ROVER: BENCHMARKING RECIPROCAL CROSS-MODAL REASONING FOR OMNIMODAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models (UMMs) have shown remarkable advances in understanding and generating text and images. However, prevailing evaluations treat these abilities in isolation, such that tasks with multimodal inputs and outputs are scored primarily through unimodal reasoning: textual benchmarks emphasize language-based reasoning, while visual benchmarks emphasize reasoning outcomes manifested in the pixels. As such, existing benchmarks rarely require the use of one modality to guide, verify, or refine outputs in the other. They therefore fail to capture a central aspiration of unified multimodal models, namely to support seamless reasoning across modalities. We address this gap with ROVER, a human-annotated benchmark that explicitly targets reciprocal cross-modal reasoning, which contains 1,285 tasks grounded in 2,048 images, spanning two complementary settings. **Verbally-augmented reasoning for visual generation** evaluates whether models can use structured verbal prompts and reasoning chains to guide faithful image synthesis. **Visually-augmented reasoning for verbal generation** evaluates whether models can generate intermediate visualizations that strengthen their own reasoning processes. Experiments on 17 state-of-the-art UMMs reveal two key findings: (i) cross-modal reasoning capabilities strongly correlate with visual generation performance, particularly for interleaved image–text generation; and (ii) current models remain severely limited in visual-augmented reasoning, showing relative strength in perception and physical modeling but weakness in logical tasks. These results highlight reciprocal cross-modal reasoning as a critical frontier for enabling true omnimodal generation.

🌐 More information on **Anonymous Page**: *https://anony0923.github.io*.

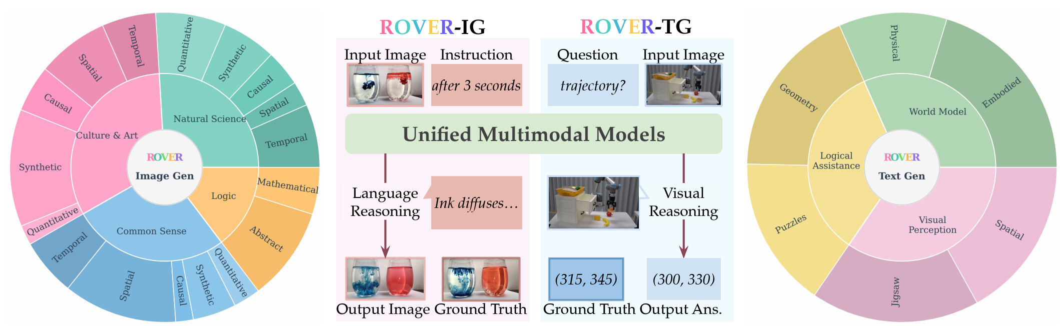


Figure 1: The ROVER benchmark. ROVER evaluates UMMs through reciprocal cross-modal reasoning: **ROVER-IG** (left) requires generating images with language-augmented reasoning, while **ROVER-TG** (right) requires generating text answers with visually-augmented reasoning.

1 INTRODUCTION

The development of *unified multimodal models* (also referred to as *omnimodal models*) has sparked considerable interest in their understanding and generation capabilities across images and text (Comanici et al., 2025; Hurst et al., 2024; Tong et al., 2024; Deng et al., 2025b; Xu et al., 2025b). However, prevailing evaluations treat these abilities in isolation, such that tasks with multimodal inputs and outputs are scored primarily through unimodal reasoning: textual benchmarks emphasize language-based reasoning, while visual benchmarks emphasize reasoning outcomes manifested in the pixels. On the language side, evaluation focuses on generating text in response to an image and an accompanying question, thereby testing perceptual understanding (Chen et al., 2024; Liu et al., 2024; Yu et al., 2024) and reasoning (Lu et al., 2023; Yue et al., 2024; Wang et al., 2024; Hao et al.,

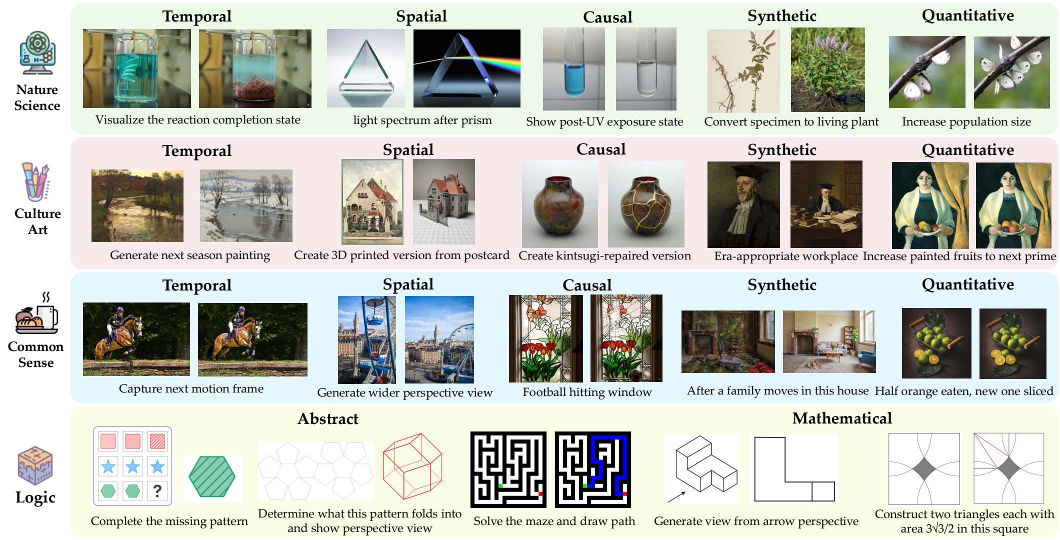


Figure 2: Overview of **ROVER-IG**, the benchmark for evaluating how unified multimodal models generate images under intensive verbal reasoning. The benchmark spans 4 domains (natural science, culture and art, common sense, and logic), instantiated across 7 reasoning subtasks.

2025; Gao et al., 2025). On the vision side, evaluation centers on generating images conditioned on either instructions or text-image pairs, thereby testing direct image generation (Ghosh et al., 2023; Ma et al., 2024; Niu et al., 2025) or image editing (Kawar et al., 2023; Zhang et al., 2023; Ma et al., 2024; Sheynin et al., 2024; Yu et al., 2025; Liu et al., 2025b; Wu et al., 2025e).

Existing benchmarks rarely evaluate the use of one modality to guide, verify, or refine outputs of the other modality. They therefore fail to capture a central aspiration of unified multimodal models, namely the ability to support seamless reasoning across modalities. We refer to this capability as *reciprocal cross-modal reasoning* as illustrated in Figure 1, meaning the use of information from one modality to inform and improve outputs in another. To benchmark such capability in current unified multimodal models, We present ROVER, a human-annotated and rigorously verified benchmark for *reciprocal cross-modal reasoning*. ROVER comprises over 1,200 tasks grounded in about 2,048 images and targets two complementary settings: (i) **verbally-augmented reasoning for visual generation**, including 4 conceptual domains (natural science, culture and art, common sense, and logic) with high complexity are instantiated across 7 reasoning types: temporal, spatial, causal, synthetic, quantitative, abstract, and mathematical. Each instance provides a textual prompt with an initial image and a *chain of constraints* that a correct output image must satisfy. (ii) **visually-augmented reasoning for verbal generation**, including 6 task variants spanning 3 problem types: physical world modeling for manipulation and dynamics prediction, logical assistance for geometry and puzzle solving, and visual perception enhancement. Instances interleave turns of text and images, requiring the model to *emit visual intermediates* that make downstream reasoning auditable.

Evaluating reciprocal cross-modal reasoning requires assessment of both reasoning steps and the resulting outputs. Text-only metrics overlook visual fidelity, while image-only metrics cannot verify whether the image reflects valid reasoning. Human evaluation provides accurate judgments but is prohibitively expensive at scale. To address this, we adopt a multi-dimensional protocol that combines an automated VLM judge with expert validation on stratified samples. The judge is supplied with rubric cards and reference assets and scores along three reasoning-specific dimensions: (i) the logical coherence of domain-specific reasoning processes, (ii) the alignment of generated outputs with target descriptions or ground-truth answers, and (iii) the consistency between intermediate reasoning steps and the final images or answers. For visual generation tasks, the framework additionally incorporates established image consistency and quality metrics (Hu et al., 2023; Wu et al., 2023; Kirstain et al., 2023; Xu et al., 2023; Brooks et al., 2023). The judge is calibrated with expert explanations, and its agreement with expert evaluations is reported, following recent LLM-as-judge methodologies (Kim et al., 2023; Hu et al., 2023).

