix A.1) into **frozen dimensions** ($\mathcal{F}$) that encode reasoning-critical information and must be preserved, and complementary **free dimensions** that can vary to introduce diversity. A predicate ϕ enforces this

partition,

$$\phi(I') = \prod_{j \in \mathcal{F}} \mathbb{I}(x_j(I') = x_j(I)), \quad (1)$$

where $x_j(\cdot)$ extracts the j -th attribute. ϕ serves both as a validation standard for generated images and as the basis for RL rubrics in Section 3.2. The **Director** then defines a Visual Augmentation Primitive p consisting of an operation from $\{\text{Add, Remove, Modify, Relate}\}$, the target element, the constraints on frozen attributes, and the corresponding answer change. The **Generator** executes the controlled edit to synthesize the edited image I' with high visual fidelity while respecting ϕ . The **Verifier** retains (I', q, a') only when the answer has changed ($a' \neq a$), the edited scene remains visually plausible, and the frozen attributes are preserved ($\phi(I') = 1$). This stage yields edited seed images with verified answers and a preliminary chain-of-thought overview capturing the question type, key reasoning regions, and the applied transformation.

Axis Diversification and CoT Annotation (Stage 2). Building on each verified edit from Stage 1,

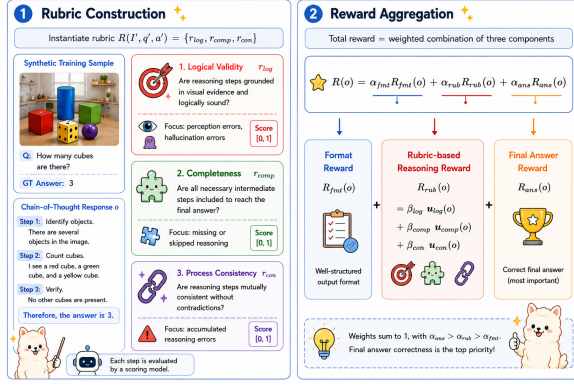


Figure 3: Overview of multimodal chain-of-thought rubric-based reinforcement learning. **(1) Rubric construction** scores each reasoning step along three dimensions (logical validity, completeness, process consistency). **(2) Reward aggregation** combines format, rubric-based reasoning, and final-answer rewards, with final-answer correctness as the top priority.

this stage produces visually diverse variants and constructs structured multimodal reasoning chains. The **Planner** selects which free dimensions to vary, such as lighting, weather, viewpoint, background, color, style, object appearance, and scene context, while keeping all frozen dimensions $\mathcal{F}$ unchanged. The **Generator** synthesizes multiple variants per edited sample that share the same answer but differ along the chosen free axes. For each variant, we construct a four-step multimodal chain-of-thought z' that decomposes the reasoning process into (1) observing the overall image structure, (2) identifying key visual elements, (3) performing task-specific reasoning such as counting or spatial analysis, and (4) concluding with a final answer. The **Verifier** validates the quality of both the generated images and the reasoning chains, ensuring that visual variations remain faithful to ϕ . The final dataset $\mathcal{D}_{\text{syn}}$ pairs each variant with its question, answer, and chain-of-thought.

3.2 Multimodal Chain-of-Thought Rubric-Based Reinforcement Learning

Standard reinforcement learning with verifiable rewards (RLVR) optimizes only final-answer correctness, providing no reward signal on the quality of intermediate reasoning steps. The multimodal reasoning chains produced by Visual Augmentation Primitives (observation, identification, task-specific operations, and conclusion) demand process-level supervision that goes beyond binary rewards over final answers. Extending rubric-based reward design (Gunjal et al., 2025; Lightman et al.,



Figure 4: UNREAL benchmark composition. (a) Sample distribution across four STEM families (inner ring, with sub-domain counts) and eight question types (outer ring). (b) Two-dimensional difficulty stratification by reasoning complexity $|\mathcal{F}|$ and visual distraction intensity, with marginal totals.

2024; Wang et al., 2024a) from the text-only domain to visual reasoning, we introduce multimodal chain-of-thought rubric-based reinforcement learning, which evaluates each intermediate reasoning step for logical validity, completeness, and consistency, as illustrated in Figure 3.

Rubric Construction. We score every response along three dimensions tailored to multimodal reasoning chains.

- **Logical validity.** Whether each step is a valid logical consequence of prior steps and visual evidence.
- **Completeness.** Whether each step in the response aligns with the corresponding step in the ground-truth reasoning chain z' , providing dense step-level rewards rather than a single trajectory-level score.
- **Process consistency.** Whether consecutive steps form a coherent chain that progresses naturally, free of contradictions or circular reasoning.

Together, these three dimensions form the rubric $\mathcal{R}(I', q', a') = \{r_{\log}, r_{\text{comp}}, r_{\text{con}}\}$, with each item scored on $[0, 1]$ by an LLM-based judge.

Reward Aggregation. The total reward combines a format reward R_{fmt} , the rubric-based reasoning reward $R_{\text{rub}}(o) = \beta_{\log} u_{\log}(o) + \beta_{\text{comp}} u_{\text{comp}}(o) + \beta_{\text{con}} u_{\text{con}}(o)$, and a final-answer

reward R_{ans} ,

$$R(o) = \alpha_{\text{fmt}} R_{\text{fmt}}(o) + \alpha_{\text{rub}} R_{\text{rub}}(o) + \alpha_{\text{ans}} R_{\text{ans}}(o), \quad (2)$$

with weights summing to one and ordered as $\alpha_{\text{ans}} > \alpha_{\text{rub}} > \alpha_{\text{fmt}}$ so that final-answer correctness remains dominant. Starting from the SFT model, we optimize the policy with GRPO (Shao et al., 2024b), using group relative advantage estimation and KL regularization toward the SFT baseline.

3.3 UNREAL Benchmark

We construct **UNREAL**, a STEM visual reasoning benchmark containing 633 samples across five STEM families (Figure 4), designed to systematically evaluate model reasoning on hyperrealistic scenes. Every question is presented in open-ended form, enabling evaluation of both hallucination robustness (Li et al., 2023b; Guan et al., 2024) and real-world logical reasoning. UNREAL is built around three design principles. **Rule competition** ensures the correct answer depends on the rule shown in the image rather than the model’s text-book prior. **Structure preservation** keeps the logical structure of each problem fixed across visual variants. **High difficulty** requires genuine visual understanding rather than language-only pattern matching.

