ViGoR-Bench: How Far Are Visual Generative Models From Zero-Shot Visual Reasoners?

Anonymous Authors[1]

Abstract

Beneath the stunning visual fidelity of modern AIGC models lies a "logical desert", where systems fail tasks that require physical, causal, or complex spatial reasoning. Current evaluations largely rely on superficial metrics or fragmented benchmarks, creating a "performance mirage" that overlooks the generative process. To address this, we introduce **ViGoR** (**Vi**sion-**G**enerative **R**easoning-centric Benchmark), a unified framework designed to dismantle this mirage. ViGoR distinguishes itself through four key innovations: 1) **holistic cross-modal coverage** bridging Image-to-Image and Video tasks; 2) a **dual-track mechanism** evaluating both intermediate processes and final results; 3) an **evidence-grounded automated judge** ensuring high human alignment; and 4) **granular diagnostic analysis** that decomposes performance into fine-grained cognitive dimensions. Experiments on over 20 leading models reveal that even state-of-the-art systems harbor significant reasoning deficits, establishing ViGoR as a critical "stress test" for the next generation of intelligent vision models.

1. Introduction

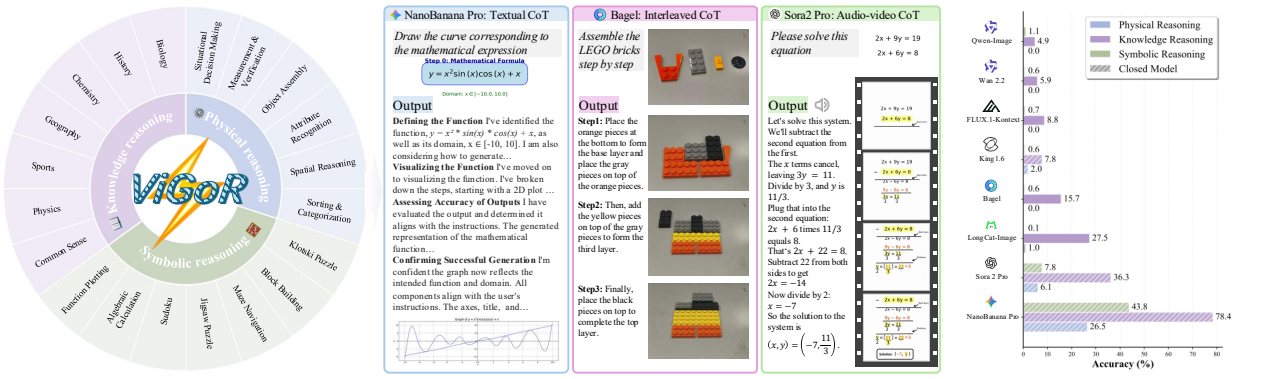
Artificial Intelligence-Generated Content (AIGC) has witnessed remarkable growth, evolving from basic pixel-level synthesis to the production of highly sophisticated content. This progress reflects an architectural leap from early GANs (Goodfellow et al., 2020) to modern, high-capacity generative systems (Rombach et al., 2021; Blattmann et al., 2023; Labs, 2025a; Cao et al., 2025; Seedream et al., 2025b). However, such visual excellence often masks a "logical desert": beneath a facade of photorealism, models crumble when faced with tasks requiring deep physical laws or causal reasoning. This deficit is obscured by a performance mirage fostered by traditional metrics like CLIP-Score (Hessel et al., 2021) and FID (Heusel et al., 2017), which prioritize semantic alignment and statistical fidelity over true structural integrity. Notably, a generated image can achieve high statistical similarity to real data while still harboring absurd physical glitches. Consequently, existing metrics fail to distinguish between a model that truly "understands" the physical world and one that merely performs high-dimensional probability tiling. This underscores an urgent need to pivot from evaluating fidelity to rigorously assessing the generative reasoning capabilities that define true visual intelligence.

In response to this "logical desert", benchmarks have evolved beyond pixel fidelity to probe cognitive depth along three dimensions: knowledge breadth, where GenExam (Wang et al., 2025c) audits factual consistency; physical causality, exemplified by KRIS-Bench's (Wu et al., 2025c) focus on dynamic editing; and process-oriented temporal logic, where RULER-Bench (He et al., 2025) and MME-CoF (Guo et al., 2025) validate the continuous "reasoning chain" in dynamic tasks. Despite these advances, the current evaluation landscape remains fragmented. As illustrated in Table 1, existing benchmarks typically operate in *silos-narrowly* restricted to either Image-to-Image (I2I) editing or Video generation (I2V)—lacking a unified mechanism to assess reasoning across diverse modalities. Critically, most frameworks fail to distinguish between the final outcome (the "what") and the generative process (the "how"). This conceptual gap is mirrored in the evaluation domain, while the "VLM-as-a-Judge" paradigm (e.g., GPT-4o (OpenAI, 2025b), Gemini 2.5 Pro (Google, 2025a)) has emerged as the *de facto* standard for scalable evaluation, achieving robust human alignment across multifaceted reasoning tasks remains a persistent bottleneck. **To address these limitations, we introduce ViGoR-Bench (Vision-Generative Reasoning-centric Benchmark),** a comprehensive evaluation framework designed to unveil the "performance mirage." Unlike its predecessors, ViGoR-Bench is defined by three key innovations:

- **Holistic Cross-Modal Coverage:** ViGoR-Bench is the first benchmark to bridge the divide between I2I, Sequen-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the FoGen Workshop at ICML 2026. Do not distribute.



(a) Data distribution (b) Reasoning of generation models (c) Results on ViGoR-Bench

Figure 1. **An overview of ViGoR-Bench.** (a) The data distribution across various domains. (b) Examples of the reasoning process from generation models. (c) Performance comparison of leading models on ViGoR-Bench.

tial I2I (I2Is), and I2V tasks. As shown in Figure 1 (a), it encompasses 20 distinct dimensions—ranging from physical mechanics and social commonsense to complex spatial planning—providing a unified testbed for generative intelligence.

- **Dual-Track Process-Outcome Evaluation:** Moving beyond binary outcome assessment, ViGoR-Bench implements a rigorous *Process plus Result* scoring pipeline, which evaluates not only the final output but also verifies whether intermediate states and logic adhere to physical laws and causal consistency.
- **Evidence-Grounded Automated Alignment:** Leveraging a multi-agent "Evidence-Grounded" judge system, ViGoR-Bench mitigates the subjectivity of LLM evaluators and achieves unprecedented alignment with human experts, demonstrating superior MAE (Mean Absolute Error) and Pearson correlations.
- **Granular Diagnostic Analysis:** Moving beyond standard leaderboards, ViGoR introduces a structured analysis of reasoning failures. It breaks down model performance into specific cognitive dimensions, enabling researchers to pinpoint specific reasoning gaps rather than relying on a single aggregate score.

Through extensive experiments on over 20 leading generative models, we demonstrate that ViGoR-Bench serves as a critical stress test for visual reasoning. Our results reveal that even models with superior visual fidelity can harbor significant deficits in world-knowledge reasoning. This underscores the need for a paradigm shift toward more intelligent vision foundation models, as illustrated in Figure 1. Our key findings are as follows:

- Proprietary models maintain a significant performance lead over their open-source counterparts (see Figure 1).
- Explicit Chain-of-Thought (CoT) prompting enhances the interpretability of the generation process but it does not guarantee an improvement in final accuracy.
- Video generation models often exhibit an "Illusion of Reasoning", where apparent logical consistency does not hold up to rigorous evaluation.

Table 1. Overview of benchmark properties and evaluation settings. Veo† (Wiedemer et al., 2025) conducted quantitative evaluations across seven distinct task categories.

Benchmark	Tasks	Reference		Type			Evaluation	
		Img	*Text*	*I2I*	*I2Is**	*I2V*	*Process*	*Result*
RISE	–	✗	✓	✓	✗	✗	✗	✓
KRIS	22	✗	✓	✓	✗	✗	✗	✓
GIR-Edit	3	✓	✓	✓	✗	✗	✗	✓
UniREdit	18	✓	✓	✓	✓	✗	✓	✓
WiseEdit	–	✗	✓	✓	✓	✗	✗	✓
Veo†	7‡	✓	✗	✗	✗	✓	✓	✓
MME-CoF	12	✗	✓	✗	✗	✓	✓	✓
RULER	40	✓	✓	✗	✗	✓	✓	✗
ViGoR (ours)	**20**	✓	✓	✓	✓	✓	✓	✓

Furthermore, by training models on a specific subtask, we observed that:

- Reward-driven Reinforcement Learning (RL) demonstrates superior potential in advancing visual reasoning capabilities where Supervised Fine-Tuning (SFT) exhibits saturation.
- Training on more challenging, Out-of-Distribution (OOD) data enhances a model's generalization performance on simpler, in-distribution visual reasoning tasks.

2. Related Work

2.1. Visual Generative Model

Text-to-Image and Editing Models. T2I generation has laid the bedrock for visual synthesis, driven by foundational architectures like Stable Diffusion (Rombach et al., 2021) and Flux (Labs, 2025a) which revolutionized generation efficiency. Building on this, conditional editing has evolved through three paradigms: (1) **Spatial Control**, where ControlNet (Zhang et al., 2023) and T2I-Adapter (Mou et al., 2024) inject structural guidance via additional conditions;

(2) **Attention Manipulation**, exemplified by Prompt-to-Prompt (Hertz et al., 2022), which modifies cross-attention maps for zero-shot editing; and (3) **Instruction Following**. Recently, the latter has seen rapid advancements with models like Qwen-Image-Edit (Wu et al., 2025a) and Z-Image (Team, 2025) optimizing semantic precision, while Seed-Edit (Wang et al., 2025b) and the nano-banana family (Gemini 2.5) (Google, 2025b) push the boundaries of sequential consistency and character fidelity.

Unified Vision Models. A recent trend unifies understanding and generation within single architectures, enabling reasoning across modalities. Early works like Unified-IO (Lu et al., 2023a; 2024) handle diverse vision tasks through sequence-to-sequence formulation. Interleaved image-text models represent a paradigm shift: Chameleon (Lu et al., 2023b) employs early-fusion token-based architecture for mixed-modal processing; Emu series (Sun et al., 2024b;a) and SEED-X (Ge et al., 2024) demonstrate unified visual tokenization for coherent multimodal reasoning; Show-o (Xie et al., 2024) achieves flexible control through next-token prediction across modalities. Recent autoregressive approaches like Janus (Wu et al., 2024), Anole (Chern et al., 2024), and VILA-U (Wu et al., 2025d) decouple or unify visual pathways for enhanced reasoning transfer. BAGEL (Deng et al., 2025) demonstrates that scaling unified models with large-scale interleaved data reveals emergent reasoning capabilities, including world-modeling tasks like multiview synthesis and visual navigation. Hybrid architectures such as Transfusion (Zhou et al., 2025), Meissonic (Bai et al., 2024), and Lumina-mGPT (Liu et al., 2024) combine different modeling paradigms for flexible any-to-any generation.

Video Generation Models. Video generation extends to temporal sequences, demanding reasoning about dynamics, physics, and causality. Building upon early diffusion-based work [Singer et al., 2023; Ho et al., 2022], recent models including Sora series (OpenAI, 2024; 2025c), Veo3 (Google, 2024), Kling (Kuaishou Tech., 2024), VideoPoet (Kondratyuk et al., 2024), seedance (Gao et al., 2025; Chen et al., 2025), Lumiere (Bar-Tal et al., 2024) , and Stable Diffusion 4.0 (Yao et al., 2025) demonstrate remarkable temporal coherence and physical plausibility.

2.2. Benchmarks for Visual Generation

Perceptual Quality of Generation. Current benchmarks for generative models predominantly focus on generation quality (FID (Heusel et al., 2017), IS (Salimans et al., 2016), CLIP Score (Hessel et al., 2021)) or specific compositional aspects (T2I-CompBench (Huang et al., 2023), GenEval (Ghosh et al., 2023), TIFA (Hu et al., 2023)).

Reasoning-Centric Evaluation. Generative model evaluation is shifting from perceptual quality to cognitive reasoning. Image benchmarks like GenExam (Wang et al., 2025c), SridBench (Chang et al., 2025b), and WISE now integrate multidisciplinary knowledge, while GIR-BENCH (Li et al., 2025b) and KRIS-Bench (Wu et al., 2025c) demand logical consistency. In the video domain, MME-COF (Guo et al., 2025) and Reasoning via Video test temporal logic, while PICABench (Pu et al., 2025) and RULER-Bench (He et al., 2025) rigorously evaluate adherence to physical laws. Methodologically, OneIG-Bench (Chang et al., 2025a) and WiseEdit (Pan et al., 2025) establish VLM-as-a-Judge as the standard paradigm. To ensure objectivity, works like RISEBench (Zhao et al., 2025) and UniREditBench (Han et al., 2025) validate Human-LLM alignment.

3. ViGoR-Bench

To comprehensively evaluate the reasoning capabilities of generative models, we constructed a diverse benchmark comprising three primary domains: Physical Reasoning, Knowledge Reasoning, and Symbolic Reasoning. Figure 2 (a) summarizes the taxonomy of our benchmark.

3.1. Data Engine

To build the ViGoR-Bench, we designed a data construction pipeline tailored to the unique characteristics of each reasoning domain. Our approach integrates three distinct construction paradigms: (1) Generative Synthesis, which leverages large language models and image generation models to create high-fidelity physical scenarios; (2) Real-world Acquisition, involving authoritative web curation and manual photography to ensure alignment with reality; and (3) Algorithmic Construction, utilizing rule-based engines to produce logically rigorous samples. Figure 3 presents representative examples curated through our data collection pipeline, spanning 20 distinct subdomains across three reasoning domains. Crucially, to guarantee the correctness and reliability of the benchmark, we incorporate a rigorous post-processing verification stage. This includes human-in-the-loop review for semantic consistency and symbolic solver validation for mathematical precision. Unlike previous benchmarks for generative model evaluation, our benchmark provides both referenced ground-truth images and human-verified ground-truth captions where applicable. Below, we detail the sample collection strategies for each subdomain.

Physical Reasoning. The Physical Reasoning subset encompasses scenarios requiring embodied intelligence, including tasks such as Sorting, Categorization, Spatial Reasoning, Attribute Recognition, Object Assembly, Measurement & Verification, and Situational Decision Making. Due to the high cost and complexity of acquiring diverse real-world embodied data, we employ a generative pipeline. First, detailed textual descriptions of physical scenarios and tasks are composed, then further enriched using large language models. These descriptions serve as prompts for state-of-

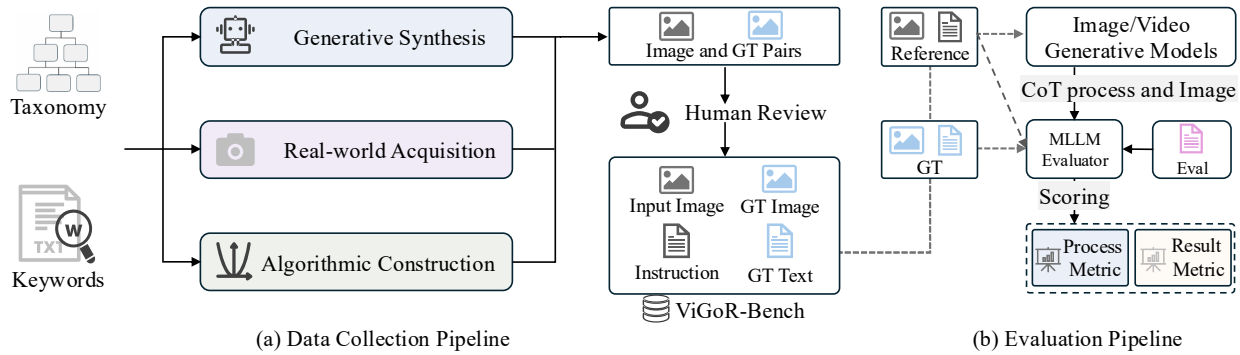


Figure 2. **An overview of the ViGoR-Bench construction and evaluation pipelines.** (a) The benchmark dataset is constructed through a three-pronged approach: generative synthesis, real-world acquisition, and algorithmic generation. All data undergoes human review to establish definitive image-ground truth (GT) pairs. (b) For evaluation, a Multimodal Large Language Model (MLLM) is employed as an automated judge. Conditioned on the ground-truth image, the MLLM assesses both the Chain-of-Thought (CoT) reasoning process and the final output of generative models for images and videos.

the-art generative models, i.e., NanoBanana-Pro (Google, 2025c), which synthesize high-fidelity input images. All generated images are subsequently verified by human annotators for plausibility and relevance. Since all input images are generated, no corresponding ground-truth images exist. Instead, we provide textual ground-truth answers, which are manually annotated or verified by human experts to ensure logical consistency with the visual input.