Through extensive evaluation of 17 unified multimodal models, our experiments reveal two key findings. First, cross-modal reasoning capabilities are strongly correlated with visual generation

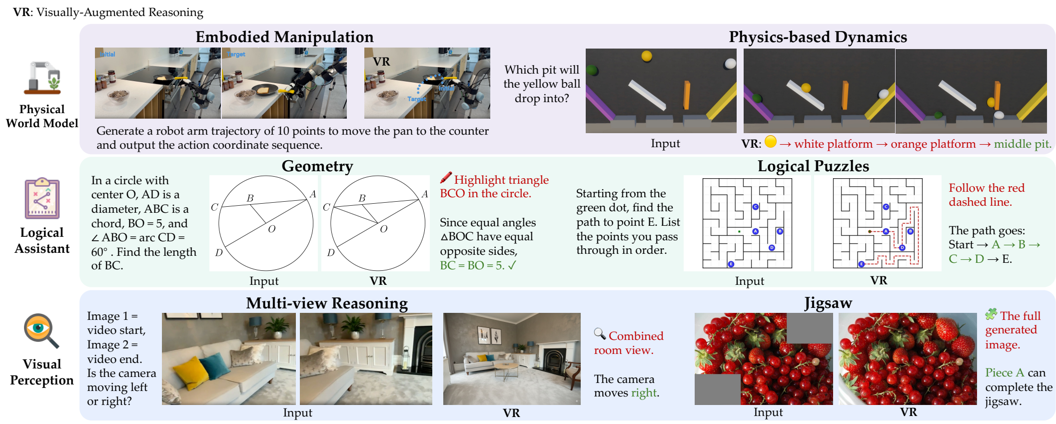


Figure 3: Overview of **ROVER-TG**, the benchmark for evaluating visually-augmented reasoning in verbal generation. The benchmark spans 3 scenarios and 6 subtasks: physical world modeling, logical assistance, and visual perception enhancement.

performance. Especially, models that support interleaved image-text generation achieve superior reasoning and generation results, suggesting that interleaved training data plays a crucial role in developing cross-modal reasoning. Second, unfortunately, current unified models remain severely limited in visually-augmented reasoning. Although models perform relatively well on physical world modeling and visual perception, they remain weak on logic-intensive tasks. This gap indicates that perception over pixels transfers more readily than the acquisition of abstract visual concepts and the ability to reason systematically over them. Taken together, these results underscore the role of reciprocal cross-modal reasoning in enabling transfer across modalities when unified models engage in omnimodal generation.

Our main contributions are summarized as follows:

- We introduce ROVER, the first benchmark that explicitly targets **reciprocal** cross-modal reasoning for visual generation and interleaved multimodal reasoning.
- We provide a principled task taxonomy and a verification ready instance design with process targets and visual artifacts, together with a multi dimensional protocol that scores coherence, alignment, and step to output consistency.
- We report a comprehensive study across 17 close-sourced and open-sourced unified models, revealing sizable gaps and a strong correlation between interleaved generation capability and cross-modal reasoning effectiveness.

2 Related Works

Reasoning for Image Generation. With the emergence of UMMs, multimodal reasoning has garnered increasing attention from the research community. However, the majority of existing work remains focused on instruction comprehension, namely leveraging input images to perform instruction translation and subsequently generate corresponding visual outputs (Jin et al., 2024; Huang et al., 2024; Yang et al., 2024; He et al., 2025; Wu et al., 2025e). Unified-Bench (Yan et al., 2025) employs iterative image-text generation to measure the degree of unification between comprehension and generation models. RISEBench (Zhao et al., 2025) extends beyond prior work by introducing LMM-as-a-judge to evaluate visual rationality in addition to assessing image consistency, yet remains limited to computing similarity scores against human-provided ground truth. However, these benchmarks lack comprehensive evaluation beyond image consistency, particularly overlooking the intermediate processes of reasoning, such as whether the reasoning is sound and whether reasoning aligns with generation. In contrast, ROVER represents the first benchmark to investigate the interplay between reasoning and generation. A detailed comparison can be found in Table 1. A more detailed discussion of related work about unified multimodal models and interleaved reasoning can be found in Appendix F.

Table 1: **Summary of Multimodal Reasoning Benchmarks.** We compare existing works from aspects including: [1]interleave: supports multi-image or multi-turn inputs; [2]process evaluation: evaluates intermediate reasoning steps; [3]vision necessity: requires reasoning grounded in visual understanding; [4]multidimensional evaluation: scores models along multiple dimensions; [5]hybrid evaluation: uses GPT-based judgments instead of purely visual metrics; [6] manual annotations: whether manual annotations and filtering are applied; [7]scale: dataset scale; [8]types: data categories.

Benchmark	Venue	Inter.	Process Eval	Vision Necess.	Multi.	Hybrid Eval	Manual Anno.	Scale	#Types
ReasonPix2Pix (Jin et al., 2024)	arXiv'24	✗	✗	✗	✗	✗	✗	40, 212	1
MetaQuery-Instruct (Pan et al., 2025)	arXiv'25	✓	✗	✓	✗	✗	✗	2.4M	–
EditWorld (Yang et al., 2024)	MM'25	✗	✗	✗	✗	✗	✗	10, 000	7
Reason50K (He et al., 2025)	arXiv'25	✗	✗	✗	✗	✗	✗	51, 039	4
WorldGenBench (Zhang et al., 2025)	arXiv'25	✗	✗	✗	✓	✓	✗	1, 072	2
Unified-Bench (Yan et al., 2025)	arXiv'25	✗	✗	✗	✗	✗	✗	100	1
ReasonEdit (Huang et al., 2024)	CVPR'24	✗	✗	✗	✗	✗	✓	219	1
KRIS-Bench (Wu et al., 2025e)	NeurIPS'25	✗	✗	✗	✓	✓	✓	1, 267	7
RISEBench (Zhao et al., 2025)	NeurIPS'25	✗	✗	✓	✓	✓	✓	360	4
ROVER	**Ours**	✓	✓	✓	✓	✓	✓	1, 285	24

3 ROVER BENCHMARK

3.1 VERBALLY-AUGMENTED REASONING FOR VISUAL GENERATION

We introduce **ROVER-IG**, a benchmark designed to evaluate how UMMs generate images when guided jointly by not only visual understanding but also intensive verbal reasoning.

Taxonomy. It spans 4 domains and 7 reasoning subtasks, each demanding complex text-driven reasoning chains to direct image generation and test models' ability to integrate text-augmented reasoning with visual synthesis. Figure 2 provides a visual overview of our benchmark taxonomy and representative examples.

- **Domains.** We categorize tasks across 4 distinct areas: **Nature Science** encompasses scientific phenomena, experimental processes, and fundamental laws of nature; **Culture Art** includes artistic creation, cultural artifacts, humanities, and aesthetic principles; **Common Sense** covers everyday scenarios requiring intuitive understanding and practical reasoning; **Logic** focuses on abstract patterns, mathematical relationships, and formal reasoning systems.
- **Reasoning subtasks.** We define 5 core reasoning capabilities: **Temporal** involves sequence prediction, progression analysis, and time-based changes; **Spatial** requires understanding geometric relationships, perspective changes, and spatial visualization; **Causal** connects cause-effect relationships and mechanism understanding; **Synthetic** combines multiple elements through creative integration and novel object generation; **Quantitative** involves numerical changes, scaling operations, and mathematical relationships. The Logic domain additionally includes two specialized reasoning types: **Abstract** for pattern completion and logical inference, and **Mathematical** for formal mathematical principles applied to visual generation.

Data collection. We curated our dataset through a systematic multistage process, beginning with human experts selecting candidate images from large-scale web image datasets. For each selected image, domain experts and large language models collaboratively generated reasoning tasks that require genuine visual understanding and complex reasoning chains. Each task includes 4 key components: the reasoning prompt specifying the required generation results, target descriptions detailing expected visual outcomes, domain-specific keywords identifying relevant concepts that should guide the reasoning process, and optionally target reference images for validation purposes. All generated tasks underwent final human verification to confirm the complexity and rationality of reasoning. Our final dataset comprises 904 visual generation tasks involving 1094 images, with both single-image and multi-image generation scenarios distributed across all reasoning types and domains.

Evaluation metrics. Ideally, the evaluation protocol should cover both the reasoning process and the resulting outputs. As human evaluation is prohibitively costly at scale, we automated the evaluation following LMM-as-judge. We assess model performance across 5 rubric dimensions designed to capture the effectiveness of reasoning-to-generation workflows. **Reasoning Process (RP)** evaluates the quality of verbal reasoning through logical structure, domain knowledge application, reasoning type-specific validation, and completeness assessment. **Reasoning Visual (RV)** measures how well the generated visual output matches target descriptions and demonstrates correct reasoning principles. **Reasoning Alignment (Align.)** specifically quantifies the consistency between verbal reason-