Data Construction and Quality Control. UNREAL is constructed through a three-phase process. First, seed questions are collected from open educational resources across five STEM families. Mathematics (44.2%) covers eight sub-domains including combinatorial geometry, metric geometry, and graph theory. Physics & Engineering (16.1%) spans mechanics, electromagnetism & optics, and fluids & thermodynamics. Reasoning & Logic (15.3%) includes logic & document knowledge and counting & numeracy. Perception (14.5%) covers spatial reasoning and visual perception. Natural Science (9.8%) addresses biology & chemistry. Second, the VAP pipeline generates hyperrealistic scene variants for each seed with difficulty controlled along two dimensions: **reasoning complexity**, determined by the number of frozen dimensions $|\mathcal{F}|$, and **visual distraction intensity**, determined by the magnitude of variation over free dimensions. Third, every generated sample undergoes human verification for answer correctness, visual plausibility, and reasoning consistency. Samples that fail any criterion are discarded, and eval-

uation questions are drawn from an independent pool fully isolated from model training.

Evaluation Protocol. The protocol requires models to generate final answers directly without multiple-choice formats or auxiliary cues. Given the evaluation set $\mathcal{D}_{\text{eval}} = \{(I'_n, q'_n, a'_n)\}_{n=1}^M$, the score of a model π is

$$\text{Score}(\pi) = \frac{1}{M} \sum_{n=1}^M \mathbb{I}(\text{match}(\pi(I'_n, q'_n), a'_n)), \quad (3)$$

where $\text{match}(\cdot, \cdot)$ denotes semantic equivalence that ignores surface differences such as numerical formatting and letter case. We report both overall and difficulty-stratified performance to analyze how model behavior scales across varying task complexity.

4 Experiments

4.1 Experimental Setup

Models and Training. We validate our approach on three model configurations of Qwen3-VL-Instruct at 4B, 8B, and 32B, comparing each against its corresponding base model for direct, scale-matched evaluation. All models are trained on the synthetic dataset $\mathcal{D}_{\text{syn}}$ produced by Visual Augmentation Primitives with the multimodal chain-of-thought rubric-based reinforcement learning described in Section 3.2, using GRPO (Shao et al., 2024b) with group relative advantage estimation and KL regularization toward the base policy. The 4B and 8B variants are trained on a single 8×GPU node, while the 32B variant uses 48 GPUs across six nodes with rollout tensor parallelism. Full hyperparameters are provided in Appendix B.4.

4.2 STEM Benchmarks

We evaluate on eight challenging STEM benchmarks spanning mathematics, science, and visual logic, namely ZeroBench (Roberts et al., 2025), ZeroBench-Sub (Roberts et al., 2025), MathVista (MINI split) (Lu et al., 2024), MMMU (DEV-VAL split) (Yue et al., 2024), MMMU-Pro, EMMA (Hao et al., 2025), LogicVista (Xiao et al., 2024), and BabyVision (Chen et al., 2026). Table 1 reports the main STEM results. Genesis delivers consistent improvements over the Qwen3-VL baselines at every model scale. Genesis-4B surpasses not only Qwen3-VL-8B but also InternVL3-8B on LogicVista (60.0 vs. 44.1), confirming that process-

Table 1: **Main results on STEM visual reasoning benchmarks.** **Bold** and underlined denote the best and second-best results among open-source models, respectively. Closed-source model results are shown for reference only and are excluded from ranking. Subscripts show the improvement over the corresponding Qwen3-VL baseline. “–” indicates a result that is still being verified.

Model	ZeroBench	ZB-Sub	MathVista	MMMU	MMMU-Pro	EMMA	LogicVista	BabyVision
<i>Closed-source models</i>								
GPT-5	2	-	81.3	84.2	78.4	-	71.8	-
Gemini-2.5-Pro	3	-	82.7	81.7	68.8	-	72.0	-
Claude-Opus-4.1	3	-	75.5	78.4	64.8	-	67.3	-
<i>Open-source models</i>								
InternVL3-8B	-	-	71.6	62.7	-	-	44.1	-
InternVL3-38B	-	-	75.1	<u>70.1</u>	-	-	58.4	-
LLaVA-OV-72B	-	-	67.5	56.8	-	-	-	-
Qwen3-VL-4B	0	13.2	74.3	63.3	36.0	30.5	50.3	15.0
Qwen3-VL-8B	0	18.3	78.2	63.6	41.4	31.0	55.9	10.8
Qwen3-VL-32B	<u>1</u>	<u>26.7</u>	82.9	68.2	47.9	<u>38.1</u>	<u>63.5</u>	<u>16.0</u>
Genesis-4B	0	17.1 _{+3.9}	77.6 _{+3.3}	63.4 _{+0.1}	53.9 _{+17.9}	32.8 _{+2.3}	60.0 _{+9.7}	16.8_{+1.8}
Genesis-8B	<u>1</u> ₊₁	24.9 _{+6.6}	78.8 _{+0.6}	66.0 _{+2.4}	<u>56.7</u> _{+15.3}	35.8 _{+4.8}	59.7 _{+3.8}	15.2 _{+4.4}
Genesis-32B	3 ₊₂	29.9 _{+3.2}	<u>82.4</u> _{-0.5}	75.4 _{+7.2}	66.5 _{+18.6}	46.5 _{+8.4}	68.5 _{+5.0}	14.4 _{-1.6}

level supervision enables a smaller model to outperform larger baselines. At the 32B scale, Genesis-32B achieves the best results on six of eight benchmarks among open-source models, with the largest improvements over Qwen3-VL-32B on MMMU-Pro (+18.6), EMMA (+8.4), and MMMU (+7.2). The improvements are most pronounced on benchmarks that require multi-step deduction (MMMU-Pro, EMMA, LogicVista), validating the design of rubric-based rewards over intermediate multimodal reasoning.

4.3 VQA Benchmarks

We further evaluate on six visual question answering benchmarks covering hallucination robustness (HallusionBench (Guan et al., 2024)), aggregate visual skills (MMStar), and infographic understanding (InfoVQA).

As shown in Table 2, Genesis-32B achieves the best results among open-source models on MMStar (75.9) and InfoVQA (87.5). It surpasses all closed-source baselines on HallusionBench (74.6 vs. 65.7 for GPT-5) and InfoVQA (87.5 vs. 84.2 for Gemini-2.5-Pro), demonstrating that rubric-based supervision improves both hallucination robustness and fine-grained visual understanding.

4.4 UNREAL Benchmark

We evaluate on UNREAL, our STEM visual reasoning benchmark with controllable difficulty levels and strictly open-ended questions. Results are reported per STEM family and overall.

As shown in Table 3, UNREAL presents a significantly harder challenge than standard benchmarks.

Table 2: **Main results on VQA benchmarks.** **Bold** and underlined denote the best and second-best results among open-source models, respectively. Closed-source model results are shown for reference only and excluded from ranking. Subscripts show the improvement over the corresponding Qwen3-VL baseline.