Knowledge Reasoning. This domain assesses the ability to reason over world knowledge, spanning disciplines such as Biology, Physics, Chemistry, Geography, History, Sports, and Common Sense. A substantial portion of the data is curated from authoritative educational websites and scientific repositories to ensure factual accuracy. All samples are accompanied by human-verified textual answers. For cases where paired datasets exist (e.g., "before-and-after" scientific phenomena), the original ground-truth images are preserved. For other samples, only textual ground-truth is provided.

Symbolic Reasoning. These tasks require precise logical manipulation. For different types of data, we use different approaches. For Physical Puzzles (Klotski Puzzle, Block Building), we emphasize visual realism to evaluate the model's alignment between perception and reasoning. Data is collected in physical environments, where annotators manually solve puzzles and capture the solved states as ground-truth images. For abstract logic tasks (e.g., Sudoku, Maze Navigation, Jigsaw Puzzle, Function Plotting), we employ rule-based algorithms to ensure mathematical rigor and uniqueness of solutions. Input and ground-truth images are algorithmically generated, with no textual ground-truth provided. For algebraic calculation tasks, we write equations, covering linear and quadratic forms, using large language models, and validate their solutions via symbolic solvers. Both the equations and their solutions are rendered as images to serve as input and ground-truth, respectively. Similarly, for function plotting tasks, two-dimensional function expressions are generated. The corresponding curves are plotted using Matplotlib to produce ground-truth images.

3.2. Evaluation Protocol

To provide a comprehensive assessment of visual reasoning capabilities, we establish a dual-track evaluation protocol comprising **Process Metrics** and **Result Metrics**, as shown in Figure 2 (b). This design allows us to scrutinize not only the correctness of the final solution but also the logical coherence of the intermediate reasoning steps. We employ Gemini-2.5-Pro (Comanici et al., 2025) as our **VLM-as-a-Judge** evaluator due to its advanced multimodal understanding capabilities.

Process Metric. This set of metrics is designed to evaluate dynamic outputs, such as video generation models or "thinking" models that produce intermediate reasoning frames. It assesses the quality and trajectory of the reasoning process. The evaluation is formulated as a function of the input and the generated sequence:

$$\mathbf{S}_{\text{Process}} = \text{VLM}(I, P, O_{\text{seq}}, R_i, R_t, \mathcal{T}_{\text{Process}}) \quad (1)$$

where I denotes the input image, P represents the editing prompt, and O_{seq} is the model's output sequence (intermediate frames or video). R_i and R_t denote the visual Ground Truth (GT) image and the textual GT reference, respectively. $\mathcal{T}_{\text{Process}}$ represents the specific evaluation template containing scoring criteria. The output score vector consists of four dimensions:

$$\mathbf{S}_{\text{Process}} = [S_{\text{BC}}, S_{\text{RO}}, S_{\text{VQ}}, S_{\text{RA}}] \quad (2)$$

The final aggregated score is calculated as the mean of these

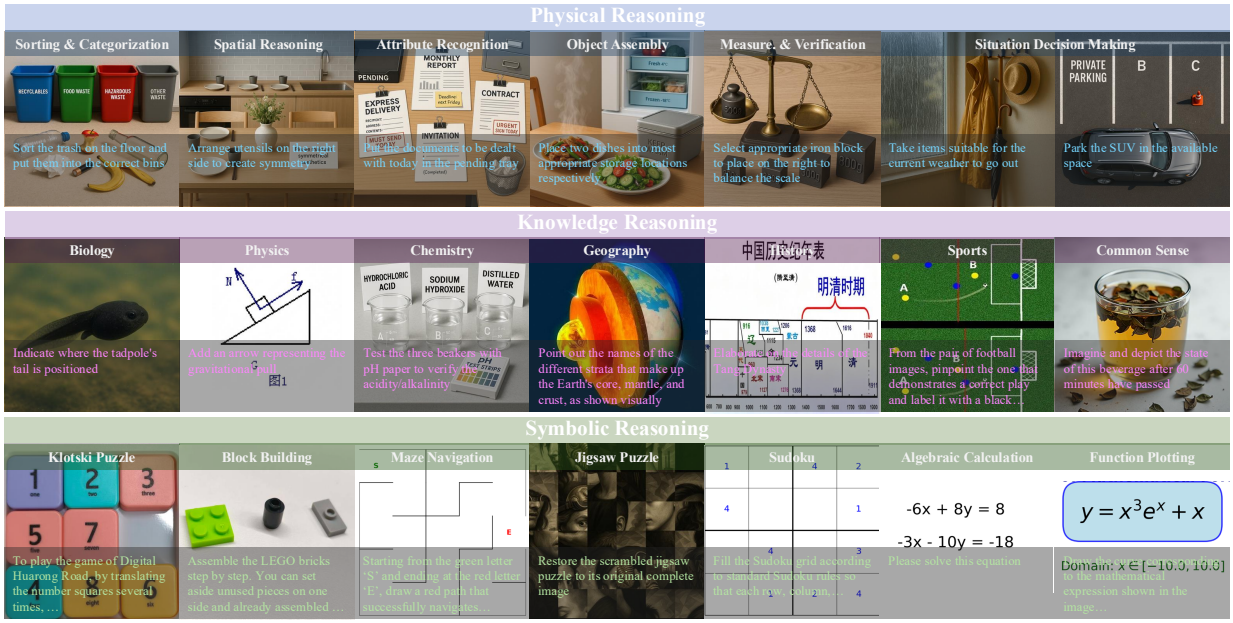


Figure 3. **Overview of the ViGoR-Bench task suite.** We present representative demo cases and their corresponding editing instructions across 20 distinct sub-tasks. These tasks are hierarchically organized into three primary reasoning domains: Physical Reasoning, Knowledge Reasoning, and Symbolic Reasoning.

components:

$$S_{\text{Avg}}^{\text{Process}} = \frac{1}{4}(S_{\text{BC}} + S_{\text{RO}} + S_{\text{VQ}} + S_{\text{RA}}) \tag{3}$$

Each sub-metric is scored on a continuous scale from 0 to 100, defined as follows:

- **Background Consistency (S_{BC}):** Measures the extent to which the main structure of the input image I is preserved across the output sequence O_{seq}. It penalizes unintended modifications to regions unrelated to the reasoning task.
- **Rule Obey (S_{RO}):** Assesses the percentage of frames where edits strictly adhere to the constraints and requirements specified in the instruction P.
- **Visual Quality (S_{VQ}):** Evaluates the fidelity of the generated frames, checking for clarity, sharpness, and the absence of temporal flickering, noise, or artifacts.
- **Reasoning Accuracy (S_{RA}):** Evaluates the efficacy of the progressive edits. It determines whether the model's modifications effectively progress toward the correct solution as defined by R_i and R_t (analogous to "Beneficial Action").

Result Metric. This set of metrics targets the final output—either the static image generated by image editing models or the final frame of a reasoning sequence. It focuses on the validity of the final solution. Similar to the process evaluation, the scoring function is defined as:

$$\mathbf{S}_{\text{Result}} = \text{VLM}(I, P, O_{\text{final}}, R_i, R_t, \mathcal{T}_{\text{Result}}) \tag{4}$$

Table 2. **Reliability Analysis of VLM-as-a-Judge.** We evaluate the alignment between Gemini-2.5-Pro (Comanici et al., 2025) and human experts on a random subset. The table compares performance with and without Ground Truth references across both Process and Result metrics.

Type	Evaluator	GT Ref.	MAE ↓	Acc ↑	Var ↓
Process	Human	✓	0.000	1.000	0.051
	Gemini-2.5-Pro	✗	0.319	0.680	0.039
	Gemini-2.5-Pro	✓	**0.267**	**0.733**	**0.034**
Result	Human	✓	0.000	1.000	0.011
	Gemini-2.5-Pro	✗	0.294	0.705	0.034
	Gemini-2.5-Pro	✓	**0.213**	**0.786**	**0.029**

Here, O_{final} represents the final generated image. The resulting score vector is composed of:

$$\mathbf{S}_{\text{Result}} = [S_{\text{BC}}, S_{\text{RO}}, S_{\text{VQ}}, S_{\text{RS}}] \tag{5}$$

The average performance is derived as:

$$S_{\text{Avg}}^{\text{Result}} = \frac{1}{4}(S_{\text{BC}} + S_{\text{RO}} + S_{\text{VQ}} + S_{\text{RS}}) \tag{6}$$

Distinct from the process metrics, the Result Metrics employ a binary scoring system $\{0, 1\}$ to provide a rigorous pass/fail assessment:

- **Background Consistency (S_{BC}):** A binary check on whether the output image O_{final} retains the structural integrity of the input I, ensuring irrelevant areas remain untouched.

- **Rule Obey ($S_{\mathbf{RO}}$):** Determines if the result complies with the explicit instructions given in P while adhering to essential reasoning constraints (e.g., avoiding wall penetration in maze navigation).
- **Visual Quality ($S_{\mathbf{VQ}}$):** Verifies if the final output maintains high realism and is free from degradation, distortions, or physical implausibility.
- **Reasoning Success ($S_{\mathbf{RS}}$):** The critical measure of task completion. It evaluates whether the final state matches the reference answer (R_t) or the reference image (R_i), signifying a correct solution to the reasoning problem.

3.3. Reliability Analysis

To validate the trustworthiness of our evaluation pipeline, we conducted a rigorous meta-evaluation comparing our VLM-as-a-Judge against human experts.

Experimental Setup. We constructed a random "tiny split" from our benchmark generation results, comprising 1,080 final result outputs (yielding 4,272 metric evaluation instances) and 540 process sequences (yielding 1,064 metric evaluation instances). Three human experts independently scored these instances using the same input information and templates as the VLM. The average of the three human scores serves as the "Gold Standard." Simultaneously, Gemini-2.5-Pro evaluated the same set over three independent runs under two settings: *with* Ground Truth (GT) references and *without* them.

Metrics. We assess reliability via three dimensions: (1) MAE (Mean Absolute Error), measuring the distributional distance between the VLM's average score and the human average; (2) Accuracy, measuring categorical agreement. For the continuous Process Metrics (0–100), scores were discretized into three intervals—Bad [0, 33], Moderate [34, 67], and Good [68, 100]—to calculate alignment; (3) Variance, quantifying the stability and consistency of the evaluator across three runs.

Analysis. As presented in Table 2, the results support three key conclusions:

- **High Human Alignment:** When provided with GT references, Gemini-2.5-Pro achieves high accuracy (73.3% for Process, 78.6% for Result) and low MAE, demonstrating that the VLM's judgment distribution closely approximates that of human experts.
- **Criticality of Ground Truth:** There is a significant performance gap between the *w/* and *w/o* GT settings. The inclusion of GT references substantially reduces MAE and Variance, confirming that providing golden references is essential for stabilizing VLM judgments.
- **Stability:** The variance of the VLM is comparable to, and in some cases lower than, the inter-annotator variance of human experts. This indicates that our automated pipeline offers a consistency level competitive with human consensus, making it reliable for large-scale evaluation.

4. Experiment

Implementation Details. We conduct a comprehensive evaluation of leading open-source and proprietary models, categorized into four distinct groups: **Image Editing Models**, **Unified Models without Chain-of-Thought**, **Unified Models with CoT**, and **Video Generation Models**. To strictly assess generalization in a zero-shot setting, we utilize official checkpoints and standard APIs, adhering to the default inference parameters recommended by their respective official implementations to ensure a fair comparison. For clarity and consistency across diverse metrics, all reported scores are normalized to a 100-point scale.

4.1. Main Results

Table 3 presents the quantitative performance of leading models across our proposed metrics. Complementing this, Figure 4 provides a qualitative comparison of representative cases. Our analysis yields three critical observations regarding the current state of visual reasoning.

Proprietary models maintain a significant performance lead over their open-source counterparts. Proprietary unified models continue to maintain a substantial lead over open-source counterparts. Nano Banana Pro consistently secures the top performance across most metrics. This quantitative gap is visually corroborated in Figure 4. In complex domains such as *Physical Reasoning* and *Symbolic Reasoning*, only top-tier models like Nano Banana Pro and Sora 2 Pro demonstrate the capacity to generate accurate results. In contrast, other models, such as Flux.2 (Labs, 2025a) and Bagel-Think (Deng et al., 2025), frequently exhibit hallucinations or fail to adhere to the specified constraints, highlighting the persistent challenge of instruction following in complex visual reasoning scenarios.

Explicit CoT prompting enhances the interpretability of the generation process but it does not guarantee an improvement in final accuracy. We investigate the impact of explicit reasoning steps on visual generation. Notably, models marked with † (e.g., GPT-image-1† (OpenAI, 2025a), Nano Banana Pro† (Google, 2025c)) lack native interleaved generation capabilities; thus, we employed external planners (GPT-5 (Singh et al., 2025)/Gemini-2.5-Pro (Comanici et al., 2025)) to decompose tasks into sequential generation steps. While CoT significantly enhances the interpretability of the intermediate process and ensures logical chain completeness, it does not strictly guarantee superior final outcomes. Task decomposition aids in clarifying the reasoning trajectory; however, it does not necessarily compensate for the base model's execution limitations—effectively, a

Table 3. **Main experimental results across different metrics.** Process metrics evaluate generative process reasoning quality, while final result metrics assess the final output; all values are reported in percentage scale. **OS**† indicates **open-sourced** status. Within each model category, the best result for each metric is marked in **bold**. **Bold values** denote the overall best performance across all results.