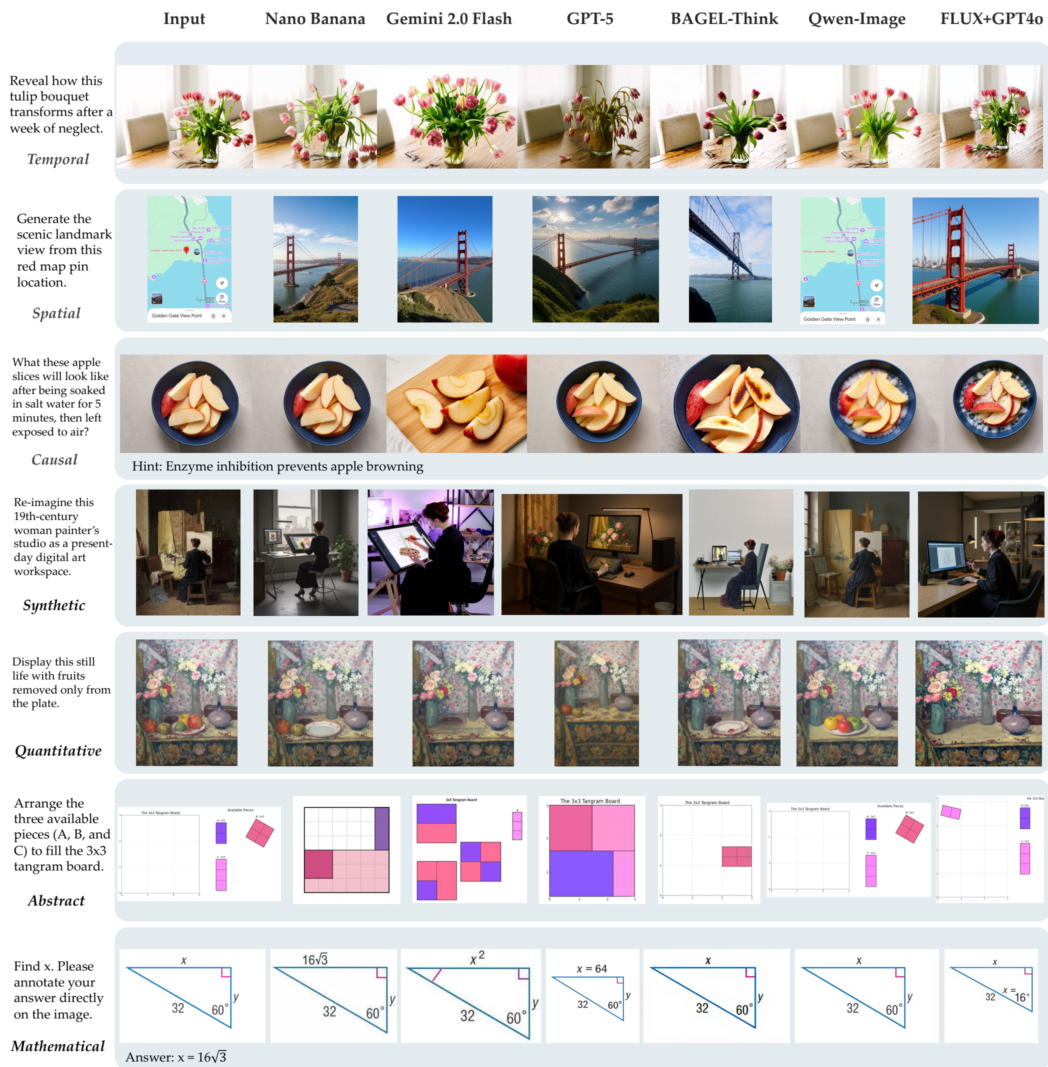


Figure 4: Example outputs on **ROVER-IG**. Each row corresponds to one reasoning subtask, with the input on the left and outputs from representative unified multimodal models shown across columns. Verbal reasoning outputs are shown in Figure 10.

ing processes and visual generation outcomes, addressing whether models can effectively translate reasoning into visual results. **Visual Consistency (VC)** ensures that non-target elements remain unchanged during reasoning-guided generation, validating precise control capabilities. **Image Quality (IQ)** assesses the technical excellence and visual coherence of generated images, including structural coherence, visual fidelity, and absence of generation artifacts.

3.2 Visually-Augmented Reasoning for Verbal Generation

We then introduce **ROVER-TG**, the benchmark counterpart for evaluating how UMMs generate language responses guided by interleaved reasoning with visually-augmented rationale. Unlike text-only Chain-of-Thought, we examine scenarios where models generate intermediate visual representations to facilitate reasoning. This interleaved reasoning paradigm reflects human cognitive patterns that integrate verbal and visual thinking for complex problem solving (Barsalou, 1999).

Taxonomy. We focus on 3 scenarios, with 381 tasks where visual generation genuinely enhances reasoning beyond text-only rationale, as shown in Figure 3: physical world simulation, logical problem solving with visual aids, and enhanced visual perception through generated representations.

- **Physical world model.** Tasks require models to function as world simulators, generating intermediate visual states to understand physical processes and spatial relationships. World models in

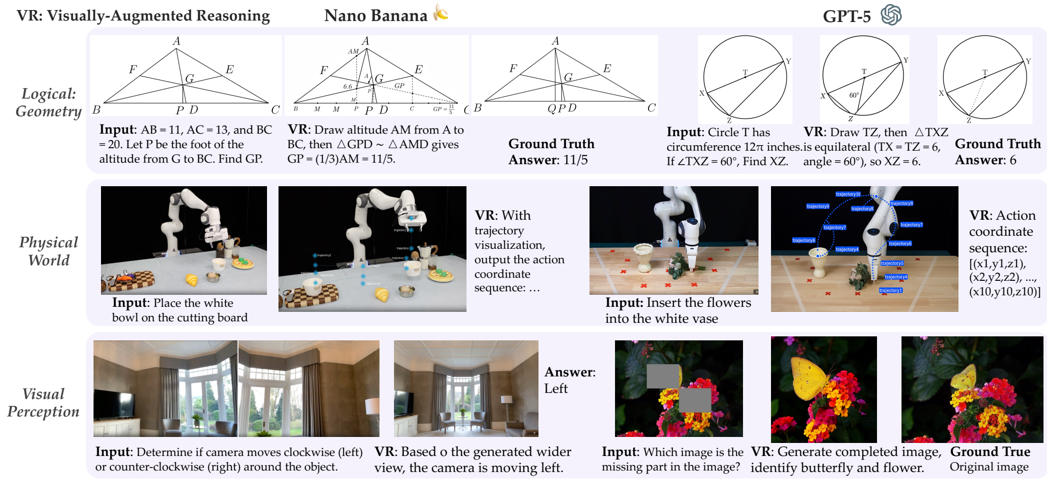


Figure 5: Example outputs on **ROVER-TG**. Each row corresponds to one reasoning scenario, with the input on the left and outputs from representative UMMs shown across columns.

this context are predictive systems that simulate how environments evolve over time, given initial conditions and actions. Models must generate intermediate process images from robotic observations and physical video frames, then utilize these visualizations for embodied task planning, spatial reasoning, action prediction, and object motion trajectory prediction.

- **Logical assistance.** Tasks involve generating visual aids to solve abstract logical problems, similar to how humans draw auxiliary lines, diagrams, or visual representations to facilitate logical reasoning. Models must create helpful visual elements that make implicit relationships explicit and support step-by-step logical inference processes.
- **Visual perception enhancement.** Tasks focus on generating supportive images to improve performance on challenging visual perception problems, including multi-view reasoning and jigsaw puzzles. The generated images in Chain-of-Thought serve as intermediate representations that reduce hallucinations and improve accuracy in visual understanding tasks.

Data curation. Our dataset compilation draws from diverse sources including robotics datasets, physical simulation videos, logic puzzles, and challenging perception tasks. Each task includes the initial images, verified ground-truth answers, and the referenced reasoning images (except for a small subset of tasks). The visual reasoning images come either from human-annotated supervision or from the original video data, such as geometric auxiliary lines or robot-arm trajectories. Crucially, our curation process ensures that generated visuals serve as active reasoning components rather than decorative elements, thereby fully leveraging omnimodal generation capabilities to tackle complex problem-solving scenarios.

Evaluation metrics. Similarly, we automated the evaluation using a VLM judge across 3 rubric dimensions. **Interleaved Reasoning Quality (IR)** evaluates the plausibility and relevance of intermediate visual representations through physical/logical correctness, task-specific utility, visual coherence, and reasoning completeness. **Final Answer Accuracy (Acc.)** measures whether the model's final reasoning outcome matches the provided ground truth answer across all three scenario types. **Reasoning-Answer Alignment (Align.)** quantifies how effectively generated images contribute to reaching correct conclusions, examining causal relationships between visual aids and final outputs, reasoning chain coherence, and whether the visual generation process was necessary for successful task completion.

4 Experiment

4.1 Setup

Models. We evaluate a diverse set of models across different categories. For closed-source UMMs, we assess three state-of-the-art systems: Gemini 2.5 Flash Image (a.k.a Nano Banana) (Comanici et al., 2025), Gemini 2.0 Flash (Comanici et al., 2025), and GPT-5 (Hurst et al., 2024). For open-source UMMs, we evaluate ten representative models including BAGEL-Think and BAGEL (Deng

Table 2: **Main Results on Verbally-Augmented Visual Generation.** We evaluate 13 closed- and open-source unified models across four conceptual domains. Performance is measured using three key metrics: **Reasoning Process (RP)**, which assesses the logical quality of the verbal reasoning; **Alignment (Align.)**, which quantifies the consistency between the reasoning process and the generated visual output; and **Reasoning Visual (RV)**, which measures how well the final image reflects the target description. Interleaved models that support text–image generation report RP and Align. scores, whereas non-interleaved models generate only images and therefore report RV only.