Model	HallusionBench	MMStar	InfoVQA
<i>Closed-source models</i>			
GPT-5	65.7	76.4	79.0
Gemini-2.5-Pro	63.7	77.5	84.2
Claude-Opus-4.1	60.4	72.1	69.4
<i>Open-source models</i>			
InternVL3-8B	49.9	68.2	-
InternVL3-28B	57.1	71.5	-
LLaVA-OV-72B	-	66.1	74.9
Qwen3-VL-4B	71.2	68.5	79.7
Qwen3-VL-8B	72.2	72.0	82.0
Qwen3-VL-32B	75.0	<u>74.7</u>	<u>87.4</u>
Genesis-4B	72.5 _{+1.3}	70.4 _{+1.9}	82.9 _{+3.2}
Genesis-8B	73.0 _{+0.8}	72.0 _{+0.0}	84.7 _{+2.7}
Genesis-32B	<u>74.6</u> _{-0.4}	75.9 _{+1.2}	87.5 _{+0.1}

Among proprietary models, GPT-5.5 achieves the highest overall accuracy (59.6). Among open-source models, Qwen3.6-Plus leads at 42.1 overall, while Qwen3-VL-32B achieves 24.3. Performance is consistently lowest on the mathematics domain across all models, identifying multi-step symbolic-visual reasoning as the remaining frontier.

4.5 Ablation Studies

Rubric-Based RL vs. RLVR. Table 4 compares rubric-based RL against standard RLVR at the 32B scale, holding training data, optimizer, and compute budget constant. Rubric supervision yields consistent gains across all benchmarks, with the

Table 3: **Results on the UNREAL benchmark, stratified by domain.** Bold denotes the best result per column. Scores are accuracy (%).

Model	Overall	Math.	Nat. Sci.	Percep.	Phys. & Eng.	Reas. & Log.
<i>Closed-source models</i>						
Claude-Opus-4.7	32.9	28.6	46.8	28.3	32.4	41.2
Claude-Sonnet-4.6	30.7	23.4	45.2	32.6	35.3	36.1
Doubao-Seed-1.8	35.9	27.5	45.2	39.1	44.1	42.3
Gemini-3.1-Pro	37.1	23.2	51.6	43.5	53.9	44.3
GPT-5.5	59.6	62.9	51.6	54.3	56.9	62.9
Qwen3.6-Plus	42.1	40.7	37.7	32.6	44.1	55.7
<i>Open-source models</i>						
InternVL3.5-38B	18.3	6.8	43.5	19.6	28.4	23.7
InternVL3.5-8B	16.6	9.6	38.7	14.1	20.6	20.6
Qwen3-VL-32B	24.3	17.9	35.5	27.2	24.5	33.0
Qwen3-VL-8B	17.5	7.9	38.7	18.5	26.5	21.6
Qwen3-VL-4B	12.8	5.0	30.6	13.0	13.7	22.7
Qwen3.5-27B	22.1	9.3	41.9	37.0	19.6	35.1
Qwen3.5-9B	18.3	5.4	46.8	22.8	17.6	34.0
Qwen3.5-4B	16.6	6.4	30.6	22.8	22.5	24.7
Genesis-4B	14.2 ^{+1.4}	6.1 ^{+1.1}	32.3 ^{+1.7}	14.1 ^{+1.1}	15.7 ^{+2.0}	24.7 ^{+2.0}
Genesis-8B	19.1 ^{+1.6}	8.9 ^{+1.0}	40.3 ^{+1.6}	20.7 ^{+2.2}	28.4 ^{+1.9}	23.7 ^{+2.1}
Genesis-32B	25.7 ^{+1.4}	18.9 ^{+1.0}	37.1 ^{+1.6}	29.4 ^{+2.2}	26.5 ^{+2.0}	34.0 ^{+1.0}

Table 4: Rubric-based RL vs. RLVR at 32B, holding data and optimizer constant.

Reward (32B)	ZeroBench	ZB-Sub	MathVista	MMMU	MMMU-Pro	EMMA	LogicVista
RLVR	3	28.44	82.20	75.22	66.42	45.23	66.22
Rubric	3	29.94	82.40	75.44	66.47	46.48	68.46

Table 5: Comparison of policy-optimization algorithms under rubric-based reward (4B).

RL Algorithm (4B)	ZeroBench	ZB-Sub	MathVista	MMMU	MMMU-Pro	EMMA	LogicVista
SAPO	0	20.96	77.2	63.39	52.20	34.61	59.06
GRPO	0	17.07	77.6	63.44	53.87	32.78	59.96

largest improvements on those that demand intermediate reasoning quality (LogicVista +2.24, EMMA +1.25, ZeroBench-Sub +1.50).

Choice of RL Algorithm. We compare GRPO and SAPO under the same rubric reward at the 4B scale. As shown in Table 5, the two algorithms exhibit complementary strengths. Neither algorithm dominates across all benchmarks, indicating that the choice between GRPO and SAPO may depend on the target task distribution rather than one being universally superior under rubric-based supervision.

Training Stage Composition. We compare three training recipes on the same Qwen3-VL-32B initialization. We find that RL only achieves the strongest results, while inserting an SFT stage before RL hurts performance on math-heavy benchmarks. We attribute this to SFT pushing the policy toward imitating annotated reasoning chains too rigidly, constraining the exploration that rubric-based RL relies on for discovering alternative valid

derivations. Skipping SFT preserves a broader prior over reasoning paths and allows the rubric reward to freely shape the learned strategies.

5 Conclusion

In this paper, we propose Visual Augmentation Primitives, a generation-based data construction framework that synthesizes diverse photorealistic training samples with coherent multimodal reasoning chains, unifying visual perception with logical reasoning at every step. We further introduce multimodal chain-of-thought rubric-based reinforcement learning, which extends rubric-based reward design from the text-only domain to visual reasoning by scoring intermediate steps for logical validity, completeness, and consistency. To systematically evaluate this frontier, we construct the UNREAL benchmark covering four STEM families with controllable difficulty levels, enabling fine-grained analysis of both hallucination robustness and logical reasoning. Experiments show that

a 4B model built on this framework surpasses 8B baselines and scales consistently to 32B across visual reasoning and VQA benchmarks.

6 Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2024YFE0210600), and the Natural Science Foundation of China under Grant Nos. 72225011 and 72434005.

Limitations

Our pipeline relies on commercial models (GPT-Image, Gemini, Qwen), limiting full reproducibility. UNREAL contains 633 samples across five STEM families, and scaling to more domains remains future work. The LLM-based rubric judge may introduce evaluation noise. We validate only on Qwen3-VL (4B, 8B, 32B), and the pipeline currently handles single-image editing without video or multi-image support.