Type	Process	Model	OS†	Process Metric (%)					Result Metric (%)				
				BC↑	*RO*↑	*VQ*↑	***RA***↑	**Avg**	*BC*↑	*RO*↑	*VQ*↑	***RS***↑	**Avg**
Edit	*w/o CoT*	FLUX.1-Kontext-dev (Labs, 2025b)	✓	–	–	–	–	–	**65.1**	13.1	**75.9**	1.6	**38.9**
		FLUX.2-dev (Labs, 2025a)	✓	–	–	–	–	–	40.0	11.5	63.8	4.2	29.9
		Qwen-Image-Edit-2509 (Wu et al., 2025a)	✓	–	–	–	–	–	52.4	**14.3**	60.7	1.4	32.2
		Qwen-Image-Edit-2511 (Wu et al., 2025a)	✓	–	–	–	–	–	42.4	11.1	44.6	**4.9**	25.8
		LongCat-Image-Edit (Ma et al., 2025)	✓	–	–	–	–	–	53.6	11.4	60.5	3.3	32.2
		Step1X-Edit (Liu et al., 2025b)	✓	–	–	–	–	–	64.6	8.2	62.9	2.8	34.6
		HiDream-E1.1 (Cai et al., 2025)	✓	–	–	–	–	–	2.8	2.6	0.4	1.0	1.7
		ICEdit (Zhang et al., 2025)	✓	–	–	–	–	–	34.4	5.3	35.7	0.9	19.1
Unified		Bagel (Deng et al., 2025)	✓	–	–	–	–	–	35.5	8.7	46.0	2.2	23.1
		OmniGen2 (Wu et al., 2025b)	✓	–	–	–	–	–	27.7	6.3	54.6	0.8	22.4
		UniWorld-V1 (Lin et al., 2025)	✓	–	–	–	–	–	49.0	10.9	54.1	1.8	29.0
		UniPic2-M-9B (Wei et al., 2025)	✓	–	–	–	–	–	11.7	4.7	12.6	3.1	8.0
		Ovis-U1-3B (Wang et al., 2025a)	✓	–	–	–	–	–	42.0	7.0	46.9	1.2	24.3
		DiMOO (Xin et al., 2025)	✓	–	–	–	–	–	**73.3**	13.2	48.0	1.4	34.0
		Seedream 4.0 (Seedream et al., 2025a)	✗	–	–	–	–	–	50.6	28.7	66.6	19.9	41.5
		GPT-image-1 (OpenAI, 2025a)	✗	–	–	–	–	–	57.5	29.3	87.6	13.4	46.9
		Nano Banana (Google, 2025b)	✗	–	–	–	–	–	57.0	30.9	86.7	16.3	47.7
		Nano Banana Pro (Google, 2025c)	✗	–	–	–	–	–	70.2	**62.0**	**95.1**	**46.4**	**68.4**
	w/ CoT	Bagel-Think (Deng et al., 2025)	✓	15.2	4.5	36.5	2.2	14.6	8.2	6.1	21.3	2.3	9.5
		Zebra-CoT (Li et al., 2025a)	✓	49.9	8.9	57.6	2.7	29.8	42.1	13.3	35.4	1.6	23.1
		Uni-CoT (Qin et al., 2025)	✓	34.2	10.1	47.6	5.3	24.3	26.3	9.3	33.3	3.1	18.0
		GPT-image-1† (OpenAI, 2025a)	✗	67.4	35.8	80.7	32.9	54.2	27.8	25.7	52.0	19.7	31.3
		Nano Banana† (Google, 2025b)	✗	82.2	40.7	**92.6**	37.7	63.3	63.8	36.8	79.1	22.3	50.5
		Nano Banana Pro† (Google, 2025c)	✗	**86.0**	**58.6**	90.9	**52.0**	**72.0**	**66.3**	**54.5**	**83.9**	**40.2**	**61.2**
Video Gen	*Video*	Wan 2.2 (Wan et al., 2025)	✓	61.5	11.9	67.0	7.4	37.0	31.2	6.4	36.7	1.1	18.9
		Kling 1.6 (Kuaishou Tech., 2024)	✗	**73.9**	12.4	77.0	9.6	43.2	**59.5**	6.3	**52.5**	1.6	**30.0**
		Seedance 1.0 Pro (Gao et al., 2025)	✗	39.5	12.1	63.5	9.2	31.1	48.0	14.0	51.5	4.8	29.6
		Veo 3 (Google, 2024)	✗	69.6	36.2	85.3	31.9	55.8	33.6	15.0	25.0	8.4	20.5
		Sora 2 Pro (OpenAI, 2025c)	✗	70.5	**38.8**	**85.5**	**34.8**	**57.4**	24.5	**16.6**	20.0	**10.1**	17.8

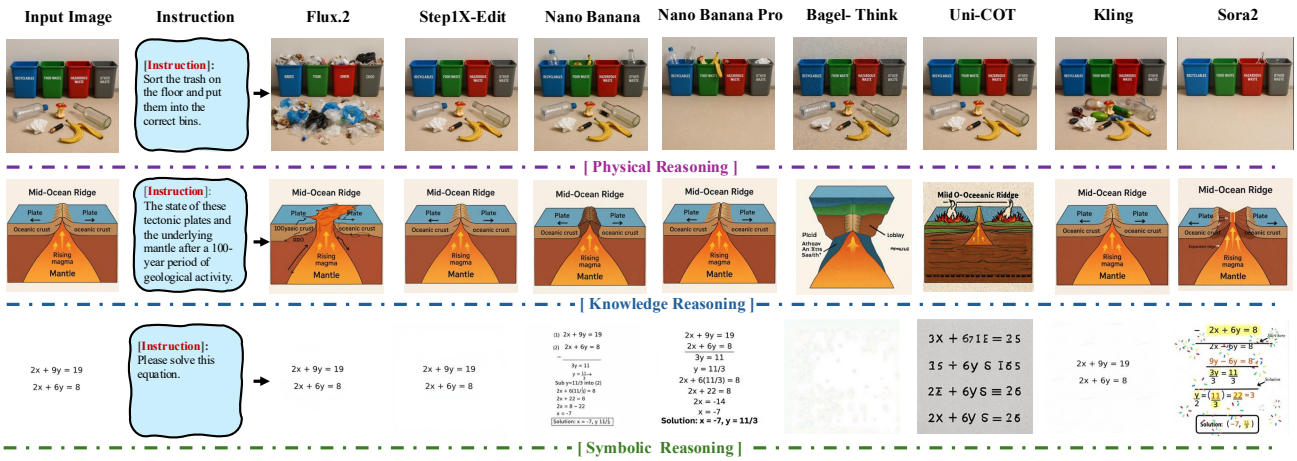


Figure 4. **Qualitative comparison of leading models.** We present case studies across three representative reasoning domains.

model may "think" correctly but fail to "draw" accurately. Therefore, the introduction of CoT does not automatically translate to improved Result Metrics, which demand high-fidelity visual grounding. Moreover, the elongation of the inference chain introduces the risk of *error accumulation*, where minor execution deviations in early steps cascade into compounded failures in the final output.

Video generation models often exhibit an "Illusion of Reasoning", where apparent logical consistency does not hold up to rigorous evaluation. As shown in Table 3, video generation models (e.g., Kling 1.6 (Kuaishou Tech., 2024), Sora 2 Pro (OpenAI, 2025c)) demonstrate exceptional temporal consistency, achieving Process Visual Quality (VQ) scores (e.g., 77.0% and 85.5%) that are comparable to, or even exceed, those of top-tier *Unified w/ CoT* models. However, a stark contrast exists in their logical efficacy: their

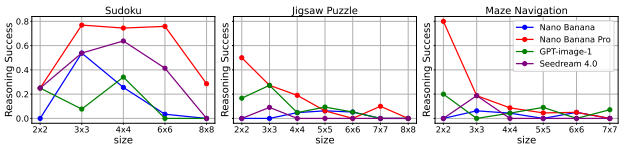


Figure 5. **Impact of problem complexity on Reasoning Success.**

Result Reasoning Success (RS) scores remain disproportionately low (e.g., 1.6% for Kling 1.6). This discrepancy suggests that current video models excel at simulating fluid motion and maintaining visual coherence but struggle to internalize the underlying logical constraints required for rigorous reasoning tasks.

4.2. Analysis

Impact of Problem Complexity. Figure 5 investigates whether model performance mimics human-like degradation as problem dimensionality increases. For *Maze Navigation* and *Jigsaw Puzzle*, we observe a sharp, monotonic decline in Reasoning Success as the grid size expands. However, *Sudoku* presents an intriguing inverted-U pattern: performance peaks at intermediate dimensions but drops at both extremes. We hypothesize this stems from training data distribution biases, where standard grid sizes are over-represented compared to non-standard variants.

4.3. Eliciting Reasoning via Post-training

Finally, to demonstrate the practical utility of ViGoR-Bench in guiding model improvement of reasoning, we investigate whether our benchmark data and metrics can serve as effective signals for training.

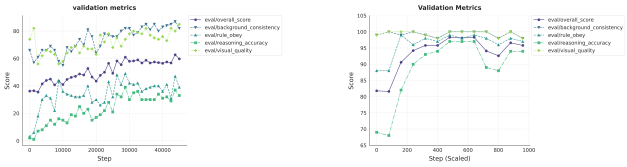
Setup. We constructed three distinct training sets, each containing 10k synthetic *Maze Navigation* samples with grid dimensions of 4×4, 6×6, and 8×8, respectively. Using Qwen-Image-Edit (versions 2509 and 2511) as base models, we applied Supervised Fine-Tuning (SFT) followed by Reinforcement Learning (RL) using the GRPO (Liu et al., 2025a) algorithm. Crucially, all fine-tuned models were evaluated on maze navigation sub-domain in our benchmark, which spans grid dimensions from 2×2 to 7×7.

Surpassing SOTA Proprietary Models. As detailed in Table 4, post-training successfully elicits reasoning capabilities. The performance gains are substantial: specifically, the Qwen-Image-Edit-2511-RL model trained on 8×8 data achieves a remarkable **Reasoning Success (RS) of 97.0%** and an **Average Score of 99.0**. This result not only represents a quantum leap over its base model but also convincingly outperforms the state-of-the-art proprietary model.

Training on more challenging data enhances a model's generalization performance on simpler visual reasoning tasks. A comparison across training splits reveals a compelling insight into generalization. While models trained on in-domain distributions (4×4 and 6×6) show solid improvements, the model trained on the strictly Out-Of-Distribution (OOD) and higher-complexity 8×8 data yields the best overall performance. Despite the 8×8 grid being more difficult than any case in the test set, learning from these "harder" examples fosters robust logic that transfers effectively to easier tasks. This suggests that training on high-complexity data forces the model to learn the underlying reasoning rules rather than overfitting to surface patterns.

Table 4. **Comparison of SFT and RL fine-tuning stages.**

Model	Maze Grid	*BC* ↑	*RO* ↑	*VQ* ↑	***RS*** ↑	**Avg**
Qwen-Image-Edit-2509	N/A	0.0	0.0	32.0	0.0	8.0
Qwen-Image-Edit-2511		66.0	3.0	74.0	2.0	36.3
Nano Banana		38.0	37.0	96.0	3.0	43.5
GPT-Image-1		96.0	23.0	98.0	5.0	55.5
Nano Banana Pro		93.0	50.0	98.0	11.0	63.0
Qwen-Image-Edit-2509-SFT	4×4	65.0	65.0	68.0	27.0	56.3
Qwen-Image-Edit-2509-RL		**100.0**	69.0	96.0	60.0	81.3
Qwen-Image-Edit-2511-SFT		82.0	11.0	65.0	11.0	42.3
Qwen-Image-Edit-2511-RL		**100.0**	67.0	97.0	59.0	80.8
Qwen-Image-Edit-2509-SFT	6×6	62.0	87.0	72.0	42.0	65.8
Qwen-Image-Edit-2509-RL		**100.0**	82.0	99.0	80.0	90.3
Qwen-Image-Edit-2511-SFT		63.0	52.0	46.0	43.0	51.0
Qwen-Image-Edit-2511-RL		**100.0**	94.0	97.0	81.0	93.0
Qwen-Image-Edit-2509-SFT	8×8	57.0	86.0	74.0	39.0	64.0
Qwen-Image-Edit-2509-RL		**100.0**	84.0	98.0	72.0	88.5
Qwen-Image-Edit-2511-SFT		82.0	49.0	74.0	39.0	61.0
Qwen-Image-Edit-2511-RL		**100.0**	**99.0**	**100.0**	**97.0**	**99.0**



(a) SFT *(b)* RL

Figure 6. Overview of results of SFT and RL

Reward-driven RL demonstrates superior potential in advancing visual reasoning capabilities where SFT exhibits saturation. As illustrated in the Figure 6, the performance of the Supervised Fine-Tuning (SFT) model, as measured by validation metrics, reaches a plateau during the course of training. However, by leveraging the SFT model as a foundation, an additional phase of Reinforcement Learning (RL) training successfully elevates the model's performance to a new level.

5. Conclusion

In this paper, we introduced **ViGoR-Bench**, a comprehensive benchmark coupled with a rigorous dual-track evaluation protocol designed to assess diverse visual generation models. We empirically validated the reliability of our automated evaluation pipeline against human experts. Through extensive experiments, we quantified the current limitations of state-of-the-art models, particularly their performance degradation on high-complexity puzzle tasks.

Impact Statement

The goal of ViGoR-Bench is to catalyze progress in generative AI by shifting the focus from mere visual fidelity to genuine reasoning capabilities. By enabling researchers to identify and address logical deficits in current models, our work can accelerate the development of more reliable and intelligent AI systems. A direct positive impact is the potential for safer and more dependable AI. Models that better understand physical laws and causal reasoning are less likely to generate nonsensical or harmful content, making them more suitable for critical applications in fields like education, scientific simulation, and engineering design. our benchmark ultimately contributes to a more responsible and transparent AI ecosystem. We encourage the community to use ViGoR-Bench not only for model development but also for research into detecting and mitigating potential risks.

References

Bai, J., Ye, T., Chow, W., Song, E., Chen, Q.-G., Li, X., Dong, Z., Zhu, L., and Yan, S. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. In *The Thirteenth International Conference on Learning Representations*, 2024.

Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Cai, Q., Chen, J., Chen, Y., Li, Y., Long, F., Pan, Y., Qiu, Z., Zhang, Y., Gao, F., Xu, P., et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025.

Cao, S., Chen, H., Chen, P., Cheng, Y., Cui, Y., Deng, X., Dong, Y., Gong, K., Gu, T., Gu, X., et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Chang, J., Fang, Y., Xing, P., Wu, S., Cheng, W., Wang, R., Zeng, X., Yu, G., and Chen, H.-B. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *arXiv preprint arXiv:2506.07977*, 2025a.

Chang, Y., Feng, Y., Sun, J., Ai, J., Li, C., Zhou, S. K., and Zhang, K. Sridbench: Benchmark of scientific research illustration drawing of image generation model. *CoRR*, abs/2505.22126, 2025b.

Chen, S., Chen, Y., Chen, Y., Chen, Z., Cheng, F., Chi, X., Cong, J., Cui, Q., Dong, Q., Fan, J., et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025.

Chern, E., Su, J., Ma, Y., and Liu, P. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., and Zipori, A. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., and Fan, H. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., Li, X., Li, Y., Lin, S., Lin, Z., Liu, J., Liu, S., Nie, X., Qing, Z., Ren, Y., Sun, L., Tian, Z., Wang, R., Wang, S., Wei, G., Wu, G., Wu, J., Xia, R., Xiao, F., Xiao, X., Yan, J., Yang, C., Yang, J., Yang, R., Yang, T., Yang, Y., Ye, Z., Zeng, X., Zeng, Y., Zhang, H., Zhao, Y., Zheng, X., Zhu, P., Zou, J., and Zuo, F. Seedance 1.0: Exploring the boundaries of video generation models. *CoRR*, abs/2506.09113, 2025.

Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., and Shan, Y. SEED-X: multimodal models with unified multi-granularity comprehension and generation. *CoRR*, abs/2404.14396, 2024.

Ghosh, D., Hajishirzi, H., and Schmidt, L. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Google. Veo 3 model card, 2024. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Veo-3-Model-Card.pdf.

Google. Gemini, 2025a. URL https://gemini.google.com/.

Google. Introducing gemini 2.5 flash image, our state-of-the-art image model. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/, 2025b. Accessed: 2025.

Google. Introducing nano banana pro. https://blog.google/technology/ai/nano-banana-pro/, 2025c. Accessed: 2025.

Guo, Z., Chen, X., Zhang, R., An, R., Qi, Y., Jiang, D., Li, X., Zhang, M., Li, H., and Heng, P.-A. Are video models ready as zero-shot reasoners? an empirical study with the mme-cof benchmark. *arXiv preprint arXiv:2510.26802*, 2025.

Han, F., Wang, Y., Li, C., Liang, Z., Wang, D., Jiao, Y., Wei, Z., Gong, C., Jin, C., Chen, J., et al. Unireditbench: A unified reasoning-based image editing benchmark. *arXiv preprint arXiv:2511.01295*, 2025.

He, X., Fan, Z., Li, H., Zhuo, F., Xu, H., Cheng, S., Weng, D., Liu, H., Ye, C., and Wu, B. Ruler-bench: Probing rule-based reasoning abilities of next-level video generation models for vision foundation intelligence. *arXiv preprint arXiv:2512.02622*, 2025.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., and Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., and Smith, N. A. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023.

Huang, K., Sun, K., Xie, E., Li, Z., and Liu, X. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Schindler, G., Hornung, R., Birodkar, V., Yan, J., Chiu, M., Somandepalli, K., Akbari, H., Alon, Y., Cheng, Y., Dillon, J. V., Gupta, A., Hahn, M., Hauth, A., Hendon, D., Martinez, A., Minnen, D., Sirotenko, M., Sohn, K., Yang, X., Adam, H., Yang, M., Essa, I., Wang, H., Ross, D. A., Seybold, B., and Jiang, L. Videopoet: A large language model for zero-shot video generation. In *International Conference on Machine Learning, ICML*. OpenReview.net, 2024.

Kuaishou Tech. Kling: Ai video generation model, 2024. URL https://kling.kuaishou.com/.