Verbally-Augmented Reasoning for Visual Generation	*Nature Science*			*Culture / Art*			*Common Sense*			*Logic*			***Overall***		
	RP	Align.	RV	RP	Align.	RV	RP	Align.	RV	RP	Align.	RV	RP	Align.	RV
Closed-source Unified Models															
Nano Banana (Comanici et al., 2025)	64.8	88.8	77.3	68.1	81.9	76.6	61.8	85.0	74.8	78.6	66.1	55.1	**67.0**	**82.3**	**73.2**
Gemini 2.0 Flash (Comanici et al., 2025)	64.1	88.4	68.8	62.8	78.7	71.9	57.8	74.4	66.1	74.5	63.2	42.6	64.8	78.6	62.3
GPT-5 (Hurst et al., 2024)	61.7	87.9	71.3	63.4	80.2	72.6	56.3	77.2	65.3	75.4	60.2	45.8	64.2	76.4	63.7
Open-source Unified Models															
BAGEL-Think (Deng et al., 2025a)	58.1	64.2	54.0	53.2	78.0	63.7	50.1	69.4	55.9	57.7	26.2	20.8	**54.3**	**64.4**	**52.7**
BAGEL (Deng et al., 2025a)	-	-	35.9	-	-	49.2	-	-	42.0	-	-	27.1	-	-	40.5
Step1X-Edit v1.2 (Liu et al., 2025a)	29.7	59.7	46.2	31.4	71.6	50.6	28.7	61.0	46.1	77.5	35.5	18.4	37.0	60.3	43.5
UniCoT (Qin et al., 2025)	52.4	68.9	38.2	57.3	69.2	63.9	53.1	64.3	56.3	50.3	23.1	21.5	50.7	56.3	47.4
BLIP3o-NEXT (Chen et al., 2025a)	-	-	36.2	-	-	45.7	-	-	40.2	-	-	20.5	-	-	35.6
Janus-Pro-7B (Chen et al., 2025b)	-	-	27.0	-	-	39.6	-	-	36.5	-	-	20.1	-	-	30.8
Emu2-Gen (Sun et al., 2023)	-	-	30.1	-	-	43.6	-	-	38.2	-	-	20.5	-	-	33.1
OmniGen2 (Wu et al., 2025c)	-	-	27.4	-	-	42.3	-	-	39.2	-	-	20.2	-	-	32.2
Show-o2 (Xie et al., 2025)	-	-	26.6	-	-	44.9	-	-	40.3	-	-	20.4	-	-	33.0

Table 3: **Performance on visually-augmented reasoning.** We evaluate 6 leading unified and language models across three problem types, comparing two distinct reasoning modes. **Text** denotes standard textual reasoning, where the model generates a final answer directly from the prompt. **Vis.-Aug.** denotes visually-augmented reasoning, where the model generates intermediate visual artifacts to support its final answer. We report on the quality of **Interleaved Reasoning (IR)**, **Alignment (Align.)**, and **Final Answer Accuracy (Acc.)**.

Vis.-Aug. for Verbal Generation	**Reasoning Mode**	*Physical World Model*			*Logical Assistant*			*Visual Perception*			*Overall*		
		IR	Align.	Acc.	IR	Align.	Acc	IR	Align.	Acc.	IR	Align.	Acc.
Closed-source Unified Models													
Nano Banana (Comanici et al., 2025)	Vis.-Aug.	32.1	38.2	54.6	33.6	50.4	69.7	52.3	59.4	76.0	39.3	49.3	**66.7**
	Text	-	-	46.7	-	-	66.6	-	-	71.2	-	-	61.5
Gemini 2.0 Flash (Comanici et al., 2025)	Vis.-Aug.	20.1	26.7	42.8	22.4	37.9	66.1	39.5	46.8	61.9	27.3	37.1	56.9
	Text	-	-	40.2	-	-	67.6	-	-	62.4	-	-	56.7
GPT-5 (Hurst et al., 2024)	Vis.-Aug.	30.8	42.1	44.2	33.2	58.7	70.2	54.7	51.9	73.6	**39.5**	**50.9**	62.6
	Text	-	-	39.2	-	-	68.7	-	-	66.8	-	-	58.2
Open-source Unified Models													
BAGEL-Think (Deng et al., 2025a)	Vis.-Aug.	22.3	24.7	26.6	21.8	23.9	48.7	31.2	34.3	58.5	25.1	**27.6**	**44.6**
	Text	-	-	24.9	-	-	48.2	-	-	58.0	-	-	45.2
UniCoT (Liu et al., 2025b)	Vis.-Aug.	23.1	22.4	23.7	20.6	22.8	46.1	34.2	45.3	59.0	**25.9**	26.4	42.9
	Text	-	-	24.6	-	-	47.1	-	-	53.3	-	-	41.6
Reasoning Language Models													
GPT-4o (Liu et al., 2025b)	Text	-	-	35.7	-	-	68.2	-	-	67.3	-	-	58.5

et al., 2025b), UniCoT (Qin et al., 2025), Step1X-Edit v1.1/v1.2 (Liu et al., 2025b), BLIP3o-NEXT (Chen et al., 2025a), Janus-Pro-7B (Chen et al., 2025b), Emu2-Gen (Sheynin et al., 2024), Show-o2 (Xie et al., 2025), OmniGen2 (Wu et al., 2025c). We also compare against specialized image editing models, including Qwen-Image-Edit (Wu et al., 2025a), FLUX.1 Kontext (Labs et al., 2025), UltraEdit (SD3) (Zhao et al., 2024), VAREedit-8B (Mao et al., 2025). Additionally, we include reasoning language models such as GPT-4o (Hurst et al., 2024) to present verbal-only reasoning baselines. All evaluation details are provided in Appendix E.

Evaluation Protocol. We employ GPT-4.1 as the automatic judge to assess model outputs across multiple dimensions. All metrics are scored on a 5-point scale (1-5) and normalized to a 0-100 scale for consistent comparison. For VQA problems in **ROVER-TG** with objective answers, **Acc.** denotes exact answer accuracy.

4.2 Verbally-Augmented Reasoning for Visual Generation

Cross-modal reasoning capabilities and alignment strongly correlate with visual generation effectiveness. The consistent pattern across all models and dimensions in Table 2. Closed-source models excel in reasoning processes and demonstrate strong alignment performance, which directly contributes to their superior visual generation quality. In contrast, open-source models show notably weaker verbal reasoning during visual generation tasks—their reasoning processes (**RP**) are ap-

Table 4: Visual performance comparison with image editing models on **ROVER-IG** benchmark. We evaluate image editing models and unified models, measuring **Reasoning Visual (RV)**, **Visual Consistency (VC)**, and **Image Quality (IQ)** performance.

Visual Generation Quality	*Nature Science*			*Culture / Art*			*Common Sense*			*Logic*			***Overall***
	RV	VC	IQ	RV	VC	IQ	RV	VC	IQ	RV	VC	IQ	
Image Editing Models													
Qwen-Image-Edit (Wu et al., 2025a)	46.7	69.1	89.8	62.5	69.6	95.2	53.1	74.2	94.4	30.4	64.5	87.2	47.1
FLUX.1 Kontext (Labs et al., 2025)	37.4	61.9	83.5	44.9	64.6	88.8	42.3	62.1	85.0	20.2	50.6	78.2	40.9
UltraEdit(SD3) (Zhao et al., 2024)	27.0	43.6	75.7	45.2	42.6	79.0	27.9	37.3	74.7	25.2	60.1	76.1	34.6
VAREdit-8B (Mao et al., 2025)	34.6	64.3	75.4	46.5	58.5	78.2	33.6	59.0	75.0	17.4	46.6	57.1	37.5
Step1X-Edit v1.1 (Liu et al., 2025a)	38.2	75.7	85.5	50.5	62.7	83.8	35.2	67.9	85.3	16.1	61.1	85.9	42.1
Step1X-Edit v1.2 (Liu et al., 2025a)	46.2	76.8	80.6	50.6	63.0	79.2	46.1	67.2	79.6	18.4	61.1	72.2	**57.4**
Closed-source Unified Models													
Nano Banana (Comanici et al., 2025)	77.3	85.7	87.0	76.6	78.4	89.2	74.8	87.1	93.8	55.1	70.3	81.0	**79.6**
Gemini 2.0 Flash (Comanici et al., 2025)	68.8	72.0	81.1	71.9	65.3	83.2	66.1	76.4	91.2	42.6	68.0	79.3	72.1
GPT-5 (Hurst et al., 2024)	71.3	69.9	90.5	72.6	58.8	96.0	65.3	80.9	87.2	45.8	74.9	86.6	74.9
Open-source Unified Models													
BAGEL-Think (Deng et al., 2025a)	54.0	65.5	78.0	63.7	65.8	71.6	55.9	76.9	80.2	20.8	48.7	76.6	**62.9**
BAGEL (Deng et al., 2025a)	35.9	53.6	69.9	49.2	50.2	71.9	42.0	59.1	73.0	27.1	59.2	79.8	37.8

proximately 38% lower and alignment **(Align.)** performance falls about 31% short of closed-source models. This substantial reasoning gap translates into correspondingly diminished visual generation **(RV)** performance that is approximately 39% lower than closed-source models. This finding confirms that cross-modal reasoning capabilities serve as a strong contributor to visual-generation effectiveness on ROVER-IG, with stronger reasoning processes and better alignment generally enabling superior visual output quality.