References

- Anthropic. 2025. [Claude 4.7](#). Model card and technical overview.
- Black Forest Labs. 2024. [FLUX.1](#). Open-weight text-to-image model family.
- Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*.
- Liang Chen, Weichu Xie, Yiyang Liang, Hongfeng He, Hans Zhao, Zhibo Yang, Zhiqi Huang, Haoning Wu, Haoyu Lu, Yiping Bao, and 1 others. 2026. Babyvision: Visual reasoning beyond language. *arXiv preprint arXiv:2601.06521*.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024b. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.
- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multi-modal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.
- Google DeepMind. 2025. [Gemini 3.5](#). Technical report.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. 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.
- Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. 2025. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*.
- Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.
- Fangyu Liu, Guy Emerson, and Nigel Collier. 2023a. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pages 23439–23554.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter

- Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521.
- OpenAI. 2025a. [Introducing 4o image generation](#). Blog post introducing native image generation in GPT-4o.
- OpenAI. 2025b. [Introducing GPT-5.5](#). Technical report.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pages 1862–1874.
- Qwen Team. 2025. [Qwen 3.7 technical report](#). Technical report.
- Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, and 1 others. 2025. Zerobench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*.
- Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024a. Visual cot: Advancing multimodal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9568–9578.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024a. [Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations](#). In *Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, and 1 others. 2024b. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. In *International Conference on Learning Representations*, volume 2024, pages 24025–24057.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits its reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*.
- Fan Yang, Shurong Zheng, Hongyin Zhao, Yufei Zhan, Xin Li, Yousong Zhu, Chaoyang Zhao Ming Tang, and Jinqiao Wang. 2026a. Tracevision: Trajectory-aware vision-language model for human-like spatial understanding. *arXiv preprint arXiv:2602.19768*.
- Fan Yang, Yousong Zhu, Xin Li, Yufei Zhan, Hongyin Zhao, Shurong Zheng, Yaowei Wang, Ming Tang, and Jinqiao Wang. 2026b. Focus: Unified vision-language modeling for interactive editing driven by referential segmentation. *Advances in Neural Information Processing Systems*, 38:33408–33437.
- Fan Yang, Yousong Zhu, Yufei Zhan, Hongyin Zhao, Xin Li, Yaowei Wang, Ming Tang, Xin Ning, and Jinqiao Wang. 2026c. Seg-llava: empowering pixel-level understanding with large vision language model. *Pattern Recognition*, page 113560.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. 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.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

A Data Sources and Construction Details

Across all sources our pipeline normalises questions into eight types: multiple-choice, numerical computation, symbolic reasoning, diagram interpretation, graph reading, spatial reasoning, process analysis, and comparative judgment. Figure 5 visualizes the frequency distribution of question keywords across the full training corpus, where word size is proportional to occurrence count. High-frequency terms such as “number”, “value”, “total”, “left”, and “select” reflect the quantitative and spatial reasoning focus of the dataset, while domain-specific terms such as “square”, “color”, “diagram”, and “pattern” illustrate the visual grounding required by the questions.

5. **Modifier verification.** Two independent Qwen-VL verifiers $\mathcal{M}_1, \mathcal{M}_2$ read I' and the rewritten question q' to extract the *actual* new answer:

$$\tilde{a}_j = \mathcal{M}_j(I', q'), \quad c_j = \text{conf}_j \in \{\text{low}, \text{med}, \text{high}\}. \quad (4)$$

We accept the record iff

$$\text{eq}(\tilde{a}_1, \tilde{a}_2) \wedge \max_j c_j \geq \text{med} \wedge \phi(I') = 1. \quad (5)$$

Stage 2: Visual Diversification and CoT Annotation. Each accepted Stage-1 record $(I', q', a', \mathcal{C})$ is then expanded in five more steps.

1. **Selector / axis planner.** A Qwen-VL planner produces K valuations $\{v^{(k)}\}_{k=1}^K$ over the free axes $\mathcal{V}$, together with slot assignments $\{s^{(k)}\}$. We optimise for inter-plan diversity in axis space while leaving every $\alpha \in \mathcal{F}$ untouched:

$$\{v^{(k)}\} = \arg \max_{\{v\}} \sum_{k \neq k'} d_{\mathcal{V}}(v^{(k)}, v^{(k')}). \quad (6)$$

2. **Template instantiation.** The question template T_q and answer template T_a stored in $\mathcal{C}$ are slot-filled with $s^{(k)}$:

$$q^{(k)} = T_q(s^{(k)}), \quad a^{(k)} = T_a(s^{(k)}, a'). \quad (7)$$

3. **Call 2b — Edit Prose.** A Qwen text model writes a 150–260-word paragraph $\pi^{(k)}$ that paints the new background, surface, lighting, viewpoint, etc. implied by $v^{(k)}$, while paraphrasing $\mathcal{F}$ into a closing integrity clause.
4. **Image generation & verification.** The same image backend produces $I''_k = \mathcal{G}_{\text{edit}}(I', \pi^{(k)})$. A Qwen-VL verifier then answers four yes/no checks $(A_1, \dots, A_4)$ (CLASS, RELATION, READABILITY, NATURAL), and we accept the variant iff $\bigwedge_{i=1}^4 A_i = \text{yes}$.
5. **Call 4 — CoT writer.** A Gemini-3.1-pro vision-language model takes $(I''_k, q^{(k)}, a^{(k)})$ together with soft hints from $\mathcal{C}$.disambiguators and emits a brief reasoning chain $z^{(k)}$ of 3–6 sentences, ending with the literal token ****Final Answer: $a^{(k)}$ ****.

The final supervised training tuple is $(I''_k, q^{(k)}, a^{(k)}, z^{(k)})$.

A.3 Prompt Templates

For reproducibility, this subsection reproduces the (system, user) prompt skeletons used at every model call in Appendix A.2. Placeholders in {braces} are filled at runtime; the slot syntax $\langle S \rangle, \langle SURF \rangle, \langle N \rangle$ is the slot vocabulary defined in Appendix A.2.

Stage 1 — Call 1: Analyzer (Qwen-VL).

System prompt

You are an analytical assistant for an image-editing data generation pipeline. Given an image and a visual question with its ground-truth answer, your job is to produce a structured analysis JSON that downstream code uses to design controlled edits.

Core framework: 5 edit ops $\times$ 4 answer_effect_types. The five ops are remove, add, modify_property, substitute_category, context_change. The four answer-effect types are delta, literal_replace, flip_boolean, categorical_shift. Every feasible edit MUST change the answer; identity-preserving operations go to Stage 2.

Test-point analysis. Classify the structural difficulty into one of {dual_value_disambiguation, real_vs_depicted, partial_visibility, category_filtering, lookalike_confusion, spatial_ordering, composition_state, text_in_context, other}. The test-point structure must remain visible after every edit.

Stage-2 schema. Produce a frozen_predicate (answer_function, must_preserve, verifier_checks, slots), a slot-templated question_template and answer_template, and an axis_table classifying each of the 19 universal axes as free/frozen/depends. Return ONLY valid JSON; no markdown fences, no commentary.

User prompt

Original question: {question}
Original answer: {answer}

Analyze this image and produce JSON with EXACTLY these fields and nothing else: image_type, scene_description, question_analysis, test_point_analysis, disambiguators_to_preserve, feasibility_assessment, feasible_edits (3–5 entries spanning ≥ 2 effect types), frozen_predicate, question_template, answer_template, axis_table (all 19 axes in fixed order).