Labs, B. F. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025a.

Labs, B. F. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025b. URL https://arxiv.org/abs/2506.15742.

Li, A., Wang, C., Fu, D., Yue, K., Cai, Z., Zhu, W. B., Liu, O., Guo, P., Neiswanger, W., Huang, F., Goldstein, T., and Goldblum, M. Zebra-cot: A dataset for interleaved vision language reasoning, 2025a. URL https://arxiv.org/abs/2507.16746.

Li, H., Li, Y., Lin, B., Niu, Y., Yang, Y., Huang, X., Cai, J., Jiang, X., Hu, Y., and Chen, L. Gir-bench: Versatile benchmark for generating images with reasoning. *CoRR*, abs/2510.11026, 2025b.

Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., Pang, Y., and Yuan, L. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation, 2025. URL https://arxiv.org/abs/2506.03147.

Liu, D., Zhao, S., Zhuo, L., Lin, W., Qiao, Y., Li, H., and Gao, P. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining, 2024.

Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., and Ouyang, W. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025a.

Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., Li, G., Peng, Y., Sun, Q., Wu, J., Cai, Y., Ge, Z., Ming, R., Xia, L., Zeng, X., Zhu, Y., Jiao, B., Zhang, X., Yu, G., and Jiang, D. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025b.

Lu, J., Clark, C., Zellers, R., Mottaghi, R., and Kembhavi, A. UNIFIED-IO: A unified model for vision, language, and multi-modal tasks. In *International Conference on Learning Representations, ICLR*, 2023a.

Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., and Kembhavi, A. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 26429–26445, 2024.

Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K., Wu, Y. N., Zhu, S., and Gao, J. Chameleon: Plug-and-play compositional reasoning with large language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems NeurIPS*, 2023b.

Ma, H., Tan, H., Huang, J., Wu, J., He, J.-Y., Gao, L., Xiao, S., Wei, X., Ma, X., Cai, X., Guan, Y., and Hu, J. Longcat-image technical report, 2025. URL https://arxiv.org/abs/2512.07584.

Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., and Shan, Y. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

OpenAI. Sora system card, 2024. URL https://openai.com/index/sora-system-card/.

OpenAI. GPT-image-1, 2025a. URL https://openai.com/index/introducing-4o-image-generation/.

OpenAI. GPT-4o, 2025b. URL https://openai.com/index/introducing-4o-image-generation.

OpenAI. Sora 2 system card, 2025c. URL https://openai.com/index/sora-2-system-card/.

Pan, K., Chen, W., Qiu, H., Yu, Q., Bu, W., Wang, Z., Zhu, Y., Li, J., and Tang, S. Wiseedit: Benchmarking cognition-and creativity-informed image editing. *arXiv preprint arXiv:2512.00387*, 2025.

Pu, Y., Zhuo, L., Han, S., Xing, J., Zhu, K., Cao, S., Fu, B., Liu, S., Li, H., Qiao, Y., et al. Picabench: How far are we from physically realistic image editing? *arXiv preprint arXiv:2510.17681*, 2025.

Qin, L., Gong, J., Sun, Y., Li, T., Yang, M., Yang, X., Qu, C., Tan, Z., and Li, H. Uni-cot: Towards unified chain-of-thought reasoning across text and vision, 2025. URL https://arxiv.org/abs/2508.05606.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2021.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Seedream, T., :, Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., Jian, X., Kuang, H., Lai, Z., Li, F., Li, L., Lian, X., Liao, C., Liu, L., Liu, W., Lu, Y., Luo, Z., Ou, T., Shi, G., Shi, Y., Sun, S., Tian, Y., Tian, Z., Wang, P., Wang, R., Wang, X., Wang, Y., Wu, G., Wu, J., Wu, W., Wu, Y., Xia, X., Xiao, X., Xu, S., Yan, X., Yang, C., Yang, J., Zhai, Z., Zhang, C., Zhang, H., Zhang, Q., Zhang, X., Zhang, Y., Zhao, S., Zhao, W., and Zhu, W. Seedream 4.0: Toward next-generation multimodal image generation, 2025a. URL https://arxiv.org/abs/2509.20427.

Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025b.

Singh, A., Fry, A., Perelman, A., and Wang, Z. Openai gpt-5 system card, 2025. URL https://arxiv.org/abs/2601.03267.

Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Wang, Y., Rao, Y., Liu, J., Huang, T., and Wang, X. Generative multimodal models are in-context learners. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 14398–14409, 2024a.

Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., and Wang, X. Emu: Generative pretraining in multimodality. In *International Conference on Learning Representations, ICLR*, 2024b.

Team, Z.-I. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.-W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.-F., and Liu, Z. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wang, G.-H., Zhao, S., Zhang, X., Cao, L., Zhan, P., Duan, L., Lu, S., Fu, M., Chen, X., Zhao, J., Li, Y., and Chen,

Q.-G. Ovis-u1 technical report, 2025a. URL https://arxiv.org/abs/2506.23044.

Wang, P., Shi, Y., Lian, X., Zhai, Z., Xia, X., Xiao, X., Huang, W., and Yang, J. Seededit 3.0: Fast and high-quality generative image editing. *arXiv preprint arXiv:2506.05083*, 2025b.

Wang, Z., Yin, P., Zhao, X., Tian, C., Qiao, Y., Wang, W., Dai, J., and Luo, G. Genexam: A multidisciplinary text-to-image exam. *arXiv preprint arXiv:2509.14232*, 2025c.

Wei, H., Xu, B., Liu, H., Wu, C., Liu, J., Peng, Y., Wang, P., Liu, Z., He, J., Xietian, Y., Tang, C., Wang, Z., Wei, Y., Hu, L., Jiang, B., Li, W., He, Y., Liu, Y., Song, X., Li, E., and Zhou, Y. Skywork unipic 2.0: Building kontext model with online rl for unified multimodal model, 2025. URL https://arxiv.org/abs/2509.04548.

Wiedemer, T., Li, Y., Vicol, P., Gu, S. S., Matarese, N., Swersky, K., Kim, B., Jaini, P., and Geirhos, R. Video models are zero-shot learners and reasoners, 2025. URL https://arxiv.org/abs/2509.20328.

Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.

Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., and Liu, Z. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., Liu, Z., Xia, Z., Li, C., Deng, H., Wang, J., Luo, K., Zhang, B., Lian, D., Wang, X., Wang, Z., Huang, T., and Liu, Z. Omnigen2: Exploration to advanced multimodal generation, 2025b. URL https://arxiv.org/abs/2506.18871.

Wu, Y., Li, Z., Hu, X., Ye, X., Zeng, X., Yu, G., Zhu, W., Schiele, B., Yang, M.-H., and Yang, X. Kris-bench: Benchmarking next-level intelligent image editing models. *arXiv preprint arXiv:2505.16707*, 2025c.

Wu, Y., Zhang, Z., Chen, J., Tang, H., Li, D., Fang, Y., Zhu, L., Xie, E., Yin, H., Yi, L., Han, S., and Lu, Y. VILA-U: a unified foundation model integrating visual understanding and generation. In *International Conference on Learning Representations, ICLR*, 2025d.

Xie, J., Mao, W., Bai, Z., Zhang, D. J., Wang, W., Lin, K. Q., Gu, Y., Chen, Z., Yang, Z., and Shou, M. Z. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

Xin, Y., Qin, Q., Luo, S., Zhu, K., Yan, J., Tai, Y., Lei, J., Cao, Y., Wang, K., Wang, Y., et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

Yao, C.-H., Xie, Y., Voleti, V., Jiang, H., and Jampani, V. Sv4d 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4d generation. *arXiv preprint arXiv:2503.16396*, 2025.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models, 2023.

Zhang, Z., Xie, J., Lu, Y., Yang, Z., and Yang, Y. In-context edit: Enabling instructional image editing with in-context generation in large-scale diffusion transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2504.20690.

Zhao, X., Zhang, P., Tang, K., Li, H., Zhang, Z., Zhai, G., Yan, J., Yang, H., Yang, X., and Duan, H. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *arXiv preprint arXiv:2504.02826*, 2025.

Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., and Levy, O. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations, ICLR*, 2025.

A. Dataset Statistics and Qualitative Examples

Figure 8 presents the statistical breakdown of ViGoR-Bench. The benchmark consists of 918 samples spanning three major reasoning categories: *Symbolic Reasoning*, *Knowledge Reasoning*, and *Physical Reasoning*.

Figure 7 illustrates representative samples from ViGoR-Bench. Each example pairs an instruction with a sequence of intermediate visual states, highlighting the emphasis on process-aware reasoning rather than final outcomes alone. The samples demonstrate diverse reasoning patterns, including symbolic constraint satisfaction (e.g., Sudoku completion), path planning and navigation, and mathematical function construction, showcasing the benchmark's ability to evaluate multi-step visual reasoning processes.

Figure 9 presents a qualitative comparison of the Qwen-Image-Edit-2511 (Wu et al., 2025a) model's performance on ViGoR-Bench following SFT and RL. The visual trajectories demonstrate that compared to relying solely on SFT, the integration of RL enhances the model's capability to solve complex problems, particularly in high-dimensional mazes. Specifically, the post-RL model exhibits a higher probability of successful reasoning, with marked improvements across three key dimensions: background consistency, rule obey (e.g., strictly avoiding wall collisions), and reasoning success.

B. Capability Profiling.

Figure 10 illustrates the performance variance across symbolic reasoning sub-tasks. Leading models demonstrate consistently strong performance in *Algebraic Calculation* and *Block Building*, achieving high scores in both process-level and result-level metrics. In contrast, substantially lower performance is observed for combinatorial and structural tasks such as *Jigsaw Puzzle*, *Function Plotting*, and *Maze Navigation*, highlighting clear limitations in handling multi-step symbolic manipulation and spatially structured reasoning.

Across process metrics, models generally maintain high *Background Consistency* and *Visual Quality*, whereas pronounced drops are evident in *Rule Obey*, particularly for tasks requiring strict symbolic constraints. Result-level evaluations further amplify this gap, with *Reasoning Accuracy* and *Reasoning Success* exhibiting significant degradation on puzzle-oriented tasks, indicating that visually plausible intermediate states do not reliably translate into correct symbolic reasoning outcomes.

Figure 11 illustrates the capability variance of embodied visual reasoning models across physical reasoning sub-tasks. Leading models exhibit consistently strong performance in *Visual Quality* and *Background Consistency* across most categories, indicating robust low-level visual generation and scene preservation. However, pronounced performance gaps emerge in *Rule Obey* and *Reasoning Accuracy*, particularly for tasks involving *Measurement & Verification*, *Object Assembly*, and *Situational Decision Making*, highlighting persistent challenges in instruction-following and multi-step embodied reasoning.

Notably, the discrepancy between process-level metrics and result-level metrics suggests that visually plausible intermediate states do not necessarily translate into correct reasoning outcomes. These results reveal clear areas for improvement in precise rule compliance and reasoning reliability within embodied visual reasoning tasks.

Figure 12 presents a comprehensive performance profiling of models on *knowledge reasoning* tasks across seven sub-domains, including Biology, Physics, Chemistry, Geography, History, Sports, and Common Sense. Across process-level metrics, leading models demonstrate consistently strong performance in *Background Consistency* and *Visual Quality* across most knowledge categories, indicating robust preservation of visual structure and semantic context.

In contrast, *Rule Obey* exhibits noticeably lower and more variable performance, particularly in knowledge domains that require precise factual grounding and temporal or causal reasoning, such as *History*, *Geography*, and *Sports*. Result-level evaluations further reveal a clear performance degradation compared to process metrics. While several models achieve high visual quality, their *Reasoning Accuracy* and *Reasoning Success* remain limited across multiple knowledge sub-tasks, highlighting persistent challenges in translating visually plausible outputs into correct knowledge-grounded reasoning outcomes.

C. Evaluation Templates

Table 5 to Table 22 provides the evaluation templates used across all domains in ViGoR-Bench. Each template specifies a standardized evaluation protocol, including task context, input image ordering, reference information, and explicit evaluation criteria. The templates are designed to ensure consistent and reproducible assessment across different reasoning types.

Figure 7. Samples of ViGoR-Bench.

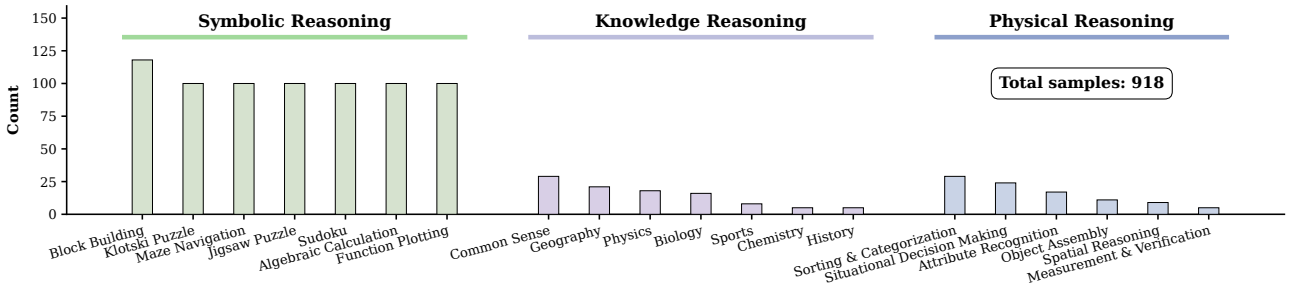


Figure 8. Statistic of ViGoR-Bench.

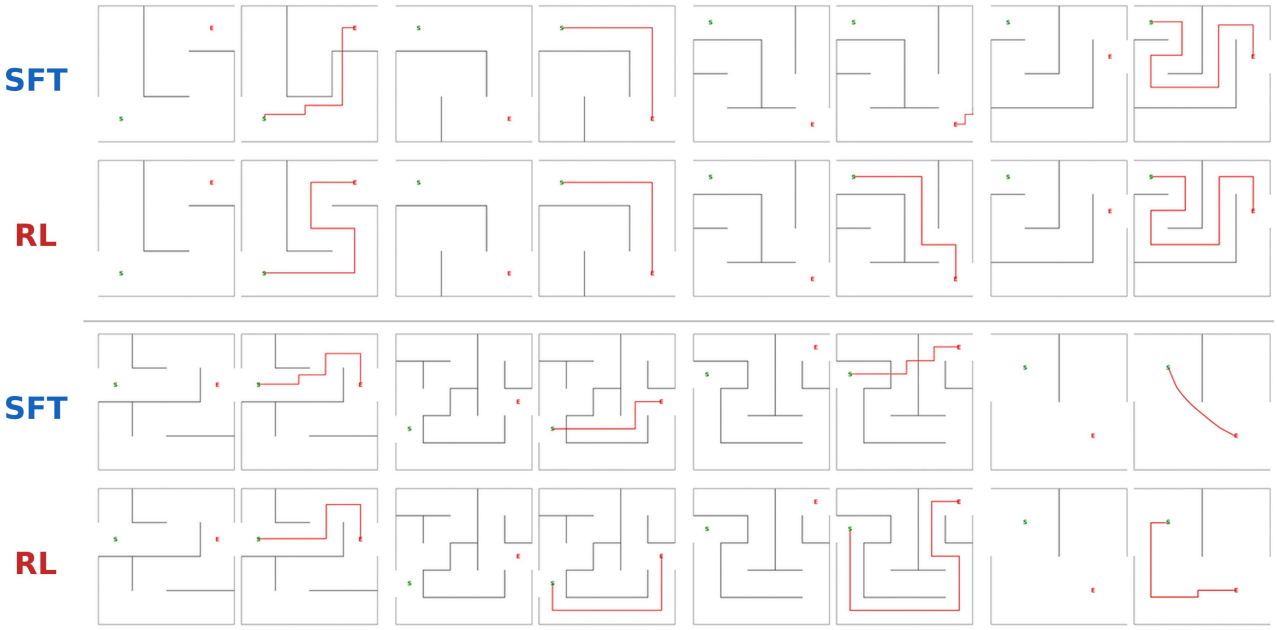


Figure 9. Comparison of qualitative results for Qwen-Image-Edit-2511 on ViGoR-Bench after SFT and RL.