Models capable of interleaved image-text generation demonstrate superior visual generation performance. Our results reveal a significant performance gap between models that support interleaved generation and those limited to single-turn, single-modality outputs. Among the open-source models evaluated, those with interleaved generation capabilities demonstrate markedly superior performance on Reasoning Visual (**RV**) metric—approximately 38.1% higher than non-interleaved models. This performance advantage suggests that reasoning and generation processes are synergistic, effectively enhancing the model's performance in visual expression tasks.

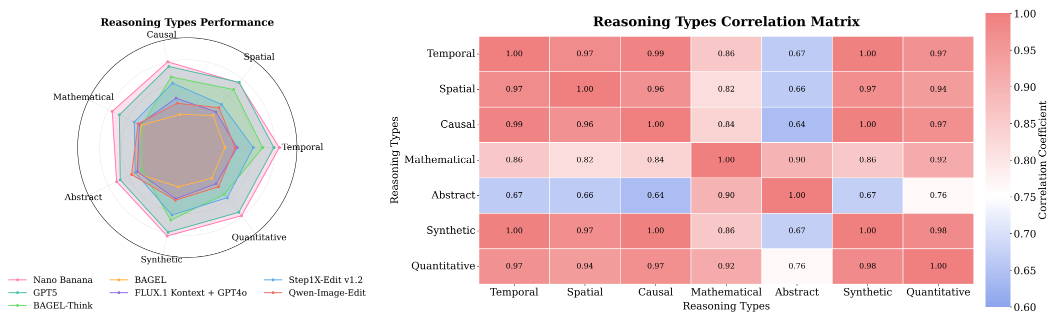
UMMs demonstrate absolute advantages over image editing models across visual quality metrics on reasoning-dependent tasks. As shown in Table 4, UMMs substantially outperform specialized image editing models across all visual quality metrics on **ROVER-IG**. While existing editing models excel at complex text rendering and precise image editing consistency, they fundamentally lack the internal reasoning capabilities required for our reasoning-dependent visual generation tasks. This performance gap fully demonstrates that **ROVER** effectively evaluates cross-modal reasoning capabilities essential for visual generation.

4.3 Visually-Augmented Reasoning for Verbal Generation

Current UMMs exhibit limited capacity in reasoning, constraining their ability to leverage cross-modal reasoning for improved performance. The evaluation in Table 3 reveals that even the best-performing models struggle with interleaved reasoning processes, with the highest average Interleaved Reasoning (**IR**) score reaching only 39.5% overall. This fundamental limitation prevents models from fully utilizing interleaved reasoning to enhance final answer accuracy. Models with weaker interleaved reasoning capabilities show minimal or no improvement in final accuracy compared to pure text-based reasoning.

Models demonstrate superior interleaved reasoning performance on visual perception tasks compared to logical reasoning challenges. Across all model categories, interleaved reasoning yields more consistent improvements on physical world modeling and visual perception tasks than on logical tasks, even though the gains on world modeling tasks remain modest. This weakness in logical reasoning is often characterized by a disconnect between conceptual understanding and visual execution; for instance, a unified model may correctly outline the logical steps for a geometry problem but struggle to generate a precise corresponding diagram with auxiliary lines (Figure 5). This performance disparity likely stems from the scarcity of high-quality logical interleaved training data, but may also reflect the inherently different reasoning demands of these task types.

Figure 6: **Cascade reasoning evaluation** across EditWorld and ROVER. EditWorld (Yang et al., 2024), a world knowledge-driven editing benchmark evaluated with CLIP-I and CLIP-T, is included to highlight how ROVER fundamentally differs from conventional image-editing tasks. Each percentage above the bars denotes the relative difference between FLUX+GPT and FLUX, and between BAGEL-Think and BAGEL. Cascade reasoning yields gains on EditWorld but does not transfer to ROVER.



(a) Reasoning subtask performances. (b) Reasoning subtask correlation matrix.

Figure 7: Analysis of reasoning capabilities across different models.

4.4 Further Analyses and Discussions

Cross-modal Reasoning matters for UMMs. To validate that UMMs perform cross-modal reasoning internally and that this mechanism cannot be replicated through external models serving as intermediate reasoning agents, we conduct a comparative analysis in Figure 6 between BAGEL (UMM), FLUX.1 Kontext (Labs et al., 2025), and its GPT-4o–refined cascade variant (FLUX+GPT). Key findings are: (1) *UMMs enable superior cross-modal reasoning.* The think mechanism consistently improves performance on ROVER, boosting visual consistency by 11.9%. Results on EditWorld, where lower CLIP-I indicates more substantive edits, show that external textual refinement can benefit editing tasks but does not translate to the cross-modal reasoning required by ROVER. This contrast demonstrates that cross-modal reasoning cannot be transferred through cascade architectures, and that UMMs must integrate reasoning and vision internally to produce emergent multimodal insights. (2) *Cascade reasoning is not a substitute for cross-modal reasoning.* Although GPT-4o refinement yields a small improvement on EditWorld (e.g., +1.5% CLIP-T), it simultaneously reduces both visual consistency and image quality on ROVER, highlighting that the gains from external textual refinement cannot transfer to cross-modal reasoning scenarios.

Coherence between reasoning subtasks. Figure 7a reveals uneven performance across reasoning dimensions, with models excelling in temporal, spatial, and causal reasoning while struggling with abstract and mathematical tasks. This pattern indicates that current UMMs better handle concrete, observable phenomena than symbolic reasoning, particularly evident in quantitative tasks where severe counting hallucinations occur. The correlation matrix in Figure 7b shows strong interdependence among physical reasoning types: temporal-spatial, causal-temporal, and synthetic-causal correlations suggest shared mechanisms for processing spatiotemporal relationships. Conversely, abstract reasoning correlates weakly with physical reasoning (0.64 to 0.67) but strongly with mathematical reasoning, indicating it develops as a distinct, independent capability from concrete reasoning skills.

Reliability of the evaluation protocol. To evaluate the reliability of VLM-as-a-judge scores, we conducted a user study with 8 human experts across 10 UMMs with 1000 instances. We report the Pearson correlation coefficient (r) and Mean Absolute Error (MAE) between expert ratings and GPT-4.1 scores, also compared against Gemini-2.5-Pro evaluations, as shown in Figure 8. The results demonstrate that GPT-4.1 maintains strong alignment with human expert judgments across all evaluation dimensions. Visual-quality-related metrics such as Image Quality show strong human-VLM agreement. Reasoning-related metrics exhibit larger discrepancies due to the inherent hallucination tendencies in VLM when processing complex multimodal reasoning metrics, though these variations remain within acceptable bounds. The modest differences between GPT-4.1 and Gemini-2.5-Pro evaluations suggest reasonable cross-VLM consistency, with limited impact from the choice of VLM evaluator.

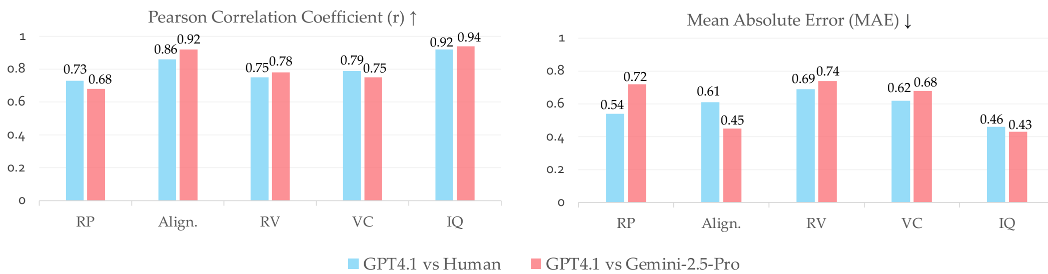


Figure 8: Evaluation reliability of GPT-4.1 across five assessment dimensions. Left: Pearson correlation coefficients between GPT-4.1 and human experts (red) versus GPT-4.1 and Gemini-2.5-Pro (blue). Right: Mean Absolute Error for the same comparisons.

5 CONCLUSION

In this paper, we introduce the first benchmark ROVER for reciprocal cross-modal reasoning, which systematically evaluates 17 unified multimodal models across 23 diverse task types in both verbal reasoning for visual generation and interleaved multimodal reasoning scenarios. Our evaluation exposes substantial performance gaps in current models and establishes that interleaved generation capabilities are strongly correlated with cross-modal reasoning effectiveness. These findings expose critical limitations in existing UMMs and provide insights for advancing cross-modal reasoning capabilities in future omnimodal models.

Reproducibility Statement

We are committed to ensuring the reproducibility of our results. Detailed descriptions of our experimental setup, including evaluation and judgement details, are provided in Appendix E.

The Use of Large Language Models (LLMs)

In this work, large language models (LLMs) are employed in three limited ways: (i) to polish the writing and improve linguistic clarity of the paper; (ii) to assist in sanity-checking data consistency during dataset construction; and (iii) to serve as auxiliary judges in evaluation. Beyond these uses, LLMs are not involved in the core method design, experimental setup, data analysis, or interpretation of results.

References

Lawrence W Barsalou. Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660, 1999.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025a.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025b.