Stage 1 — Call 2a: Edit-Prose Writer (Qwen text).

System prompt

You write image editing instructions for an AI image editor (GPT-Image-2). Your output is a SINGLE FLOWING PARAGRAPH of natural English prose, 150–250 words.

Strict format rules. No bullet points, no labels (no ‘SCENE:’, ‘EDIT:’, ‘KEEP:’), no section headers, no line breaks. Express constraints with subordinate clauses (“while keeping...”, “without altering...”, “so that...”) and emphasise critical constraints with “exactly”, “only”, “still”.

The paragraph must (1) describe the scene including the test-point structure, (2) specify the single edit operation with precise spatial detail, (3) weave the lock list into the description, and (4) close with a one-clause integrity statement.

User prompt

SCENE CONTEXT: {scene_description}
 TEST POINT: {core_challenge} (must remain visible)
 EDIT OPERATION: type={op}, params={params}, expected_delta={expected_visual_delta}
 MUST NOT CHANGE: {lock_list} ∪ {disambiguators}
 INTEGRITY CONSTRAINT: {edit_must_operate_within}
 Now write the single flowing paragraph for GPT-Image-2. Output ONLY the paragraph, no other text.

Stage 1 — Modifier: Dual-Validator (Qwen-VL × 2).

System prompt

You are reading the actual answer from an edited image.

User prompt

Background: the image shown was produced by editing an original photograph.
 The editing instruction asked to: {rationale_natural}
 If the edit succeeded perfectly, the answer would be: {designed_answer}.

However, AI image editing may not be perfect.

Your task: look at THIS image and determine the ACTUAL answer to the following question, based on what you literally see.
 Question: {new_question}

Output strictly in this format:
 ACTUAL_ANSWER: <value>
 CONFIDENCE: high | medium | low
 REASONING: <one sentence>

Stage 2 — Selector: Axis Planner (Qwen-VL).

System prompt

You are an image augmentation strategy planner. You receive an already-edited image (Stage 1 output) plus a per-image analysis (frozen_predicate + axis_table). Your job is to produce N visually-distinct edit plans for Stage 2. Each plan picks one concrete value for every FREE axis and one slot-class for every replaceable slot.

Hard constraints: (1) FROZEN axes must NOT be touched; (2) verifier_checks must still hold true; (3) stay strictly within the natural-photo distribution; (4) locked slots keep their default class.

Optimization: maximise per-axis aggressiveness (A) and inter-plan diversity in axis space (B), while keeping each plan intra-coherent (C). Also emit a 3–5-tag diversity_signature per plan. Return ONLY valid JSON.

User prompt

scene_description, test_point, frozen_predicate (answer_function, must_preserve, verifier_checks, slots), and the per-image axis_table are injected. Task: produce {N=4} maximally diverse edit plans, each with slot_assignments, axis_values, diversity_signature, and a one-sentence rationale. Output JSON only.

Stage 2 — Call 2b: Diversification Prose (Qwen text).

System prompt

You write image editing instructions for an AI image editor (GPT-Image-2). Your output is a SINGLE FLOWING PARAGRAPH of 150–260 words. Every FREE axis must be addressed with at least one concrete adjective or noun phrase.

Strength rules. Strictly avoid softeners (“subtle”, “slightly”, “a touch of”, “gentle”, “softly”, ...) and pre-fer amplifiers (“completely”, “entirely”, “dramatically”, “fully”, “transformed into”).

Natural-photo rule. The result must look like a real photograph; no illustration / cel-shaded / surreal style, no dramatic single-color stage lighting, no extreme weather.

Frozen predicate. Paraphrase the must_preserve list into the closing of the paragraph using positive language (“the scene still shows...”). Do NOT mention the question or the answer.

User prompt

BASE SCENE: {scene_description}
 AXIS VALUES + SLOT CLASSES TO APPLY: {slot_assignments} ∪ {axis_values} (every entry must appear in the prose)
 FROZEN PREDICATE — MUST be preserved (answer logic = {answer_function}): {must_preserve}

Now write the single flowing paragraph. Open with the new background and surface plus the slot subjects; cover every free axis with a concrete adjective; close with a positive paraphrase of the must_preserve constraints. Output ONLY the paragraph.

Stage 2 — Verifier (Qwen-VL, 4 yes/no checks).

System prompt

You are a strict visual verifier. You will be shown an image and 4 short yes/no questions. Answer each one with a single word (yes or no). Do not explain. Do not hedge. Use only the visual evidence in the image. If a question’s premise can’t be evaluated, answer “no”.

User prompt

Look at the image and answer each question with one word: yes or no.

Q1 (CLASS): {class_check}
 Q2 (RELATION): {relation_check}
 Q3 (READABILITY): {readability_check}
 Q4 (NATURAL): “Is this a plausible real photograph (not illustration / surreal)?”

Output EXACTLY A1: ... A2: ... A3: ... A4: ... (one yes/no per line).

Stage 2 — Call 4: CoT Writer (Gemini-3.1-pro).

System prompt

You are a careful visual reasoning assistant. You produce a brief, accurate chain of thought for a visual question, ending with the final answer.

Style requirements. Write in natural, flowing English (not numbered steps, not bullet lists). Be concise (3–6 short sentences). Your reasoning must lead to the supplied final answer. You may briefly note distractors or features that distinguish the target if it makes the reasoning clearer. End with the line ****Final Answer: <answer>****.

User prompt

Question: {final_question}

Look at the image and reason briefly to the answer.

Helpful hints (use only if natural): {disambiguators[:3]}; {core_challenge}.

End your response with: ****Final Answer: {final_answer}****.

B Multimodal Chain-of-Thought Rubric-Based Reinforcement Learning

This appendix provides the full mathematical formulation of our multimodal chain-of-thought rubric-based reinforcement learning framework. We first introduce the preliminaries of reinforcement learning with verifiable rewards (Section B.1), then formalize rubric-based reinforcement learning (Section B.2), and finally present our multimodal extension with detailed derivations (Section B.3).

B.1 Preliminaries: Reinforcement Learning with Verifiable Rewards

We formulate visual question answering as a token-level Markov decision process. Given an image I and a question q , the policy π_θ autoregressively generates a response $o = (o_1, o_2, \dots, o_T) \in \mathcal{V}^{\leq T_{\max}}$ over a vocabulary $\mathcal{V}$, with

$$\pi_\theta(o \mid I, q) = \prod_{t=1}^T \pi_\theta(o_t \mid I, q, o_{<t}). \quad (8)$$

A scalar reward $R : \mathcal{V}^{\leq T_{\max}} \rightarrow \mathbb{R}$ is assigned to the complete response.