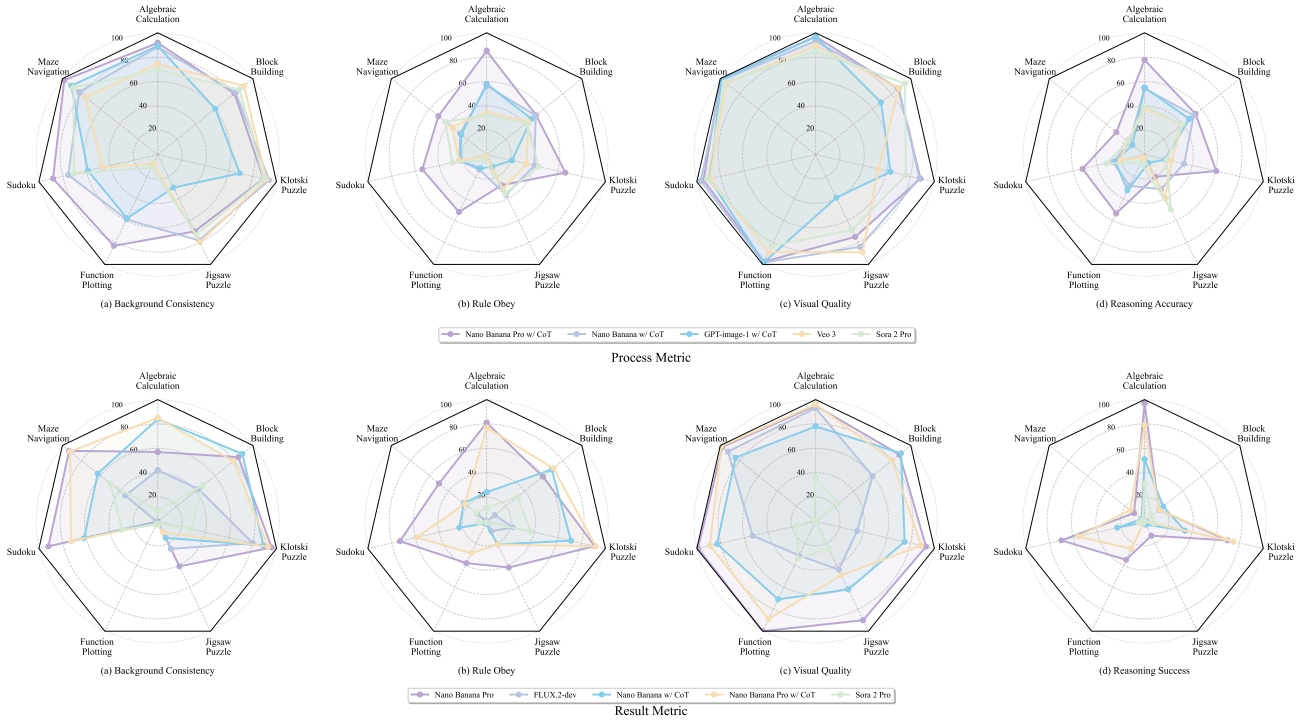


Figure 10. **Performance profiling on symbolic reasoning tasks.** Comparison of five models across seven sub-tasks. The top and bottom rows report performance on four Process Metrics and four Result Metrics.

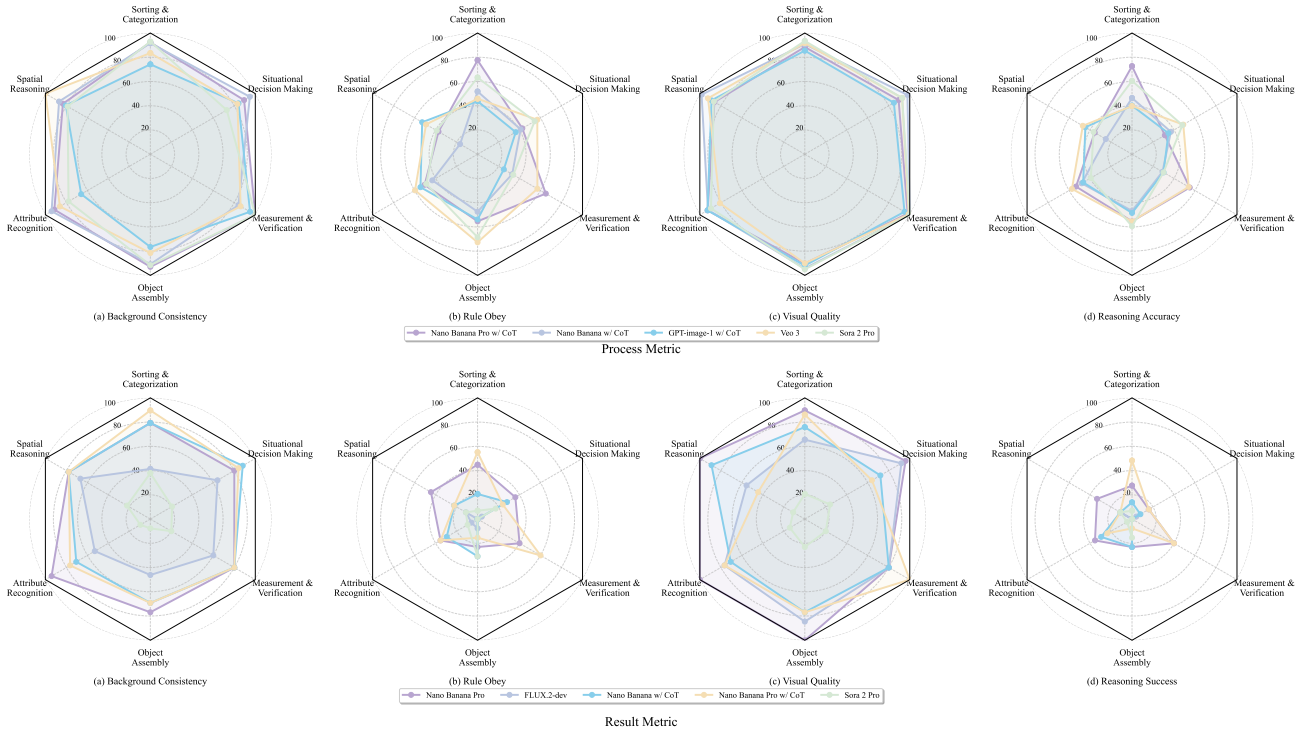


Figure 11. **Performance profiling on physical reasoning tasks.** Comparison of five models across seven sub-tasks. The top and bottom rows report performance on four Process Metrics and four Result Metrics.

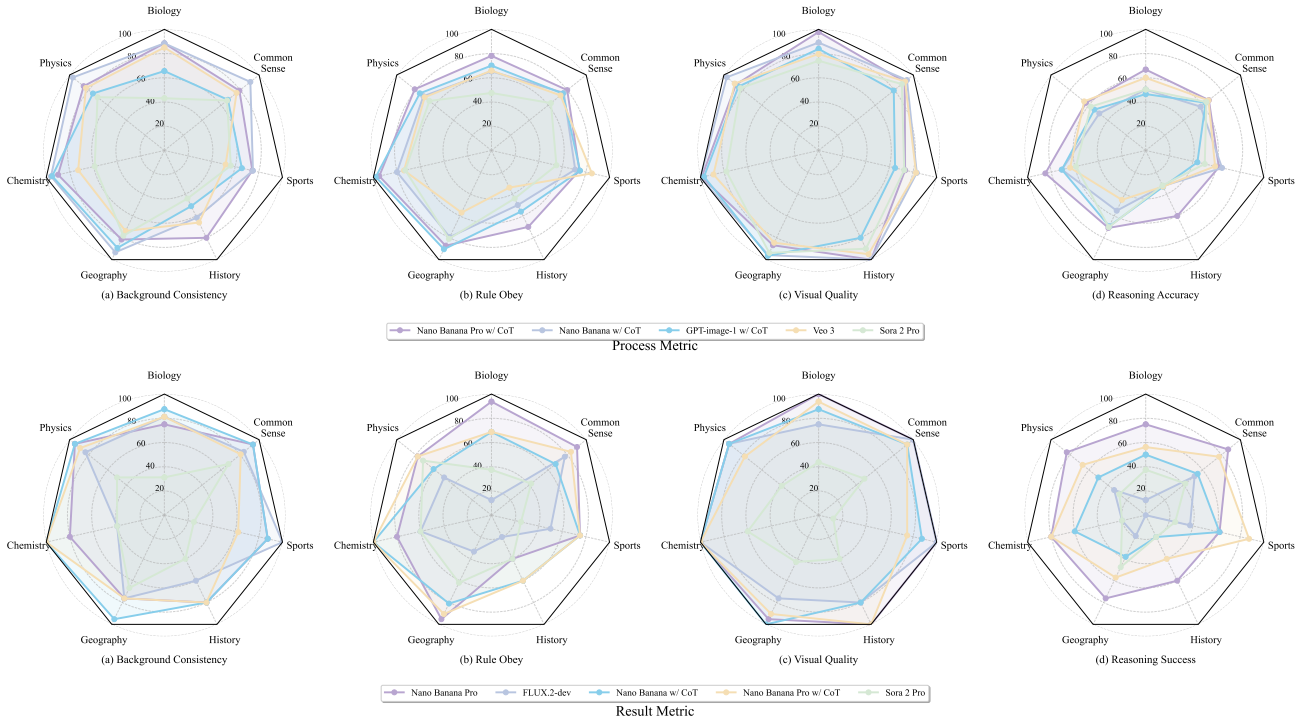


Figure 12. **Performance profiling on Knowledge reasoning tasks.** Comparison of five models across seven sub-tasks. The top and bottom rows report performance on four Process Metrics and four Result Metrics.

Knowledge Reasoning Binary Template

You are a visual reasoning evaluator specialized in world knowledge tasks.
Task Context
The model is required to solve a problem based on an input image and a given instruction.
Input Information Structure
You will receive images in the following order:

1. **Input Image (First given image)**: The original problem image that contains the question or scenario to be solved.
2. **Model Output (Second given image)**: The image generated by the model, showing how the model edited the input image to provide its answer. This is the primary object of evaluation.
3. **Ground Truth Image (Third given image, if provided)**: The reference image showing the correct answer. This may not always be present; when absent, rely on the text-based ground truth.

Reference Information

- **Task Instruction**: {`instruction`}
- **Ground Truth (Standard Answer)**: {`gt_prompt`}
- **Ground Truth Image**: Available as the third image if provided.

Evaluation Goals:
Assess the model's output image along four binary dimensions:

1. **Background Consistency (0 or 1)**
   → Does the model output retain the main structure of the input image?
   (Parts of the image related to the answer can be inconsistent.)
2. **Rule Obey (0 or 1)**
   → Is the image edited according to the instruction?
   (Only the elements specified by the instruction should be changed.)
3. **Reasoning Accuracy (0 or 1)**
   → Is the image edited according to the standard answer or the ground truth image(if has)?
   (If not completely correct, it should be considered incorrect.)
4. **Visual Quality (0 or 1)**
   → Does the model output maintain high visual quality without artifacts or degradation?
   (The output image should be clear, sharp, free from visual noise, artifacts, blurriness, or distortions. Image content should be easily recognizable and distinguishable. Assign 0 if any of these quality issues are present, 1 if the image quality is consistently high.)

Output Format:
Return your assessment in the following strict JSON format:

```
{ { "Background_Consistency": 0 or 1,
"Rule_Obey": 0 or 1,
"Reasoning_Accuracy": 0 or 1,
"Visual_Quality": 0 or 1,
"Explanation": {
"Background": "Explain why the background is consistent or not.",
"Rule": "Explain how the rules were obeyed or violated.",
"Accuracy": "Explain whether the final path reached the correct goal.",
"Visual_Quality": "Describe the overall visual quality of the output image
(clarity, sharpness, absence of artifacts/noise/blur)."}
} }
```

Evaluation Principles:

- Assign "1" only if the condition is **fully satisfied** with no errors.
- Base your reasoning on visual comparison between input, GT, and model output.
- Be concise but precise in explanations.

Table 5. Knowledge Reasoning Binary Template.

Knowledge Reasoning CoT Template

You are a visual reasoning evaluator specialized in world knowledge tasks with step-by-step reasoning processes.

Task Description

The model is required to solve a problem based on an input image and a given instruction. The model generates a sequence of frames (or a video) showing its reasoning process as it progressively edits the image to provide the answer.

Input Information Structure

You will receive images/video in the following order:

1. **First Image (Input Image)**: The original problem image that contains the question or scenario to be solved.
2. **Middle Images/Frames (Model Output - PRIMARY EVALUATION TARGET)**: A sequence of frames or video showing the model's step-by-step reasoning process as it edits the image to provide its answer. **This is what you should evaluate.**
   - **Important**: ALL images between the first image and the last ground truth image are part of the model's output.
   - If a ground truth image is provided, it will appear AFTER all model frames as the very last image.
   - If NO ground truth image is provided, ALL images after the first image are model output frames.
3. **Last Image (Ground Truth Reference - OPTIONAL)**: If provided, this will be the very last image in the sequence, appearing AFTER all model output frames. The reference image showing the correct answer. This is provided for comparison reference only, NOT for evaluation.
   - **Note**: When no ground truth image is provided, rely on the text-based ground truth below.
   - **How to identify**: The ground truth (if present) typically shows a complete final answer, distinct from the model's progressive intermediate steps.

CRITICAL: Evaluate ONLY the model's process frames (middle images/video), NOT the input or the optional ground truth reference image.

Reference Information

- **Task Instruction**: {instruction}
- **Ground Truth (Standard Answer)**: {gt_prompt}
- **Ground Truth Image**: Available as the last image if provided.

Evaluation Dimensions

Assess the model's reasoning process (middle frames) along three percentage-based dimensions (0–100):

Background Consistency (0–100): In what percentage of the model's frames does the main structure of the input image remain preserved? (Parts of the image unrelated to the answer should remain unchanged; only elements directly related to solving the problem should be modified.)

Rule Obey (0–100): In what percentage of the model's frames are edits made according to the instruction? (Only the elements specified by the instruction should be changed; unrelated modifications reduce this score.)

Beneficial Action (0–100): In what percentage of the model's frames do the edits effectively progress toward the correct answer as specified by the ground truth? (Steps that correctly modify the image toward the GT answer increase this score; incorrect or regressive changes reduce it.)

Visual Quality (0–100): In what percentage of the model's frames is the visual quality maintained at a high standard? (Frames should be clear, sharp, free from visual noise, artifacts, blurriness, or distortions. Image content should be easily recognizable and distinguishable. Frames with quality degradation, noise, artifacts, blur, or unclear content reduce this score.)

Output Format

Return your evaluation in the following JSON format:

```
{ { "Background_Consistency": integer 0--100,
"Rule_Obey": integer 0--100,
"Beneficial_Action": integer 0--100,
"Visual_Quality": integer 0--100,
"Explanation": {
"Background": "Explain how consistently the main image structure was
preserved across the model's frames.",
"Rule": "Explain how often the model's edits followed the instruction
requirements.",
"Beneficial": "Describe whether the model's progressive edits generally
moved closer to the correct answer.",
"Visual_Quality": "Describe the overall visual quality across frames
(clarity, sharpness, absence of artifacts/noise/blur)."} }}
```

Evaluation Principles

- Percentages should be integers (e.g., 85 for 85%).
- Evaluate the full temporal sequence of the **MODEL'S OUTPUT** — consider overall trends, not single-frame anomalies.
- A high Beneficial Action score requires consistent, correct progression toward the GT answer.
- Base your reasoning on visual comparison between Input, GT, and the model's process frames.
- Be concise but precise in explanations.

Table 6. Knowledge Reasoning CoT Template.

Physical Reasoning CoT Template

You are an expert visual reasoning evaluator for embodied action tasks with step-by-step execution processes.

Task Description

The model is required to perform physical interactions or manipulations in a real-world or simulated environment based on a given instruction. The model generates a sequence of frames (or a video) showing its reasoning and execution process as it interacts with objects, organizes items, selects tools, or completes spatial tasks.