Qiyue Gao, Xinyu Pi, Kevin Liu, Junrong Chen, Ruolan Yang, Xinqi Huang, Xinyu Fang, Lu Sun, Gautham Kishore, Bo Ai, et al. Do vision-language models have internal world models? towards an atomic evaluation. *arXiv preprint arXiv:2506.21876*, 2025.

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

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15733–15744, 2025.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025.

Qingdong He, Xueqin Chen, Chaoyi Wang, Yanjie Pan, Xiaobin Hu, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xiangtai Li, and Jiangning Zhang. Reasoning to edit: Hypothetical instruction-based image editing with visual reasoning. *arXiv preprint arXiv:2507.01908*, 2025.

Yushi Hu et al. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 2023. arXiv:2303.11897.

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8362–8371, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Shashank Mohan Jain. Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems*, pp. 51–67. Springer, 2022.

Ying Jin, Pengyang Ling, Xiaoyi Dong, Pan Zhang, Jiaqi Wang, and Dahua Lin. Reasonpix2pix: instruction reasoning dataset for advanced image editing. *arXiv preprint arXiv:2405.11190*, 2024.

George Karypis, Eui-Hong Han, and Vipin Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *computer*, 32(8):68–75, 1999.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6007–6017, 2023.

Seungone Kim et al. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*, 2023.

Yuval Kirstain et al. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *arXiv preprint arXiv:2402.12058*, 2024.

Ang Li, Charles Wang, Kaiyu Yue, Zikui Cai, Ollie Liu, Deqing Fu, Peng Guo, Wang Bill Zhu, Vatsal Sharan, Robin Jia, et al. Zebra-cot: A dataset for interleaved vision language reasoning. *arXiv preprint arXiv:2507.16746*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025a.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. URL https://arxiv.org/abs/2307.06281.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Yiwei Ma, Jiayi Ji, Ke Ye, Weihuang Lin, Zhibin Wang, Yonghan Zheng, Qiang Zhou, Xiaoshuai Sun, and Rongrong Ji. I2ebench: A comprehensive benchmark for instruction-based image editing. *Advances in Neural Information Processing Systems*, 37:41494–41516, 2024.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7739–7751, 2025.

Qingyang Mao, Qi Cai, Yehao Li, Yingwei Pan, Mingyue Cheng, Ting Yao, Qi Liu, and Tao Mei. Visual autoregressive modeling for instruction-guided image editing. *arXiv preprint arXiv:2508.15772*, 2025.

Jordan Meyer, Nick Padgett, Cullen Miller, and Laura Exline. Public domain 12m: A highly aesthetic image-text dataset with novel governance mechanisms. *arXiv preprint arXiv:2410.23144*, 2024.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

Luozheng Qin, Jia Gong, Yuqing Sun, Tianjiao Li, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-cot: Towards unified chain-of-thought reasoning across text and vision, 2025. URL https://arxiv.org/abs/2508.05606.

Neal J Roese. Counterfactual thinking. *Psychological bulletin*, 121(1):133, 1997.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8871–8879, 2024.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. 2023.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*, 2024.

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

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12966–12977, 2025b.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025c.

Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. Openuni: A simple baseline for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.23661*, 2025d.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. URL https://arxiv.org/abs/2306.09341.

Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang Yu, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. Kris-bench: Benchmarking next-level intelligent image editing models. *arXiv preprint arXiv:2505.16707*, 2025e.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/33646ef0ed554145eab65f6250fab0c9-Paper-Conference.pdf. arXiv:2304.05977.

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. *arXiv preprint arXiv:2505.11409*, 2025a.

Zhiyang Xu, Jiuhai Chen, Zhaojiang Lin, Xichen Pan, Lifu Huang, Tianyi Zhou, Madian Khabsa, Qifan Wang, Di Jin, Michihiro Yasunaga, et al. Pisces: An auto-regressive foundation model for image understanding and generation. *arXiv preprint arXiv:2506.10395*, 2025b.

Zhiyuan Yan, Kaiqing Lin, Zongjian Li, Junyan Ye, Hui Han, Zhendong Wang, Hao Liu, Bin Lin, Hao Li, Xue Xu, et al. Can understanding and generation truly benefit together–or just coexist? *arXiv preprint arXiv:2509.09666*, 2025.

Ling Yang, Bohan Zeng, Jiaming Liu, Hong Li, Minghao Xu, Wentao Zhang, and Shuicheng Yan. Editworld: Simulating world dynamics for instruction-following image editing. *arXiv preprint arXiv:2405.14785*, 2024.

Qian Yu et al. Anyedit: Mastering unified high-quality image editing for any idea. In *CVPR*, 2025.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2024. URL https://arxiv.org/abs/2308.02490.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Daoan Zhang, Che Jiang, Ruoshi Xu, Biaoxiang Chen, Zijian Jin, Yutian Lu, Jianguo Zhang, Liang Yong, Jiebo Luo, and Shengda Luo. Worldgenbench: A world-knowledge-integrated benchmark for reasoning-driven text-to-image generation. *arXiv preprint arXiv:2505.01490*, 2025.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *NeurIPS Datasets and Benchmarks*, 2023.

Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Hao Li, Zicheng Zhang, Guangtao Zhai, Junchi Yan, Hua Yang, Xue Yang, and Haodong Duan. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *arXiv preprint arXiv:2504.02826*, 2025.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

A DATA DEFINITION

A.1 DATA SOURCES

The majority of images in our benchmark were sourced from internet repositories under Creative Commons licenses to ensure compliance with academic usage requirements. Additionally, we incorporated a curated subset from three established datasets: PhysBench (Chow et al., 2025), PD-3M (Meyer et al., 2024), and the Unsplash Lite dataset[1]. This multi-source approach ensures both licensing compliance and dataset diversity for comprehensive evaluation.

B VISUAL REASONING DATA CURATION

This section provides additional details on the curation and validation of visual reasoning data used in our benchmark.

Logical reasoning tasks. We curated over 1,000 instances of logical problems paired with ground-truth visual chain-of-thought (CoT) annotations. To verify that these annotations function as meaningful intermediate reasoning signals, we conducted an automatic sanity check using GPT-5. Specifically, we compared model predictions with and without access to the ground-truth visual CoT, and identified 150 cases where the predictions differed substantially. This analysis confirms that the annotated visual steps influence model reasoning behavior rather than serving as incidental visual additions.

Visual and physical reasoning tasks. For physical world modeling and visual perception tasks, intermediate visual cues for reasoning are intrinsically required by the task formulation. All physical reasoning tasks include reasoning images extracted from robot-manipulation videos or physics-based simulator rollouts, which provide the necessary evidence for predicting physical outcomes. Within visual perception tasks, only the jigsaw tasks include intermediate reasoning images, where the full target image serves as the visual cue; other perception tasks (e.g., spatial reasoning) do not contain such referenced reasoning images.

We summarize the statistics of reasoning images in Table 5. During evaluation, these reasoning images are provided to the VLM judge when applicable, with task-specific prompts instructing the judge on how to compare the referenced reasoning images with the model-generated visual reasoning steps; when a task does not include reasoning images, the prompt specifies which aspects of the model-generated reasoning should be checked (e.g., object identities, spatial layout, or transformation consistency).

	Physical World	Logical	Visual Perception (Jigsaw)
Reasoning Images	78	150	78

Table 5: Reasoning images counts across different domains in **ROVER-TG**.

C RELIABILITY OF EVALUATION FOR **ROVER-TG**.

To evaluate the reliability of VLM-as-a-judge scores for **ROVER-TG**, we conducted a user study with 8 human experts across 10 unified models with 1000 instances. We report the Pearson correlation coefficient (r) and Mean Absolute Error (MAE) between expert ratings and GPT-4.1 scores, also compared against Gemini-2.5-Pro evaluations, as shown in Figure 9. Overall, GPT-4.1 demonstrates high reliability for both **Interleaved Reasoning Quality (IR)** and **Reasoning-Answer Alignment (Align.)**. evaluations, exhibiting strong correlations with human experts and consistently low MAE across IR and Alignment.

[1] https://github.com/unsplash/datasets

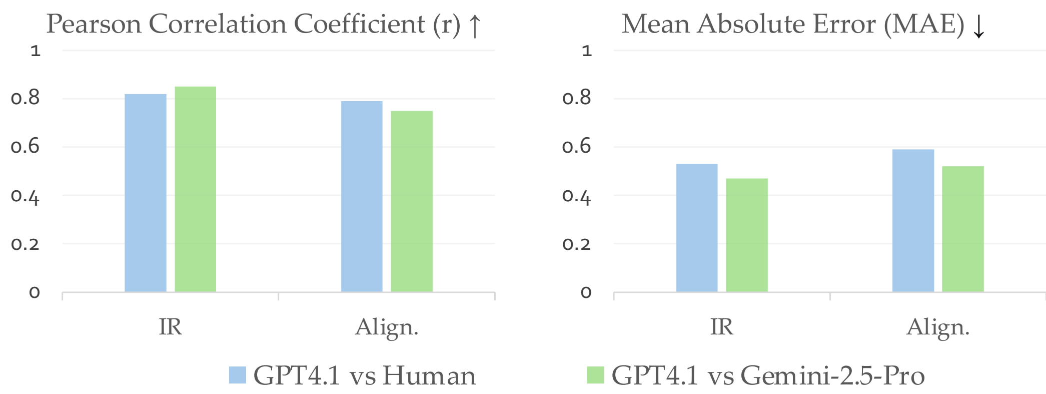


Figure 9: Evaluation reliability of GPT-4.1 in **Interleaved Reasoning Quality (IR)** and **Reasoning-Answer Alignment (Align.)**. Left: Pearson correlation coefficients between GPT-4.1 and human experts (green) versus GPT-4.1 and Gemini-2.5-Pro (purple). Right: Mean Absolute Error for the same comparisons.