Policy Optimization. Starting from a supervised fine-tuned (SFT) reference $\pi_{\text{ref}} := \pi_{\text{sft}}$, we maximise the expected reward subject to a per-prompt KL constraint to the reference policy:

$$\begin{aligned} \mathcal{J}(\theta) = & \mathbb{E}_{(I,q) \sim \mathcal{D}} \mathbb{E}_{o \sim \pi_\theta(\cdot \mid I,q)} [R(o)] \\ & - \beta_{\text{kl}} \mathbb{E}_{(I,q) \sim \mathcal{D}} D_{\text{KL}}(\pi_\theta(\cdot \mid I,q) \parallel \pi_{\text{ref}}(\cdot \mid I,q)), \end{aligned} \quad (9)$$

with $\beta_{\text{kl}} > 0$ controlling the strength of the regulariser and $\mathcal{D}$ the training distribution.

Group Relative Policy Optimization. For each prompt (I, q) we sample a group $\mathcal{O} = \{o^{(i)}\}_{i=1}^G$ from the behaviour policy π_{old} . The (response-level) advantage is normalised within the group,

$$\hat{A}^{(i)} = \frac{R(o^{(i)}) - \mu_{\mathcal{O}}}{\sigma_{\mathcal{O}} + \epsilon}, \quad (10)$$

where $\mu_{\mathcal{O}} = \frac{1}{G} \sum_{i=1}^G R(o^{(i)})$, $\sigma_{\mathcal{O}} = \sqrt{\frac{1}{G} \sum_{i=1}^G (R(o^{(i)}) - \mu_{\mathcal{O}})^2}$, and $\epsilon > 0$ is a numerical stabiliser. The policy is updated by minimising the PPO-style clipped surrogate

$$\begin{aligned} \mathcal{L}_{\text{GRPO}}(\theta) = & -\mathbb{E}_{(I,q), \mathcal{O}} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|o^{(i)}|} \right. \\ & \left. \sum_{t=1}^{|o^{(i)}|} \min \left(\rho_t^{(i)}(\theta) \hat{A}^{(i)}, \right. \right. \\ & \left. \left. \text{clip}(\rho_t^{(i)}(\theta), 1-\epsilon_c, 1+\epsilon_c) \hat{A}^{(i)} \right) \right], \end{aligned} \quad (11)$$

with per-token importance ratio

$$\rho_t^{(i)}(\theta) = \frac{\pi_\theta(o_t^{(i)} \mid I, q, o_{<t}^{(i)})}{\pi_{\text{old}}(o_t^{(i)} \mid I, q, o_{<t}^{(i)})}, \quad (12)$$

and clip range $\epsilon_c \in (0, 1)$.

Standard Verifiable Rewards. RLVR uses a binary outcome signal on the extracted final answer,

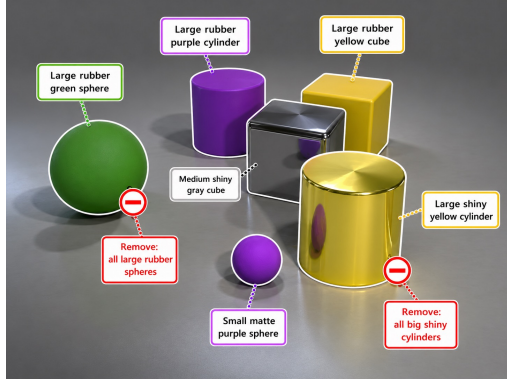
$$R_{\text{RLVR}}(o) = \mathbf{1}[\text{extract}(o) \equiv a^*], \quad (13)$$

where $\text{extract}(\cdot)$ parses the answer span from o and $\equiv$ denotes semantic equivalence. This signal is silent about the quality of intermediate reasoning.

B.2 Rubric-Based Reinforcement Learning

Rubric-based reinforcement learning (Gunjal et al., 2025; Lightman et al., 2024; Wang et al., 2024a) extends RLVR with fine-grained reward signals over the quality of the reasoning process. Specifically, a set of rubrics $\mathcal{R} = \{c_1, \dots, c_M\}$ is predefined, and a judge model $\mathcal{J}$ scores the policy model’s complete output o against each rubric independently:

$$r_m = \mathcal{J}(o \mid c_m, q, a^*), \quad m = 1, \dots, M, \quad (14)$$



Question

Subtract all large rubber spheres. Subtract all big shiny cylinders. How many objects are left?

Genesis

Overall observation: The image contains 6 objects with different shapes, colors, sizes, and material properties.

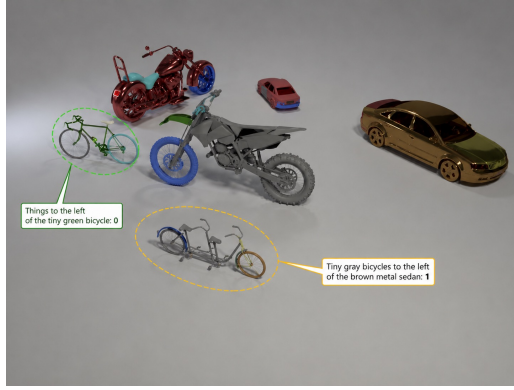
Key element identification: The objects are A green sphere (large, rubber-like) A purple cylinder (large, matte/rubber) A yellow cube (large, matte/rubber) A gray cube (medium, shiny/metallic) A yellow cylinder (large, shiny) A small purple sphere (small, matte)

Reasoning based on the question: First, subtract all large rubber spheres. The large rubber-like green sphere matches this condition, so it is removed. Next, subtract all big shiny cylinders. The large shiny yellow cylinder matches this condition, so it is removed.

The remaining objects are the purple cylinder, the yellow cube, the gray cube, and the small purple sphere. Therefore, **4** objects are left.

<answer>**4**</answer>

Figure 6: An example of a compositional visual reasoning dataset sample, where answering the question requires reasoning over object attributes such as color, shape, size, and material.



Question

Is the number of tiny gray bicycles that are on the left side of the brown metal sedan greater than the number of things that are to the left of the tiny green bicycle?

Genesis

Overall observation: The image shows several vehicles and bicycles arranged around a brown metal sedan on the right side.

Key element identification: The brown metal sedan is on the right. To its left, there are several objects, including a red motorcycle, a small red car, a gray dirt bike, a tiny green bicycle, and a tiny gray bicycle/tandem bicycle.

Reasoning based on the question: First, count the tiny gray bicycles on the left side of the brown metal sedan. There is 1 tiny gray bicycle visible to the left of the sedan. Next, count the things to the left of the tiny green bicycle. The tiny green bicycle is already near the left side of the scene, and there are no objects clearly to its left, so the count is 0.

Since 1 is greater than 0, the number of tiny gray bicycles on the left side of the brown sedan is greater.