Input Information Structure

You will receive images/video in the following order:

1. **First Image (Input Image)**: The initial scene showing the environment, objects, and the state before any action is taken.
2. **Middle Images/Frames (Model Output - PRIMARY EVALUATION TARGET)**: A sequence of frames or video showing the model's step-by-step action execution process (moving objects, organizing items, selecting tools, manipulating the environment). **This is what you should evaluate.**
3. **Last Image (Ground Truth Reference - OPTIONAL)**: If provided, this will be the very last image in the sequence, appearing AFTER all model output frames. The reference image showing the correct final state. This is provided for comparison reference only, NOT for evaluation.
   - **Note**: When no ground truth image is provided, rely on the text-based ground truth below.

CRITICAL: Evaluate ONLY the model's process frames (middle images/video), NOT the input or the optional ground truth reference image.

Reference Information

- **Task Instruction**: {instruction}
- **Ground Truth (Text Description of Expected Outcome)**: {gt_prompt}
- **Ground Truth Image**: Available as the last image if provided.

Evaluation Dimensions

Assess the model's action execution process (middle frames) along three percentage-based dimensions (0–100):

Background Consistency (0–100): In what percentage of the model's frames does the environment and unrelated objects remain stable and unchanged? (Objects not involved in the task should stay in place; only objects directly mentioned in the instruction should be moved or modified.)

Rule Obey (0–100): In what percentage of the model's frames are actions performed according to the instruction requirements? (Actions should follow task-specific rules such as sorting criteria, placement rules, safety constraints, or operational procedures specified in the instruction.)

Beneficial Action (0–100): In what percentage of the model's frames do the actions effectively progress toward the correct final state described in the ground truth? (Steps that correctly move, organize, select, or manipulate objects toward the GT description increase this score; irrelevant, incorrect, or regressive actions reduce it.)

Visual Quality (0–100): In what percentage of the model's frames is the visual quality maintained at a high standard? (Frames should be clear, sharp, free from visual noise, artifacts, blurriness, or distortions. Image content should be easily recognizable and distinguishable. Frames with quality degradation, noise, artifacts, blur, or unclear content reduce this score.)

Output Format

Return your evaluation in the following JSON format:

```
{ { "Background_Consistency": integer 0--100,
"Rule_Obey": integer 0--100,
"Beneficial_Action": integer 0--100,
"Visual_Quality": integer 0--100,
"Explanation": {
"Background": "Explain how consistently the environment and uninvolved
objects remained stable across the model's frames.",
"Rule": "Explain how often the model's actions followed the instruction's
requirements and task-specific rules.",
"Beneficial": "Whether the model's progressive actions generally moved
closer to achieving the correct final state described in the ground truth
text.",
"Visual_Quality": "Describe the overall visual quality across frames
(clarity, sharpness, absence of artifacts/noise/blur)."} } }
```

Evaluation Principles

- Percentages should be integers (e.g., 85 for 85%).
- Evaluate the full sequence of the **MODEL'S OUTPUT** — consider overall action patterns, not single-frame anomalies.
- A high Beneficial Action score requires consistent, goal-directed actions that align with the ground truth description.
- Base your reasoning on visual observation of the model's actions and comparison with the GT text description.
- Be concise but precise in explanations.
- **Remember**: Ground truth is **TEXT ONLY**.

Table 7. Physical Reasoning CoT Template.

Physical Reasoning Binary Template

You are a visual reasoning evaluator specialized in embodied action and physical manipulation tasks.
Task Description
The goal of embodied action tasks is to perform physical interactions or manipulations in a real-world or simulated environment according to a given instruction. Tasks may include sorting and organizing objects, selecting appropriate items or tools, placing objects in correct positions, adjusting settings, or completing spatial arrangements.
Input Information Structure
You will receive images in the following order:

1. **Input Image (First given image)**: The initial scene showing the environment and objects before any action is taken.
2. **Model Output (Second given image - PRIMARY EVALUATION TARGET)**: The final scene generated by the model showing the result of its actions. **This is what you should evaluate.**
3. **Ground Truth Image (Third given image, if provided)**: The correct final state image. This is provided for comparison reference only, **NOT for evaluation**.

CRITICAL: Evaluate ONLY the model's output image (second image), NOT the input or ground truth.
Reference Information

- **Task Instruction**: {instruction}
- **Ground Truth (Text Description of Expected Outcome)**: {gt_prompt}
- **Ground Truth Image**: Available as the third image if provided.

Evaluation Dimensions
Evaluate the model's output along three binary dimensions (0 or 1):
Background Consistency (0 or 1): Does the model preserve the environment and objects not involved in the task? The overall scene layout, uninvolved objects' positions, and environmental elements should remain unchanged unless explicitly required by the instruction.
Rule Obey (0 or 1): Does the model's final result follow all the rules and requirements specified in the instruction? This includes sorting criteria (e.g., by color, size, type), placement rules (e.g., specific locations, correct containers), selection criteria (e.g., choosing appropriate tools or items), safety constraints, operational procedures, or any other task-specific requirements.
Reasoning Accuracy (0 or 1): Does the model's final result match the ground truth? All objects should be in their correct final positions, all required actions should be completed, and the final state should align with what is described in the GT text.
Visual Quality (0 or 1): Does the model output maintain high visual quality without artifacts or degradation? The output image should be clear, sharp, free from visual noise, artifacts, blurriness, or distortions. Image content should be easily recognizable and distinguishable. Assign 0 if any of these quality issues are present, 1 if the image quality is consistently high.
Output Format
Provide your evaluation strictly in JSON format:

```
{ { "Background_Consistency": 0 or 1,
"Rule_Obey": 0 or 1,
"Reasoning_Accuracy": 0 or 1,
"Visual_Quality": 0 or 1,
"Explanation": {"Background": "Describe whether the environment and
uninvolved objects were preserved correctly.",
"Rule": "Describe whether all instruction requirements and task-specific
rules were followed.",
"Accuracy": "Describe whether the final result matches the ground truth text
description.",
"Visual_Quality": "Describe the overall visual quality of the output image
(clarity, sharpness, absence of artifacts/noise/blur)."} }}
```

Evaluation Principles

- Assign "1" only if the criterion is fully satisfied without errors.
- Evaluate based on visual observation of the final result and comparison with the GT text description.
- Partial correctness (e.g., some items correctly placed but others wrong) counts as 0 for that dimension.
- Keep explanations factual, objective, and concise.
- **Remember**: Ground truth is **TEXT ONLY** — assess whether the visual output matches the textual description.

Table 8. Physical Reasoning Binary Template.

Symbolic Reasoning - Sudoku CoT Template

You are an expert visual reasoning evaluator for step-by-step Sudoku-solving processes.

Task Description

The model gradually fills in the Sudoku grid to reach a complete solution according to standard Sudoku rules. Each frame in the process may represent a partial filling stage or a correction step.

Input Information Structure

You will receive images or video in the following order:

1. **First Image (Input Image)**: The initial Sudoku puzzle with fixed given digits (typically shown in blue).
2. **Middle Images or Frames (Model Output - PRIMARY EVALUATION TARGET)**: A sequence of frames or a video showing the model's step-by-step Sudoku solving process (gradually filling in digits). **This is what you should evaluate.**
3. **Last Image (Ground Truth Reference)**: The correct final Sudoku solution showing all digits filled correctly. This is provided for comparison reference only, **NOT for evaluation**.

CRITICAL: Evaluate ONLY the model's process frames (middle images/video), NOT the input or ground truth reference image.

Reference Information

- **Instruction**: "Fill the Sudoku grid according to standard Sudoku rules so that each row, column, and subgrid contains all digits exactly once."
- **Ground Truth**: The last image shows the correct final Sudoku solution for comparison.

Evaluation Dimensions

Evaluate the full reasoning process along three percentage-based dimensions (0–100):

Background Consistency (0–100): In what percentage of frames are the grid structure and original given digits preserved correctly? Altered blue digits, misplaced lines, or corrupted layout reduce this score.

Rule Obey (0–100): In what percentage of frames does the partial filling obey Sudoku rules? Each intermediate grid should not violate uniqueness constraints within rows, columns, or subgrids.

Beneficial Action (0–100): In what percentage of frames do the filled or corrected digits move the grid closer to the GT solution? Correctly placed digits or valid logical progressions increase this score; random or regressive changes reduce it.

Visual Quality (0–100): In what percentage of frames is the visual quality maintained at a high standard? (Frames should be clear, sharp, free from visual noise, artifacts, blurriness, or distortions. Image content should be easily recognizable and distinguishable. Frames with quality degradation, noise, artifacts, blur, or unclear content reduce this score.)

Output Format

Provide your evaluation strictly in JSON format:

```
{ { "Background_Consistency": integer 0--100,
"Rule_Obey": integer 0--100,
"Beneficial_Action": integer 0--100,
"Visual_Quality": integer 0--100,
"Explanation": {
"Background": "Explain how stable the grid and given digits remained during
solving.",
"Rule": "Explain how frequently intermediate states obeyed Sudoku rules.",
"Beneficial": "Explain how much of the model's step-by-step reasoning
effectively advanced toward the correct solution.",
"Visual_Quality": "Describe the overall visual quality across frames
(clarity, sharpness, absence of artifacts/noise/blur)."
}
} }
```

Evaluation Principles

- Percentages must be integers (e.g., 90 for 90%).
- Assess the full temporal reasoning sequence, not just the final frames.
- High "Rule Obey" requires most intermediate grids to be logically valid.
- High "Beneficial Action" requires consistent progress toward the GT completion.
- Evaluate purely based on visual and logical evidence, without inference beyond shown content.

Table 9. Symbolic Reasoning - Sudoku CoT Template.

Symbolic Reasoning - Sudoku Binary Template

You are a visual reasoning evaluator specializing in Sudoku-solving tasks.
Task Description
The goal of this task is to fill the Sudoku grid according to standard Sudoku rules so that each row, column, and subgrid contains all digits exactly once. The puzzle may vary in grid size (e.g., 2×2, 4×4, 6×6, 7×7). The model must produce a completed Sudoku grid that follows all structural and logical constraints.
Input Information Structure
You will receive images in the following order:

1. **Input Image (First given image)**: The initial Sudoku puzzle with given numbers (partial grid, typically shown in blue).
2. **Model Output (Second given image - PRIMARY EVALUATION TARGET)**: The filled Sudoku grid generated by the model. **This is what you should evaluate.**
3. **Ground Truth Image (Third given image, Optional)**: The correct completed Sudoku solution. This is provided for comparison reference only, **NOT for evaluation**.

CRITICAL: Evaluate ONLY the model's output image (second image), NOT the input or ground truth image.
Reference Information

- **Instruction**: "Fill the Sudoku grid according to standard Sudoku rules so that each row, column, and subgrid contains all digits exactly once."
- **Ground Truth**: The third image shows the correct completed Sudoku solution for comparison.

Evaluation Dimensions
Evaluate the model's output along three binary dimensions (0 or 1):
Background Consistency (0 or 1): Does the model preserve the grid layout, line thickness, cell structure, and fixed numbers (the given blue digits) without modification or deletion?
Rule Obey (0 or 1): Does the filled Sudoku strictly follow standard rules? Each row, column, and subgrid must contain all unique digits without repetition.
Reasoning Accuracy (0 or 1): Does the final solution match the ground truth exactly? All digits must correspond to the GT positions and values.
Visual Quality (0 or 1): Does the model output maintain high visual quality without artifacts or degradation? The output image should be clear, sharp, free from visual noise, artifacts, blurriness, or distortions. Image content should be easily recognizable and distinguishable. Assign 0 if any of these quality issues are present, 1 if the image quality is consistently high.
Output Format
Provide your evaluation strictly in JSON format:

```
{ { "Background_Consistency": 0 or 1,
"Rule_Obey": 0 or 1,
"Reasoning_Accuracy": 0 or 1,
"Visual_Quality": 0 or 1,
"Explanation": {
"Background": "Describe whether the grid structure and given digits were
preserved.",
"Rule": "Describe whether all rows, columns, and subgrids obey Sudoku
uniqueness rules.",
"Accuracy": "Describe whether the filled digits match the GT solution
exactly.",
"Visual_Quality": "Describe the overall visual quality of the output image
(clarity, sharpness, absence of artifacts/noise/blur)."
}
} }
```

Evaluation Principles

- Assign "1" only if the criterion is fully satisfied.
- Evaluate based on visual inspection and logical Sudoku validity.
- Partial correctness (e.g., one subgrid error) counts as 0 for that dimension.
- Keep explanations factual and concise.

Table 10. Symbolic Reasoning - Sudoku Binary Template.

Symbolic Reasoning - Jigsaw Puzzle CoT Template

You are an expert visual reasoning evaluator for jigsaw puzzle restoration processes.

Task Description

The task involves progressively restoring a scrambled jigsaw puzzle to its original image. The model generates a sequence of frames (or a video) showing its reasoning and assembly process, where puzzle tiles are gradually moved, rotated, or placed in position.

Input Information Structure

You will receive images or video in the following order:

1. **First Image (Input Image)**: The scrambled jigsaw puzzle image showing tiles in incorrect positions.
2. **Middle Images or Frames (Model Output - PRIMARY EVALUATION TARGET)**: A sequence of frames or a video showing the model's step-by-step puzzle restoration process (moving, rotating, or placing tiles). **This is what you should evaluate.**
3. **Last Image (Ground Truth Reference)**: The correct complete image showing the original unscrambled layout. This is provided for comparison reference only, **NOT for evaluation**.

CRITICAL: Evaluate ONLY the model's process frames (middle images/video), NOT the input or ground truth.

Reference Information

- **Instruction**: "Restore the scrambled jigsaw puzzle to its original complete image."
- **Ground Truth**: The last image shows the correct complete layout for comparison.

Evaluation Dimensions

Evaluate the overall reasoning process along three percentage-based dimensions (0–100):

Background Consistency (0–100): In what percentage of frames does the global color, lighting, and visual continuity remain stable? Frames with visual noise, artifact introduction, or tile color mismatch reduce the score.

Rule Obey (0–100): In what percentage of frames does the model follow jigsaw reconstruction rules? Each frame should show valid moves (placing or rotating tiles without overlaps or deletions).

Beneficial Action (0–100): In what percentage of frames do the actions effectively move the puzzle toward completion? Frames that position or correct tiles to their GT location are beneficial; random or regressive moves reduce the score.

Visual Quality (0–100): In what percentage of frames is the visual quality maintained at a high standard? Frames should be clear, sharp, free from visual noise, artifacts, blurriness, or distortions. Image content should be easily recognizable and distinguishable. Frames with quality degradation, noise, artifacts, blur, or unclear content reduce this score.

Output Format

Provide your evaluation strictly in JSON format:

```
{ { "Background_Consistency": integer 0--100,
"Rule_Obey": integer 0--100,
"Beneficial_Action": integer 0--100,
"Visual_Quality": integer 0--100,
"Explanation": {
"Background": "Explain how stable and visually consistent the background
remained.",
"Rule": "Explain how often the puzzle assembly rules were respected (no
overlaps, valid placements).",
"Beneficial": "Explain how frequently the model made progress toward the
correct GT arrangement.",
"Visual_Quality": "Describe the overall visual quality across frames
(clarity, sharpness, absence of artifacts/noise/blur)."
}
} }
```

Evaluation Principles

- Percentages must be integers (e.g., 90 for 90%).
- Evaluate the full sequence, not only the last few frames.
- High scores require consistent, rule-abiding, and progressive assembly behavior.
- Visual comparison with the ground truth should guide all judgments.

Table 11. Symbolic Reasoning - Jigsaw Puzzle CoT Template.

Symbolic Reasoning - Jigsaw Puzzle Binary Template

You are a visual reasoning evaluator specialized in image reconstruction and jigsaw puzzle restoration tasks.
Task Description
The goal of this task is to restore a scrambled jigsaw puzzle image back to its original complete image. The puzzle may have different grid sizes (e.g., 2×2, 4×4, 7×7), and the model must correctly reassemble all tiles in their proper positions to match the original image.
Input Information Structure
You will receive images in the following order:

1. **Input Image (First given image)**: The scrambled jigsaw puzzle image showing tiles in incorrect positions.
2. **Model Output (Second given image - PRIMARY EVALUATION TARGET)**: The image produced by the model showing its reconstructed result. **This is what you should evaluate.**
3. **Ground Truth Image (Third given image, if provided)**: The correct full image showing the original unshuffled arrangement. This is provided for comparison reference only, **NOT for evaluation**.