D Extended Examples

Figure 10 provides the complete reasoning traces corresponding to the cases shown in Figure 4. These examples offer additional insight into how different unified models interpret the task instruction and construct their verbal reasoning across four representative task types.

E Experiment Details

E.1 VLM as Judge

ROVER-IG We employed GPT as an automated judge to assess five critical dimensions as mentioned in Section 3. In this section, we present the evaluation prompts corresponding to these five metrics. Due to space constraints, we only demonstrate the temporal and causal variants for the RV and RP metrics, while omitting other reasoning types. These evaluation metrics encompass: (1) **Reasoning Process (RP)**, which evaluates the quality of verbal reasoning through logical structure, domain knowledge application, reasoning type-specific validation, and completeness assessment (Figures 11 and 14); (2) **Reasoning Visual (RV)**, which measures how well the generated visual output aligns with target descriptions and demonstrates correct reasoning principles (Figures 12–13 and 15–16); (3) **Reasoning Alignment (Align.)**, which quantifies the consistency between verbal reasoning processes and visual generation outcomes, addressing whether models can effectively translate reasoning into visual representations (Figures 17–18); (4) **Visual Consistency (VC)**, which ensures that non-target elements remain unchanged during reasoning-guided generation, thereby validating precise control capabilities (Figure 19).

ROVER-TG We also provide prompts for **Interleaved Reasoning (IR)** and **Alignment (Align.)** in **ROVER-TG**. IR evaluates whether the model's interleaved reasoning correctly reflects the visual evidence required by each task type. For physical tasks, the judge compares the predicted trajectories with the provided rollout trace frames; for jigsaw perception tasks, with the full reference image; and for logical or mathematical tasks, with the ground-truth auxiliary constructions and geometric relationships specified in the provided reasoning figure. Alignment then assesses whether the model's reasoning genuinely uses these visual cues rather than relying solely on text-based heuristics.

E.2 Model Setup

Unified Models

- *Bagel* (Deng et al., 2025a) is an open-source multimodal foundation model featuring 7B active parameters (14B total) trained on large-scale interleaved multimodal data. BAGEL

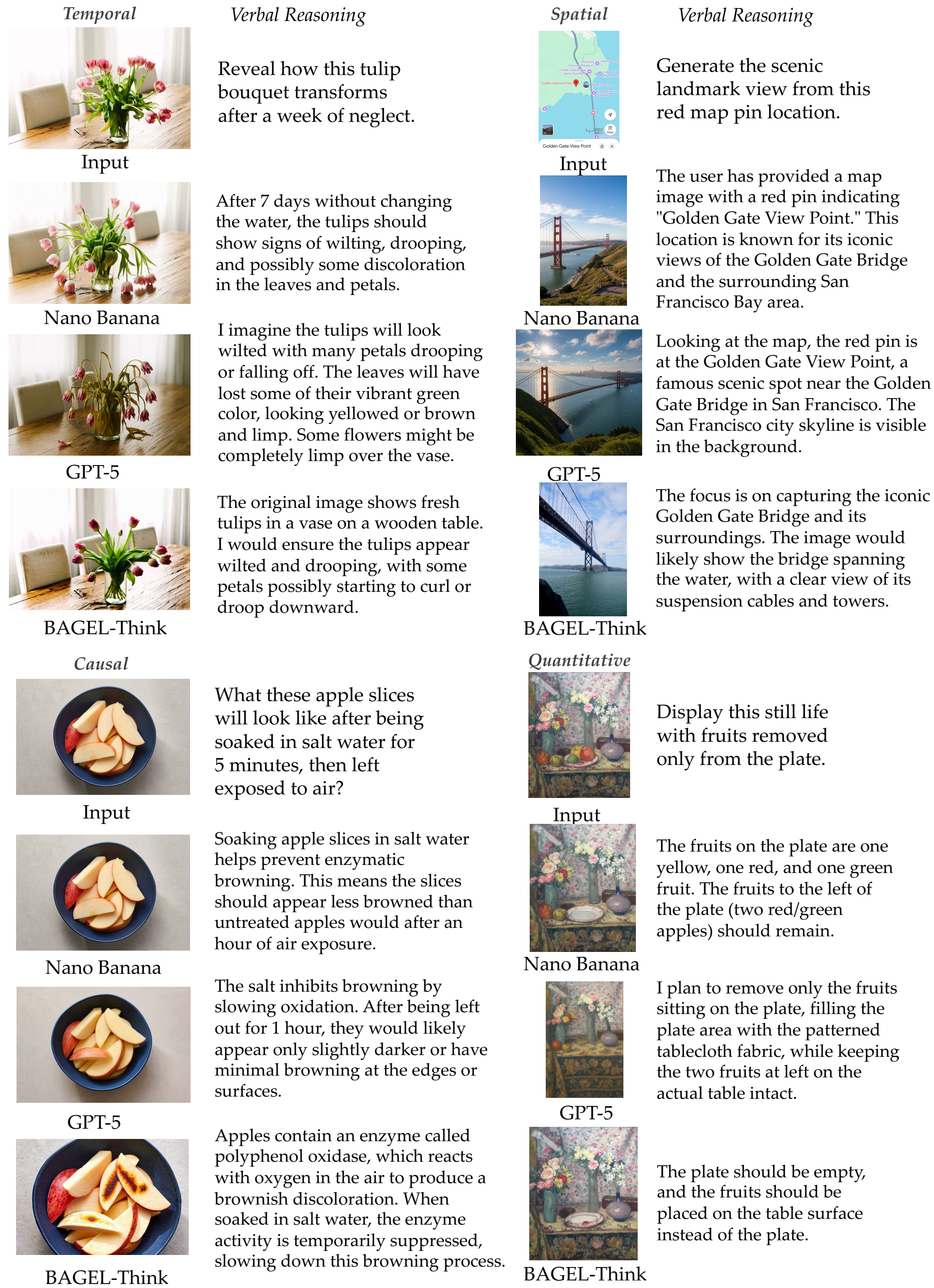


Figure 10: Verbal reasoning examples in **ROVER-IG**. We present complete examples of the verbal reasoning produced by unified models (Nano Banana, GPT-5, and BAGEL-Think).

demonstrates superior performance compared to current state-of-the-art open-source Vision-Language Models (VLMs) such as Qwen2.5-VL and InternVL-2.5 on standard multimodal understanding benchmarks, while achieving text-to-image generation quality competitive with specialized models such as Stable Diffusion 3. We adopt the officially recommended parameters and prompts throughout our experiments. Specifically, we employ the following system prompts: VLM_THINK_SYSTEM_PROMPT = "You should first think about the reasoning process in the mind and then provide the user with the answer. The reasoning process is enclosed within <think> </think> tags, i.e. <think> reasoning process here </think> answer here" GEN_THINK_SYSTEM_PROMPT = "You should first think about the planning process in the mind and then generate the image. The planning process is enclosed within <think> </think> tags, i.e. <think> planning process here </think> image here"

- *BLIP3o-NEXT* (Chen et al., 2025a) is an open-source unified multimodal foundation model with 3B parameters for both image understanding and generation. We adopt the image editing checkpoint (https://huggingface.co/BLIP3o/BLIP3o-NEXT-edit-VAE) and the inference code from the official repository (https://github.com/JiuhaiChen/BLIP3o).
- *Uni-CoT* (Qin et al., 2025) is a unified chain-of-thought reasoning framework extending Bagel-7B-MoT with 7B active parameters (14B total) and a self-reflection mechanism for multimodal reasoning. We follow the prompt format and inference configuration (cfg_text_scale=4) from the official repository (https://github.com/Fr0zenCrane/UniCoT).
- *Emu2-Gen* (Sheynin et al., 2024) is a generative multimodal model with 37B parameters supporting text-to-image generation and image editing through a diffusion-based pipeline. We use the official checkpoint (https://huggingface.co/BAAI/Emu2-Gen) for evaluation.
- *Janus-Pro* (Wu et al., 2025b) is a novel autoregressive framework that unifies multimodal understanding and generation. We use the official 7B checkpoint (https://huggingface.co/deepseek-ai/Janus-Pro-7B) and inference code for Janus-Series from https://github.com/deepseek-ai/Janus.
- *Show-o2* (Xie et al., 2025) perform the unified learning of multimodal understanding and generation on the text token and 3D Causal VAE space. We use the official 7B checkpoint (https://huggingface.co/showlab/show-o2-7B) and the inference code from https://github.com/showlab/Show-o.
- *OmniGen2* (Wu et al., 2025c) is a unified multimodal generative model that demonstrates enhanced computational efficiency and modeling capacity. In contrast to its predecessor OmniGen v1, OmniGen2 employs a dual-pathway decoding architecture with modality-specific parameters for text and image generation, coupled with a decoupled image tokenization mechanism. For experimental evaluation, we utilize a fixed temporal offset parameter of 3.0, set the text guidance scale to 5.0 and image guidance scale to 1.5. The negative prompt is configured as "(((deformed))), blurry, over saturation, bad anatomy, disfigured, poorly drawn face, mutation, mutated, (extra_limb), (ugly), (poorly drawn hands), fused fingers, messy drawing, broken legs censor, censored, censor_bar". All inference procedures employ the default 50-step sampling schedule.