<answer>**A**</answer>

Figure 7: An example of a spatial visual reasoning dataset sample, where answering the question requires reasoning over relative object positions and counting objects satisfying spatial constraints.

and the trajectory-level process reward is a weighted aggregation of the per-rubric scores:

$$R_{\text{rub}}(o) = \sum_{m=1}^M \beta_m r_m. \quad (15)$$

Compared with the binary signal R_{RLVR} that only checks final-answer correctness, rubric-based rewards can distinguish between outputs that are answer-correct but reasoning-flawed and those that are reasoning-complete and well-grounded, thereby providing denser and more informative gradient signals for policy optimization.

B.3 Multimodal Chain-of-Thought Rubric-Based RL

We extend rubric-based reinforcement learning to the multimodal reasoning setting. For a training instance (I', q', a') with a reference reasoning chain $z' = (z'_1, \dots, z'_K)$, a model response is decomposed as $o = (s_1, \dots, s_N, a)$. We use Qwen3-32B

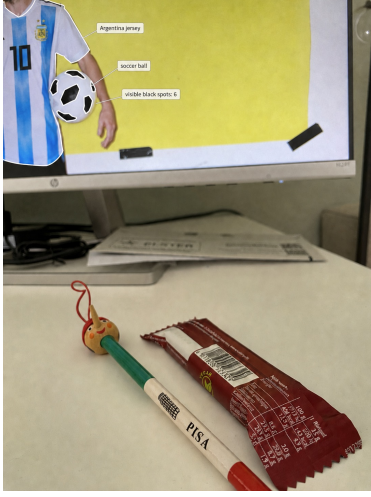
as the judge model to score o along three dimensions.

(1) Logical Validity. The judge model examines each individual reasoning step s_k and evaluates whether the conclusion drawn at that step actually follows from the preceding premises or results, producing an overall logical validity score $u_{\text{log}} \in [0, 1]$.

(2) Completeness. This dimension measures how well o covers the reference reasoning chain z' . Let $\Pi(N, K) = \{\sigma : [K] \rightarrow [N] \mid \sigma \text{ monotone, partial, injective}\}$ be the set of monotone partial injections and $\text{sim}(\cdot, \cdot) \in [0, 1]$ a semantic similarity. We first solve the optimal alignment

$$\sigma^* = \arg \max_{\sigma \in \Pi(N, K)} \sum_{j \in \text{dom}(\sigma)} \text{sim}(s_{\sigma(j)}, z'_j), \quad (16)$$

and compute the completeness score $u_{\text{comp}} \in [0, 1]$ as the mean coverage of reference steps under σ^* ;



Question

What is the number of black spots on the ball?

Genesis

Overall observation: The image shows a soccer ball on the left side of the screen, being held by a person wearing an Argentina jersey.

Key element identification: The ball has a classic black-and-white pattern, with black pentagons and white hexagons. The relevant task is to count the black spots, or pentagons, on the visible part of the ball.

Reasoning based on the question: Looking closely at the visible portion of the ball, I can see 6 distinct black pentagons. Since only part of the ball is visible from this angle, I should count only the spots that can be clearly seen in the frame.

There are 6 visible black spots on the soccer ball.

<answer>6</answer>

Figure 8: An example of a visual counting dataset sample, where answering the question requires identifying and counting target objects in the scene.

any reference step not matched by a response step contributes zero.

(3) Process Consistency. The judge model checks whether adjacent reasoning steps contain logical contradictions or circular reasoning, producing an overall consistency score $u_{\text{con}} \in [0, 1]$.

Reward Aggregation. The three dimension scores are combined into a scalar rubric reward via weighted summation:

$$R_{\text{rub}}(o) = \beta_{\text{log}} u_{\text{log}} + \beta_{\text{comp}} u_{\text{comp}} + \beta_{\text{con}} u_{\text{con}}, \quad (17)$$

where $\beta_{\text{log}} + \beta_{\text{comp}} + \beta_{\text{con}} = 1$ and $\beta_{\bullet} \geq 0$. The full reward signal sums three components,

$$R(o) = \alpha_{\text{ans}} R_{\text{ans}}(o) + \alpha_{\text{rub}} R_{\text{rub}}(o) + \alpha_{\text{fmt}} R_{\text{fmt}}(o), \quad (18)$$

with

$$R_{\text{ans}}(o) = \mathbf{1}[\text{extract}(o) \equiv a'], \quad R_{\text{fmt}}(o) \in \{0, 1\}, \quad (19)$$

and weights satisfying $\alpha_{\text{ans}} > \alpha_{\text{rub}} > \alpha_{\text{fmt}} > 0$, so that final-answer correctness remains the dominant objective while reasoning quality receives a meaningful gradient.

Training Objective. Starting from $\pi_{\text{ref}} = \pi_{\text{sft}}$ trained on the synthetic corpus $\mathcal{D}_{\text{syn}}$, we minimise the GRPO loss in Eq. (11) with the group advantages $\hat{A}^{(i)}$ of Eq. (10) computed from the total reward $R(o^{(i)})$ in Eq. (18), and the per-prompt KL

Hyperparameter	4B	8B	32B
Learning rate	1×10^{-6}	1×10^{-6}	1×10^{-6}
Train batch size	32	32	96
PPO mini-batch size	16	16	48
PPO micro-batch / GPU	4	4	1
Rollout group size n	8	8	8
Max prompt length	4096	8192	4096
Max response length	8192	8192	8192
Max tokens / GPU	16,384	16,384	32,768
GPUs	8 (1 node)	8 (1 node)	48 (6 nodes)
Rollout TP	1	1	2

Table 8: Training hyperparameters at three model scales. The 4B and 8B variants share the same base configuration, while the 32B variant uses larger batch sizes, higher per-GPU token budgets, and multi-node training with rollout tensor parallelism.

anchor of Eq. (9):

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{GRPO}}(\theta) + \beta_{\text{kl}} \mathbb{E}_{(I', q') \sim \mathcal{D}_{\text{syn}}} D_{\text{KL}}(\pi_{\theta}(\cdot | I', q') \parallel \pi_{\text{ref}}(\cdot | I', q')). \quad (20)$$

B.4 Training Configuration

Table 8 lists the full hyperparameters used for multi-modal chain-of-thought rubric-based reinforcement learning across all model scales.