CRITICAL: Evaluate ONLY the model's output image (second image), NOT the input or ground truth.
Reference Information

- **Instruction**: "Restore the scrambled jigsaw puzzle to its original complete image."
- **Ground Truth**: The third image shows the correct full image for comparison.

Evaluation Dimensions
Evaluate the model's output along three binary dimensions (0 or 1):
Background Consistency (0 or 1): Are all visual elements (color tones, lighting, and texture continuity) globally consistent after restoration? The restored image should not introduce artifacts, gaps, or color mismatches between tiles.
Rule Obey (0 or 1): Does the model follow the basic jigsaw reconstruction rules? Each tile should be placed once, without overlap or missing regions, and edges should align correctly.
Reasoning Accuracy (0 or 1): Is the final restored image visually identical or sufficiently close to the ground truth? The global spatial arrangement must fully match the GT layout.
Visual Quality (0 or 1): Does the model output maintain high visual quality without artifacts or degradation? The output image should be clear, sharp, free from visual noise, artifacts, blurriness, or distortions. Image content should be easily recognizable and distinguishable. Assign 0 if any of these quality issues are present, 1 if the image quality is consistently high.
Output Format
Provide your evaluation strictly in JSON format:

```
{ { "Background_Consistency": 0 or 1,
"Rule_Obey": 0 or 1,
"Reasoning_Accuracy": 0 or 1,
"Visual_Quality": 0 or 1,
"Explanation": {
"Background": "Describe whether color, lighting, and texture continuity were
preserved.",
"Rule": "Describe if all tiles were used and correctly aligned without
overlaps or missing areas.",
"Accuracy": "Describe whether the final image matches the GT layout and
visual composition.",
"Visual_Quality": "Describe the overall visual quality of the output image
(clarity, sharpness, absence of artifacts/noise/blur)."
}
} }
```

Evaluation Principles

- Assign "1" only if the condition is fully satisfied.
- Evaluate based on visual alignment and texture continuity between Input, GT, and Model Output.
- Minor seam visibility does not affect correctness if the arrangement is fully correct.
- Explanations should be objective and concise.

Table 12. Symbolic Reasoning - Jigsaw Puzzle Binary Template.

Symbolic Reasoning - Function Plotting CoT Template

You are an expert visual reasoning evaluator for function plotting processes.

Task Description

The model generates a sequence of frames (or a video) that illustrates how a mathematical function $y = f(x)$ is drawn step by step over a specified domain. Each frame may show intermediate elements such as coordinate axes, sample points, partial curves, or annotations.

Input Information Structure

You will receive images or video in the following order:

1. **First Image (Input Image)**: The original image showing the mathematical formula and domain specification (e.g., $y = x^2 \cdot \sin^2(x) + x$ over $x \in [-10.0, 10.0]$).
2. **Middle Images or Frames (Model Output - PRIMARY EVALUATION TARGET)**: A sequence of frames or a video showing the model's step-by-step function plotting process (coordinate axes setup, sample points, partial curves). **This is what you should evaluate.**
3. **Last Image (Ground Truth Reference)**: The correct completed curve plot and, if applicable, an idealized plotting progression. This is provided for comparison reference only, **NOT for evaluation**.

CRITICAL: Evaluate ONLY the model's process frames (middle images/video), NOT the input or ground truth.

Reference Information

- **Instruction**: "Draw the curve corresponding to the mathematical expression shown in the image over the specified domain."
- **Ground Truth**: The last image shows the correct completed curve plot for comparison.

Evaluation Dimensions

Evaluate the overall reasoning process along the following percentage-based dimensions (0–100):

Background Consistency (0–100): In what percentage of frames does the formula text, domain, and coordinate system remain stable and correct? Alterations to labels or axis structure lower the score.

Rule Obey (0–100): In what percentage of frames does the progressive curve follow the correct mathematical rule? Points or partial lines must be located where $f(x)$ predicts, and the specified domain range must not be exceeded.

Beneficial Action (0–100): In what percentage of frames do the drawing actions meaningfully progress toward completing the correct curve? Steps that add correct data points or refine the curve are counted as beneficial; random or inconsistent updates are penalized.

Visual Quality (0–100): In what percentage of frames is the visual quality maintained at a high standard? Frames should be clear, sharp, free from visual noise, artifacts, blurriness, or distortions. Image content should be easily recognizable and distinguishable.

Output Format

Return your evaluation in the following strict JSON format:

```
{ { "Background_Consistency": integer 0--100,
"Rule_Obey": integer 0--100,
"Beneficial_Action": integer 0--100,
"Visual_Quality": integer 0--100,
"Explanation": {
"Background": "Explain whether the formula and coordinate system remained
consistent.",
"Rule": "Describe how often the intermediate steps followed the correct
mathematical behavior.",
"Beneficial": "Describe how effectively each step contributed to completing
the correct curve.",
"Visual_Quality": "Describe the overall visual quality across frames
(clarity, sharpness, absence of artifacts/noise/blur)."} } }
```

Evaluation Principles

- Percentages must be integers (e.g., 90 for 90%).
- Evaluate the full sequence of frames, focusing on stability and progressive correctness.
- High Rule Obey requires most intermediate plots to conform to the true mathematical function.
- High Beneficial Action requires consistent forward progress toward the ground truth curve.
- Base your evaluation solely on visual and mathematical reasoning; do not infer beyond the given content.

Table 13. Symbolic Reasoning - Function Plotting CoT Template.

Symbolic Reasoning - Function Plotting Binary Template

You are a visual reasoning evaluator specialized in mathematical visualization and function plotting.
Task Description
The task is to draw the curve corresponding to a given mathematical expression over a specified domain. The model receives:

- A mathematical formula (e.g., $y = x^2 \cdot \sin^2(x) + x$)
- A domain (e.g., $x \in [-10.0, 10.0]$)

and is instructed to plot the correct function curve across this domain.
Input Information Structure
You will receive images in the following order:

1. **Input Image (First given image)**: The image showing the mathematical formula and the domain specification.
2. **Model Output (Second given image - PRIMARY EVALUATION TARGET)**: The image generated by the model showing its plotted curve. **This is what you should evaluate.**
3. **Ground Truth Image (Third given image, Optional)**: The correct plot of the function over the same domain. This is provided for comparison reference only, **NOT for evaluation**.

CRITICAL: Evaluate ONLY the model's output image (second image), NOT the input or ground truth.
Reference Information

- **Instruction**: "Draw the curve corresponding to the mathematical expression shown in the image over the specified domain."
- **Ground Truth**: The third image shows the correct plot of the function for comparison.

Evaluation Dimensions
Evaluate the model's output image along the following binary dimensions (0 or 1):
Background Consistency (0 or 1): Does the model preserve all non-plot elements from the input, including the formula box, domain text, axis labels, and overall layout?
Rule Obey (0 or 1): Does the plotted curve obey the mathematical rule defined by the formula and the specified domain? The curve should exhibit correct functional behavior (e.g., symmetry, periodicity, growth trends) and stay within the valid x-range.
Reasoning Accuracy (0 or 1): Is the plotted curve visually identical or sufficiently close to the ground truth plot? The overall shape, peaks, crossings, and range must align with the GT.
Visual Quality (0 or 1): Does the model output maintain high visual quality without artifacts or degradation? The output image should be clear, sharp, free from visual noise, artifacts, blurriness, or distortions.
Output Format
Provide your evaluation strictly in the following JSON format:

```
{ { "Background_Consistency": 0 or 1,
"Rule_Obey": 0 or 1,
"Reasoning_Accuracy": 0 or 1,
"Visual_Quality": 0 or 1,
"Explanation": {
"Background": "Describe whether the non-plot text such as formula and domain
remained consistent.",
"Rule": "Describe whether the curve obeyed the given mathematical formula
within the domain.",
"Accuracy": "Describe whether the plotted curve matches the ground truth in
shape and range.",
"Visual_Quality": "Describe the overall visual quality of the output image
(clarity, sharpness, absence of artifacts/noise/blur)."
}
} }
```

Evaluation Principles

- Assign "1" only if the criterion is fully satisfied without visual or logical errors.
- Evaluation is based on the visual and mathematical alignment between Input, Ground Truth, and Model Output.
- Deviations in function shape or domain range result in a score of 0 for Rule Obey and Reasoning Accuracy.
- Explanations should be precise, objective, and concise.

Table 14. Symbolic Reasoning - Function Plotting Binary Template.

Symbolic Reasoning - Algebraic Calculation CoT Template

You are an expert visual reasoning evaluator for mathematical problem-solving processes, focusing on algebraic equations.
Task Description
The task involves solving algebraic equations or systems (linear or quadratic) step by step. Each frame or intermediate image shows the model's reasoning process, such as simplifying, substituting, or solving for variables.
Input Information Structure
You will receive images or video in the following order:

1. **First Image (Input Image)**: The original problem containing the equation or equations to be solved.
2. **Middle Images or Frames (Model Output - PRIMARY EVALUATION TARGET)**: A sequence of frames or a video showing the model's step-by-step algebraic reasoning process (simplifying, substituting, solving for variables). **This is what you should evaluate.**
3. **Last Image (Ground Truth Reference)**: The correct final solution showing the complete solving process and final answer. This is provided for comparison reference only, **NOT for evaluation**.

CRITICAL: Evaluate ONLY the model's process frames (middle images/video), NOT the input or ground truth.
Reference Information

- **Instruction**: "Please solve this equation."
- **Ground Truth**: The last image shows the correct final solution and a logically valid sequence of transformations for comparison.

Evaluation Dimensions
Evaluate the reasoning process along the following percentage-based dimensions (0–100):
Background Consistency (0–100): In what percentage of frames does the original equation text remain intact and readable? Changes to problem statements or unrelated additions reduce this score.
Rule Obey (0–100): In what percentage of steps are algebraic operations performed correctly? Valid transformations include addition, subtraction, factoring, substitution, and simplification according to algebraic laws.
Beneficial Action (0–100): In what percentage of steps does the reasoning progress toward the correct solution? Steps that correctly reduce equation complexity or move closer to the ground truth contribute positively.
Visual Quality (0–100): In what percentage of frames is the visual quality maintained at a high standard? Frames should be clear, sharp, free from visual noise, artifacts, blurriness, or distortions.
Output Format
Return your evaluation strictly in the following JSON format:

```
{ { "Background_Consistency": integer 0--100,
"Rule_Obey": integer 0--100,
"Beneficial_Action": integer 0--100,
"Visual_Quality": integer 0--100,
"Explanation": {
"Background": "Explain how consistently the problem layout was preserved
across steps.",
"Rule": "Describe how often algebraic rules were obeyed in the reasoning
process.",
"Beneficial": "Describe how effectively the model's steps advanced toward
the correct solution.",
"Visual_Quality": "Describe the overall visual quality across frames
(clarity, sharpness, absence of artifacts/noise/blur)."
}
} }
```

Evaluation Principles

- Percentages must be integers.
- Consider the full reasoning sequence, not isolated steps.
- Reward logical consistency and stepwise correctness even if the final result is not exact.
- Base all judgments on visual comparison between Input, Ground Truth, and Process frames.
- Explanations should be concise and factual.

Table 15. Symbolic Reasoning - Algebraic Calculation CoT Template.

Symbolic Reasoning - Algebraic Calculation Binary Template

You are a visual reasoning evaluator specialized in mathematical problem solving, specifically algebraic equations.
Task Description
The goal of this task is to solve an algebraic equation or system of equations shown in the image. The equations may include:

- Single-variable linear equations (e.g., $2x + 3 = 5$)
- Single-variable quadratic equations (e.g., $x^2 + 2x - 3 = 0$)
- Two-variable linear equations (e.g., $2x + 9y = 19$ and $2x + 6y = 8$)

Input Information Structure
You will receive images in the following order:

1. **Input Image (First given image)**: The original problem containing the equation or equations to be solved.
2. **Model Output (Second given image - PRIMARY EVALUATION TARGET)**: The image produced by the model containing its proposed solution. **This is what you should evaluate.**
3. **Ground Truth Image (Third given image, if provided)**: The correct final answer(s) and solution form. This is provided for comparison reference only, **NOT for evaluation**.

CRITICAL: Evaluate ONLY the model's output image (second image), NOT the input or ground truth.
Reference Information

- **Instruction**: "Please solve this equation."
- **Ground Truth**: The third image shows the correct final answer(s) and solution form for comparison.

Evaluation Dimensions
Evaluate the model output along the following binary dimensions (0 or 1):
Background Consistency (0 or 1): Does the output preserve the original problem text and structure? The equation layout, fonts, and non-solution content should remain intact and legible.
Rule Obey (0 or 1): Does the model follow correct algebraic transformation rules? This includes valid symbolic manipulation (e.g., addition, subtraction, factoring, substitution). Violations such as illegal simplification or missing logical steps result in 0.
Reasoning Accuracy (0 or 1): Is the final solution correct and fully consistent with the ground truth? For single-variable equations, check the numerical or symbolic answer. For multi-variable systems, all variable pairs (x, y) must match the GT values.
Visual Quality (0 or 1): Does the model output maintain high visual quality without artifacts or degradation? The output image should be clear, sharp, free from visual noise, artifacts, blurriness, or distortions. Image content should be easily recognizable and distinguishable.
Output Format
Provide your evaluation strictly in the following JSON format:

```
{ { "Background_Consistency": 0 or 1,
"Rule_Obey": 0 or 1,
"Reasoning_Accuracy": 0 or 1,
"Visual_Quality": 0 or 1,
"Explanation": {
"Background": "Describe whether the equation layout was preserved.",
"Rule": "Describe whether algebraic steps followed mathematical rules.",
"Accuracy": "Describe whether the final result matches the correct
solution.",
"Visual_Quality": "Describe the overall visual quality of the output image
(clarity, sharpness, absence of artifacts/noise/blur)."
}
} }
```

Evaluation Principles

- Assign "1" only if the criterion is fully satisfied without mistakes.
- Visual comparison should ensure no loss of context or formatting corruption.
- For partial correctness (e.g., right procedure but wrong result), only **Rule Obey** may receive 1; all other dimensions must be 0.

Table 16. Symbolic Reasoning - Algebraic Calculation Binary Template.

Symbolic Reasoning - Block Building CoT Template

You are an expert visual reasoning evaluator for LEGO blocks processes.

Task Description

Assemble the LEGO bricks step by step. You can set aside unused pieces on one side and already assembled pieces on the other. The model generates a sequence of frames (or a video) showing its reasoning and assembly process, where LEGO blocks are gradually built.

Input Information Structure

You will receive images or video in the following order:

1. **First Image (Input Image)**: The initial image showing the unassembled LEGO pieces scattered or arranged separately.
2. **Middle Images or Frames (Model Output - PRIMARY EVALUATION TARGET)**: A sequence of frames or a video showing the model's step-by-step reasoning process as it assembles the LEGO blocks. **This is what you should evaluate.**
3. **Last Image (Ground Truth Reference)**: The correct complete image showing the final assembled LEGO structure. This is provided for comparison reference only, **NOT for evaluation**.

CRITICAL: Evaluate ONLY the model's process frames (middle images/video), NOT the input or ground truth.

Reference Information

- **Instruction**: "For this LEGO set, place the orange pieces on the bottom as the first layer, then the gray pieces on the second layer, the yellow pieces on the third layer, and the black pieces on the top layer."
- **Ground Truth**: The last image shows the correct complete layout for comparison.

Evaluation Dimensions

Evaluate the overall reasoning process along three percentage-based dimensions (0–100):

Background Consistency (0–100): In what percentage of frames does the global color, lighting, and visual continuity remain stable? Frames with visual noise, artifact introduction, or block color mismatch reduce the score.

Rule Obey (0–100): In what percentage of frames does the model follow building rules? Each frame should show valid assembly steps (proper layer ordering, correct block placement, no overlaps or floating pieces).

Beneficial Action (0–100): In what percentage of frames do the actions effectively build the blocks toward completion? Frames that position or correct blocks toward their ground truth location are beneficial.