C UNREAL Benchmark Details

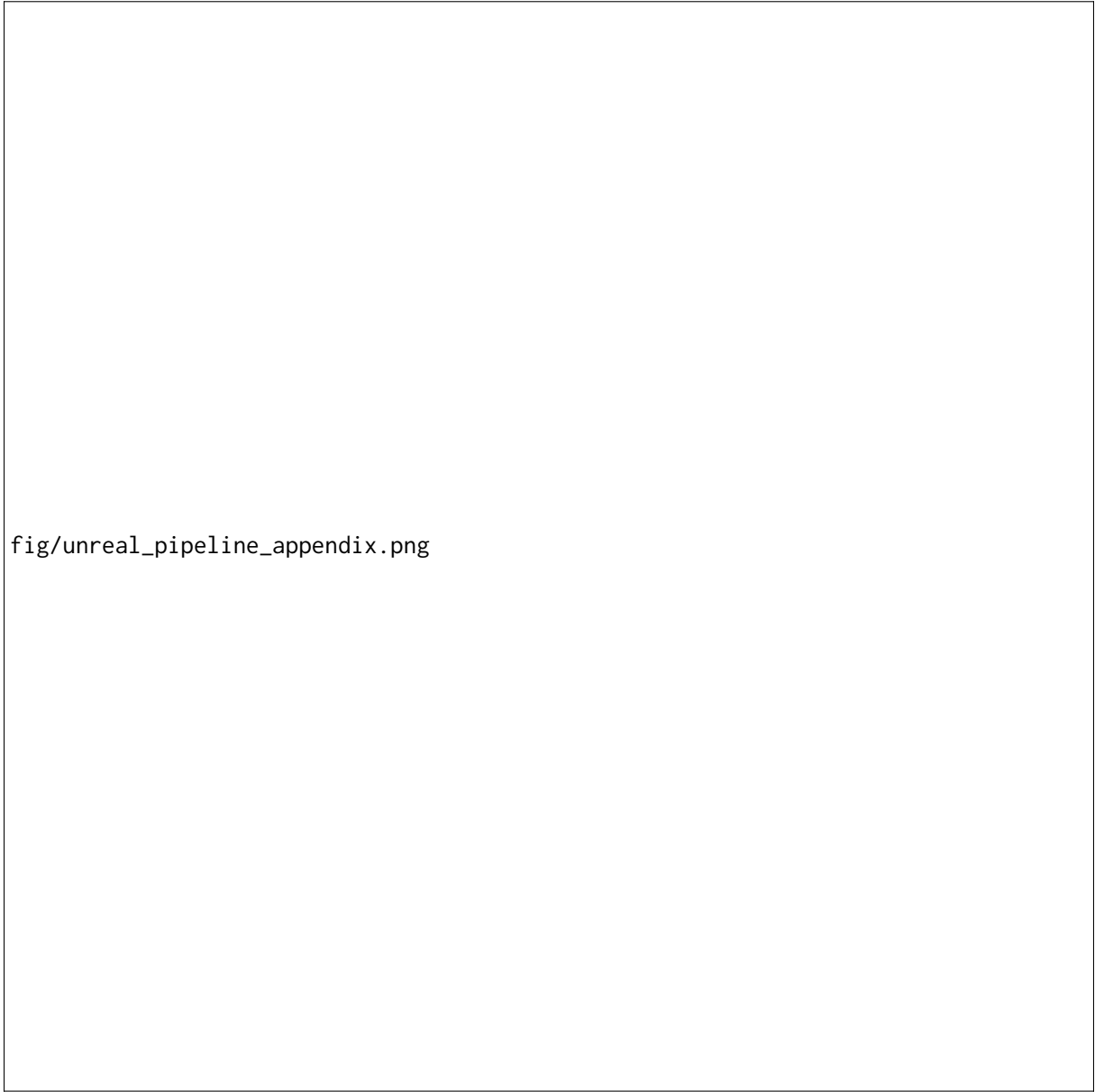
This appendix provides the full topic distribution, mechanism taxonomy, and construction details of the UNREAL benchmark. A high-level summary of the entire construction pipeline, the family composition, and the key design characteristics is given in Figure 9.

Family	Seeds	Stage 1 Records
Physics	455	805
Biology	600	–
Chemistry	693	–
Mathematics	~940	–

Table 9: UNREAL benchmark composition by STEM family.

C.1 Overview

UNREAL contains questions across four STEM families with a total of 633 evaluation samples via Stage 1 of the Visual Augmentation Primitives pipeline. All questions are open-ended and require models to generate answers directly without multiple-choice options. The end-to-end pipeline is illustrated in Figure 9. Figure 10 shows some examples.



fig/unreal_pipeline_appendix.png

Figure 9: Overview of the UNREAL benchmark construction pipeline. **Top:** the four sequential stages — (1) data collection and organisation across the four STEM families, (2) LLM-driven question generation with MCQ→OE conversion, (3) structure-preserving image-to-image augmentation using the Axis Bank, and (4) human and automated quality control. **Middle:** family composition (Physics / Biology / Chemistry / Mathematics) together with high-level statistics. **Bottom:** the five key characteristics — rule competition, structure preservation, high difficulty, fully open-ended format, and broad STEM coverage.

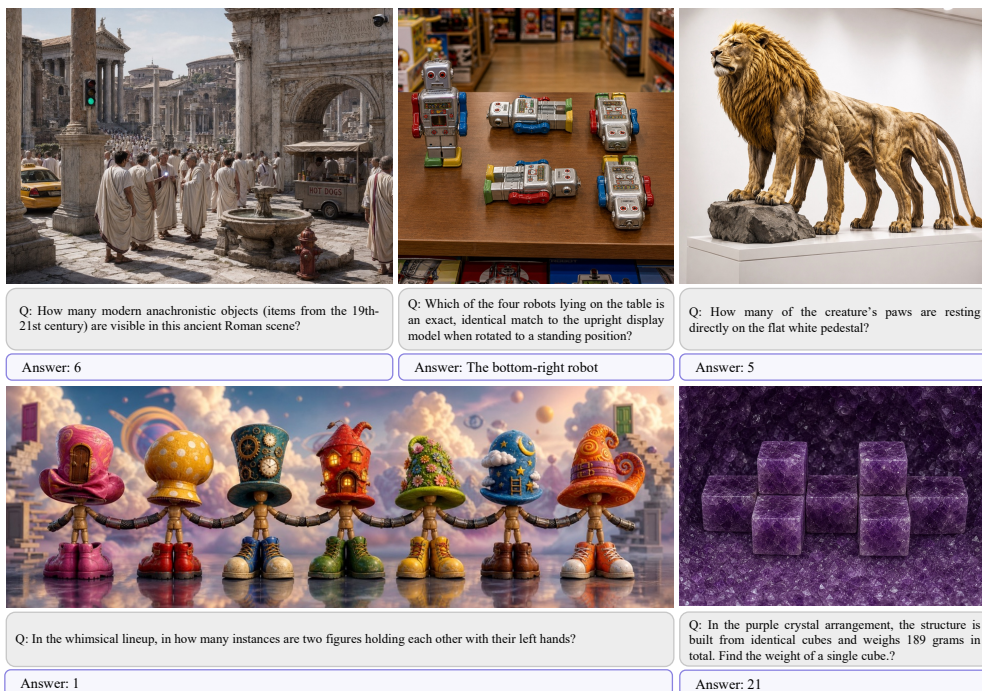


Figure 10: UNREAL benchmark illustration.