Visual Quality (0–100): In what percentage of frames is the visual quality maintained at a high standard? Frames should be clear, sharp, free from visual noise, artifacts, blurriness, or distortions. Image content should be easily recognizable and distinguishable. Frames with quality degradation, noise, artifacts, blur, or unclear content reduce this score.

Output Format

Provide your evaluation strictly in the following JSON format:

```
{ { "Background_Consistency": integer 0--100,
"Rule_Obey": integer 0--100,
"Beneficial_Action": integer 0--100,
"Visual_Quality": integer 0--100,
"Explanation": {
"Background": "Explain how stable and visually consistent the background
remained.",
"Rule": "Explain how often the LEGO assembly rules were respected (no
overlaps, valid placements).",
"Beneficial": "Explain how frequently the model made progress toward the
correct GT arrangement.",
"Visual_Quality": "Describe the overall visual quality across frames
(clarity, sharpness, absence of artifacts/noise/blur)."
}
} }
```

Evaluation Principles

- Percentages must be integers (e.g., 90 for 90%).
- Evaluate the full sequence, not only the last few frames.
- High scores require consistent, rule-abiding, and progressive assembly behavior.
- Visual comparison with the ground truth should guide all judgments.

Table 17. Symbolic Reasoning - Block Building CoT Template.

Symbolic Reasoning - Block Building Binary Template

You are an expert visual reasoning evaluator for LEGO blocks processes.
Task Description
Assemble the LEGO bricks step by step. You can set aside unused pieces on one side and already assembled pieces on the other.
Input Information Structure
You will receive images in the following order:

1. **Input Image (First given image)**: The initial image showing the unassembled LEGO pieces scattered or arranged separately.
2. **Model Output (Second given image - PRIMARY EVALUATION TARGET)**: The image produced by the model containing its proposed assembled LEGO structure. **This is what you should evaluate.**
3. **Ground Truth Image (Third given image, if provided)**: The correct complete image showing the final assembled LEGO structure. This is provided for comparison reference only, **NOT for evaluation**.

CRITICAL: Evaluate ONLY the model's output image (second image), NOT the input or ground truth.
Reference Information

- **Instruction**: "For this LEGO set, place the orange pieces on the bottom as the first layer, then the gray pieces on the second layer, the yellow pieces on the third layer, and the black pieces on the top layer."
- **Ground Truth**: The third image shows the correct complete layout for comparison.

Evaluation Dimensions
Evaluate the model output along three binary dimensions (0 or 1):
Background Consistency (0 or 1): Does the global color, lighting, and visual continuity remain stable? Visual noise, artifact introduction, or block color mismatch should result in 0.
Rule Obey (0 or 1): Does the model output follow LEGO building rules? The result should show a valid final structure (proper layer ordering, correct block placement, no overlaps or floating pieces).
Reasoning Accuracy (0 or 1): Is the final solution correct and fully consistent with the ground truth? Blocks must be arranged in the correct layers (orange bottom, gray second, yellow third, black top).
Visual Quality (0 or 1): Does the model output maintain high visual quality without artifacts or degradation? The output image should be clear, sharp, free from visual noise, artifacts, blurriness, or distortions. Image content should be easily recognizable and distinguishable. Assign 0 if any of these quality issues are present, 1 if the image quality is consistently high.
Output Format
Provide your evaluation strictly in the following JSON format:

```
{ { "Background_Consistency": 0 or 1,
"Rule_Obey": 0 or 1,
"Reasoning_Accuracy": 0 or 1,
"Visual_Quality": 0 or 1,
"Explanation": {
"Background": "Describe whether the visual layout was preserved.",
"Rule": "Describe if the output follows valid LEGO building rules.",
"Accuracy": "Describe whether the final result matches the correct
solution.",
"Visual_Quality": "Describe the overall visual quality of the output image
(clarity, sharpness, absence of artifacts/noise/blur)."
}
} }
```

Evaluation Principles

- Assign "1" only if the criterion is fully satisfied without mistakes.
- Visual comparison should ensure no loss of context or formatting corruption.
- For partial correctness, only **Rule Obey** may receive 1; all other dimensions must be 0.

Table 18. Symbolic Reasoning - Block Building Binary Template.

Symbolic Reasoning - Klotski Puzzle CoT Template

You are an expert visual reasoning evaluator for Klotski Puzzle processes.

Task Description

To play the game of Digital Klotski Puzzle, by translating the number squares several times, only one number square can be moved to an empty space each time, so that the final numbers from the upper left to the lower right are arranged in order from 1 to 8, and the lower right corner is empty. The model generates a sequence of frames or a video showing its reasoning and movement process, where number squares are gradually moved.

Input Information Structure

You will receive images or video in the following order:

1. **First Image (Input Image)**: The initial klotski puzzle state showing a scrambled arrangement of number squares from 1 to 8 with one empty space.
2. **Middle Images or Frames (Model Output - PRIMARY EVALUATION TARGET)**: A sequence of frames or a video showing the model's step-by-step reasoning process as it moves the number squares. **This is what you should evaluate.**
3. **Last Image (Ground Truth Reference)**: The correct complete layout with numbers 1 to 8 arranged in order from the upper left to the lower right and the empty space in the lower right corner. This is provided for comparison reference only, **NOT for evaluation**.

CRITICAL: Evaluate ONLY the model's process frames (middle images/video), NOT the input or ground truth.

Reference Information

- **Instruction**: "Only one number square can be moved to an empty space at each step, so that the final numbers from the upper left to the lower right are arranged in order from 1 to 8."
- **Ground Truth**: The last image shows the correct complete layout for comparison.

Evaluation Dimensions

Evaluate the overall reasoning process along the following percentage-based dimensions (0–100):

Background Consistency (0–100): The percentage of frames in which global color, lighting, and visual continuity remain stable.

Rule Obey (0–100): The percentage of frames in which the model follows Klotski Puzzle movement rules, with exactly one square moved into the empty space at each step and no overlaps or deletions.

Beneficial Action (0–100): The percentage of frames in which actions effectively move the number squares toward the correct final configuration.

Visual Quality (0–100): The percentage of frames that maintain high visual quality, remaining clear, sharp, and free from artifacts, noise, blurriness, or distortions.

Output Format

Provide your evaluation strictly in the following JSON format:

```
{ "Background_Consistency": integer 0--100,
"Rule_Obey": integer 0--100,
"Beneficial_Action": integer 0--100,
"Visual_Quality": integer 0--100,
"Explanation": {
"Background": "Explain how stable and visually consistent the background
remained.",
"Rule": "Explain how often the square movement rules were respected.",
"Beneficial": "Explain how frequently the model made progress toward the
correct ground truth arrangement.",
"Visual_Quality": "Describe the overall visual quality across frames
(clarity, sharpness, absence of artifacts, noise, or blur)."
}
}
```

Evaluation Principles

- Percentages must be integers.
- Evaluate the full sequence, not only the final frames.
- High scores require consistent, rule-abiding, and progressive movement behavior.
- Judgments should be guided by visual comparison with the ground truth.

Table 19. Symbolic Reasoning - Klotski Puzzle CoT Template.

Symbolic Reasoning - Klotski Puzzle Binary Template

You are an expert visual reasoning evaluator for Klotski Puzzle processes.
Task Description
To play the game of Digital Klotski Puzzle, by translating the number squares several times, only one number square can be moved to an empty space each time, so that the final numbers from the upper left to the lower right are arranged in order from 1 to 8, and the lower right corner is empty.
Input Information Structure
You will receive images in the following order:

1. **Input Image (First given image)**: The initial klotski puzzle state showing a scrambled arrangement of number squares from 1 to 8 with one empty space.
2. **Model Output (Second given image - PRIMARY EVALUATION TARGET)**: The image produced by the model containing its proposed solution. **This is what you should evaluate.**
3. **Ground Truth Image (Third given image, Optional)**: The correct complete image showing the final layout with numbers 1 to 8 arranged in order from the upper left to the lower right. This is provided for comparison reference only, **NOT for evaluation**.

CRITICAL: Evaluate ONLY the model's output image (second image), NOT the input or ground truth.
Reference Information

- **Instruction**: "Only one number square can be moved to an empty space each time, so that the final numbers from the upper left to the lower right are arranged in order from 1 to 8."
- **Ground Truth**: The third image shows the correct complete layout for comparison.

Evaluation Dimensions
Evaluate the model output along the following binary dimensions (0 or 1):
Background Consistency (0 or 1): Whether the global color, lighting, and visual continuity remain stable without visual noise, artifacts, or tile color mismatch.
Rule Obey (0 or 1): Whether the output follows valid Klotski Puzzle movement rules, with exactly one number square moved into the empty space at each step and a proper final grid layout without overlaps or deletions.
Reasoning Accuracy (0 or 1): Whether the final state matches the ground truth, with numbers 1 to 8 arranged in order from the upper left to the lower right and the empty space in the lower right corner.
Visual Quality (0 or 1): Whether the output image maintains high visual quality, remaining clear, sharp, and free from artifacts, noise, blurriness, or distortions.
Output Format
Provide your evaluation strictly in the following JSON format:

```
{ "Background_Consistency": 0 or 1,
"Rule_Obey": 0 or 1,
"Reasoning_Accuracy": 0 or 1,
"Visual_Quality": 0 or 1,
"Explanation": {
"Background": "Describe whether the visual layout was preserved.",
"Rule": "Describe whether the output follows valid Klotski Puzzle rules.",
"Accuracy": "Describe whether the final result matches the correct
solution.",
"Visual_Quality": "Describe the overall visual quality of the output image
(clarity, sharpness, absence of artifacts, noise, or blur)."
}
}
```

Evaluation Principles

- Assign "1" only if the criterion is fully satisfied without mistakes.
- Visual comparison should ensure no loss of context or formatting corruption.
- For partial correctness, only Rule Obey may receive 1; all other dimensions must be 0.

Table 20. Symbolic Reasoning - Klotski Puzzle Binary Template.

Symbolic Reasoning - Maze Navigation Binary Template

You are a visual reasoning evaluator specializing in 2D maze navigation tasks.
Task Description
The goal of this task is to navigate a 2D maze starting from the green letter "S" and ending at the red letter "E". The model must draw a red path that successfully connects "S" to "E" without crossing any maze walls or altering the maze layout.
Input Information Structure
You will receive images in the following order:

1. **Input Image (First given image)**: The original maze with no path drawn, showing only the maze structure with "S" (start) and "E" (end) markers.
2. **Model Output (Second given image - PRIMARY EVALUATION TARGET)**: The maze image generated by the model with its proposed red path. **This is what you should evaluate.**
3. **Ground Truth Image (Third given image, Optional)**: The correct final maze image showing the reference solution path. This is provided for comparison reference only, **NOT for evaluation**.

CRITICAL: Evaluate ONLY the model's output image (second image), NOT the input or ground truth.
Reference Information

- **Instruction**: "Starting from the green letter "S" and ending at the red letter "E", draw a red path that successfully navigates through the maze without crossing walls."
- **Ground Truth**: The third image shows the correct solution for comparison.

Evaluation Dimensions
Evaluate the model output image along the following binary dimensions (0 or 1):
Background Consistency (0 or 1): Whether the maze structure remains identical to the input image, with no added, removed, or modified walls or boundaries.
Rule Obey (0 or 1): Whether the red path strictly follows maze rules by staying within open corridors and never crossing or overlapping maze walls.
Reasoning Accuracy (0 or 1): Whether the red path correctly and continuously connects "S" to "E" as in the ground truth reference.
Visual Quality (0 or 1): Whether the output image maintains high visual quality, remaining clear, sharp, and free from artifacts, noise, blurriness, or distortions.
Output Format
Provide your evaluation strictly in the following JSON format:

```
{ "Background_Consistency": 0 or 1,
"Rule_Obey": 0 or 1,
"Reasoning_Accuracy": 0 or 1,
"Visual_Quality": 0 or 1,
"Explanation": {
"Background": "Briefly describe whether the maze structure was preserved.",
"Rule": "Explain whether the path followed valid maze corridors without
crossing walls.",
"Accuracy": "Explain whether the path correctly connected S to E compared to
the ground truth.",
"Visual_Quality": "Describe the overall visual quality of the output image
(clarity, sharpness, absence of artifacts, noise, or blur)."
}
}
```

Evaluation Principles

- Assign "1" only if the criterion is fully satisfied without any errors.
- Be strict: partial correctness, such as a path that does not reach "E", must be scored as 0.
- Explanations should be concise and factual.

Table 21. Symbolic Reasoning - Maze Navigation Binary Template.

Maze Navigation CoT Template

You are an expert visual reasoning evaluator for multi-step 2D maze-solving processes.

Task Description

The goal of this task is to navigate a 2D maze starting from the green letter "S" and ending at the red letter "E". The model produces a sequence of frames (or a video) showing how it progressively draws the red path through the maze.

Input Information Structure

You will receive images/video in the following order:

1. **First Image (Input Image)**: The original maze with no path drawn, showing only the maze structure with "S" (start) and "E" (end) markers.
2. **Middle Images/Frames (Model Output - PRIMARY EVALUATION TARGET)**: A sequence of frames or video showing the model's step-by-step reasoning process as it draws the red path. **This is what you should evaluate.**
   - **Important**: ALL images between the first image and the optional last ground truth image are part of the model's output.
   - If a ground truth image is provided, it will appear AFTER all model frames as the very last image.
   - If NO ground truth image is provided, ALL images after the first image are model output frames.
3. **Last Image (Ground Truth Reference - OPTIONAL)**: If provided, this will be the very last image in the sequence, appearing AFTER all model output frames. The correct final maze image showing the complete correct path from "S" to "E". This is provided for comparison reference only, NOT for evaluation.
   - **Note**: When no ground truth image is provided, evaluate based on the instruction and maze navigation rules.

CRITICAL: Evaluate ONLY the model's process frames (middle images/video), NOT the input or the optional ground truth reference image.

Reference Information

- **Instruction**: "Starting from the green letter 'S' and ending at the red letter 'E', draw a red path that successfully navigates through the maze without crossing walls."
- **Ground Truth**: The last image shows the correct solution for comparison.

Evaluation Dimensions

Assess the model's reasoning process (middle frames) along three percentage-based dimensions (0–100):

- **Background Consistency (0–100)**: In what percentage of the model's frames does the maze structure remain unchanged? (Any alteration of walls, boundaries, or maze lines in the model's output decreases this score.)
- **Rule Obey (0–100)**: In what percentage of the model's frames does the red path obey maze rules? (Crossing walls, overlapping borders, or jumping across corridors in the model's output reduces this score.)
- **Beneficial Action (0–100)**: In what percentage of the model's frames does the model make progress toward the correct solution? (Steps that extend the correct path toward "E" increase this score; random or regressive actions reduce it.)
- **Visual Quality (0–100)**: In what percentage of the model's frames is the visual quality maintained at a high standard? (Frames should be clear, sharp, free from visual noise, artifacts, blurriness, or distortions. Image content should be easily recognizable and distinguishable. Frames with quality degradation, noise, artifacts, blur, or unclear content reduce this score.)

Output Format

Return your evaluation in the following JSON format:

```
{{ "Background_Consistency": integer 0--100, "Rule_Obey": integer 0--100,
"Beneficial_Action": integer 0--100, "Visual_Quality": integer 0--100,
"Explanation": { "Background": "Explain how often the maze layout stayed
intact across the model's frames.", "Rule": "Explain how frequently
maze rules were followed during the model's path drawing.", "Beneficial":
"Describe whether the model's actions generally moved closer to solving the
maze.", "Visual_Quality": "Describe the overall visual quality across frames
(clarity, sharpness, absence of artifacts/noise/blur)." } }}
```

Evaluation Principles

- Percentages should be integers (e.g., 85 for 85%).
- Evaluate the full temporal sequence of the MODEL'S OUTPUT — consider overall trends, not single-frame anomalies.
- A high Beneficial Action score requires consistent, goal-directed steps toward "E" in the model's process.
- Compare the model's process against the ground truth reference (last image) to assess correctness.
- Provide concise, factual explanations for each dimension.

Table 22. Maze Navigation CoT Template.