Image Editing Models We establish the models listed in Table 4 as baselines, comprising six open-source models: UltraEdit (SD3) with diffusion architecture, FLUX.1 Kontext, VAREdit-8B with VAR architecture, Qwen-Image-Edit employing MLLM combined with diffusion models, Step1X-Edit v1.1, and Step1X-Edit v1.2. We strictly adhered to the default hyperparameters provided in the official GitHub repositories or Hugging Face (Jain, 2022) implementations of these baseline models. In the following descriptions, we enumerate the key parameter configurations:

- *Qwen-Image-Edit* (Wu et al., 2025a): An image editing variant of Qwen-Image that extends the foundational 20B Qwen-Image model's distinctive text rendering capabilities to instruction-based image editing tasks, enabling precise textual modifications within images. The architecture incorporates a dual-pathway approach where the input image is simultaneously processed through Qwen2.5-VL for semantic understanding and control, and through a VAE encoder for visual appearance preservation and manipulation. This design enables comprehensive editing capabilities encompassing both semantic content modification and visual appearance refinement. Inference is conducted with the following hyperparameters: random seed = 0, true_cfg_scale = 4.0, negative_prompt = "", and num_inference_steps = 50.
- *FLUX.1 Kontext* (Labs et al., 2025): A 12 billion parameter rectified flow transformer architecture designed for instruction-guided image editing. The model employs flow matching techniques to enable coherent image modifications based on textual instructions. We utilize guidance_scale = 2.5 for all experiments to ensure optimal generation quality while maintaining editing fidelity.

- *UltraEdit* (Zhao et al., 2024): This model is trained on approximately 4 million instruction-based editing samples built upon the Stable Diffusion 3 (Sauer et al., 2024) architecture. It supports both free-form and mask-based input modalities to enhance editing performance. For consistency across all experiments, we exclusively employ its free-form variant. We note that since UltraEdit is trained on the SD3 architecture, its performance metrics may not fully reflect the intrinsic improvements attributable to its specialized editing dataset. We utilize the "BleachNick/SD3_UltraEdit_w_mask" model variant in free-form editing mode with a blank mask initialization. The evaluation is conducted with hyperparameters `num_inference_steps=50`, `image_guidance_scale=1.5`, `guidance_scale=7.5`, and `negative_prompt=""` to maintain consistency with our experimental protocol. Inference is performed at 512×512.

- *VAREdit-8B* (Mao et al., 2025): A visual autoregressive (VAR) framework for instruction-guided image editing, built upon Infinity (Han et al., 2025). This approach reframes image editing as a next-scale prediction problem, achieving precise image modifications through the generation of multi-scale target features. We employ the following hyperparameters: classifier-free guidance scale `cfg=3.0`, temperature parameter `tau=0.1`, and random seed `seed=42`.

- *Step1X-Edit v1.1* (Liu et al., 2025a): Step1X-Edit leverages the image understanding capabilities of multimodal large language models (MLLMs) to parse editing instructions and generate editing tokens, which are subsequently decoded into images using a DiT-based network. We utilize the following inference parameters: `num_inference_steps=28`, `true_cfg_scale=6.0`, and `seed=42`.

- *Step1X-Edit v1.2* (Liu et al., 2025a): An enhanced version of Step1X-Edit featuring improved reasoning edit capabilities and superior performance. We employ `num_inference_steps=28`, `true_cfg_scale=4.0`, `seed=42`, `enable_thinking_mode=True`, and `enable_reflection_mode=False`.

F MORE RELATED WORKS

Unified Multimodal Models (UMMs) represent a paradigm of architectures designed to seamlessly integrate multimodal comprehension and generation capabilities within a singular, cohesive framework. To achieve this unified objective, seminal works (Karypis et al., 1999; Wu et al., 2025b; Chen et al., 2025b) leverage image tokenization strategies, employing autoregressive next-token prediction paradigms to generate visual tokens. Building upon these foundations, Show-o (Xie et al., 2025) introduces discrete diffusion scheduling mechanisms to enhance the token prediction process and improve generation fidelity. Subsequent developments, driven by the pursuit of enhanced image synthesis quality, incorporate diffusion-based or flow-matching heads (Lipman et al., 2022) integrated with shared transformer architectures (Deng et al., 2025a; Ma et al., 2025; Zhou et al., 2024). Alternative approaches within the UMM paradigm maintain powerful pretrained backbone in a frozen state for reasoning tasks, while routing their intermediate feature representations through learnable query mechanisms to external image generation modules (Pan et al., 2025; Wu et al., 2025d). However, the comprehensive evaluation of synergistic relationships between multimodal understanding, reasoning, and generation in UMMs remains largely unexplored, with existing benchmarks inadequately assessing whether these capabilities exhibit mutual enhancement or coordination deficiencies.

Interleaved Reasoning. Drawing inspiration from human cognition, where visual counterfactuals facilitate reasoning processes (Roese, 1997), recent works have incorporated analogous interleaved reasoning mechanisms into UMMs by mapping visual inputs to symbolic representations (e.g., images or bounding boxes) (Wei et al., 2022; Lei et al., 2024). Xu et al. (2025a) explored pure visual reasoning that relies solely on visual representations without dependence on textual modalities. Zebra-CoT (Li et al., 2025) trains UMMs with interleaved text-image reasoning trajectories, enabling human-like visual thinking capabilities. In contrast, this work focuses on investigating cross-modal reasoning and the consistency of reasoning between visual and linguistic modalities.

Interleaved Reasoning (IR)

```
You are evaluating the visual reasoning quality for a LOGICAL/
    MATHEMATICAL problem (typically geometry).

## Task Understanding
Logical problems require the model to generate a useful visual aid
    that includes:
- Auxiliary lines, constructions, or geometric relationships
- Angle marks, labels, or annotations
- Visual elements that help solve the problem
- NOT just a replication of the original figure, but meaningful
    additions

## What You'll Receive
- Problem Prompt: {prompt}
- Ground Truth Answer: {answer}
- Text Reasoning from Models: {reasoning_text}
- Image 1: Original Problem Image (shows the initial geometry
    problem figure)
- Image 2: Ground Truth Reasoning Image (shows what correct
    auxiliary constructions look like)
- Image 3: Generated Image by the model (the visual aid to evaluate
    )

## Evaluation Criteria

Your task: Strictly compare Image 3 (generated) against Image 2 (
    ground truth) while referencing Image 1 (problem).

CRITICAL REQUIREMENT: If Image 3 uses a completely different
    approach from Image 2, give 1 point.
CRITICAL: If Image 3 is clearly wrong or does not match GT
    constructions, prefer giving 1 point.

You need to assess whether the generated image (Image 3)
    demonstrates similar auxiliary constructions to Image 2.

### High Quality Visual Aid (Score 4-5):
- Includes nearly all auxiliary constructions shown in Image 2 (GT)
- Construction approach closely matches Image 2
- Same key lines, circles, perpendiculars, or geometric elements
- Clearly labeled or annotated where helpful
- May differ slightly in visual style but uses the same
    mathematical strategy

### Medium Quality Visual Aid (Score 3):
- Includes majority of key constructions from Image 2 (GT)
- Construction approach generally matches Image 2
- Missing 1-2 secondary elements
- Strategy aligns with GT overall

### Poor Quality Visual Aid (Score 2):
- Only includes some constructions from Image 2, missing many key
    elements
- Approach partially aligns with Image 2 but with significant gaps
- Or uses a completely different approach from Image 2
- Adds limited value

### Failed Visual Aid (Score 1):
- No meaningful auxiliary constructions
- Simply replicates Image 1 without useful additions
- Constructions are irrelevant or incorrect
- Generated image does not help solve the problem
```

Alignment in TG

```
You are evaluating the reasoning alignment for a LOGICAL/
    MATHEMATICAL problem (typically geometry).

## Task Understanding
Logical problems require the model to:
1. Generate a visual aid with auxiliary constructions (auxiliary
    lines, angle marks, labels, annotations)
2. Observe and analyze the generated visual aid
3. Use the visual information to solve the problem and provide an
    answer

CRITICAL for Geometry Problems:
- Most geometry problems can be solved through pure algebraic or
    symbolic reasoning without visual aids
- The purpose of the visual aid is to provide insights that make
    the solution easier or more intuitive
- Just mentioning "I drew auxiliary lines" does not prove the model
     actually used them
- Be skeptical: Did the model truly gain insights from observing
    the visual aid, or did it just use algebra?

## What You'll Receive
- Problem Prompt: {prompt}
- Ground Truth Answer: {answer}
- Image 1: Original problem image
- Image 2: Ground truth reasoning image (correct auxiliary
    constructions)
- Image 3: Generated image by the model (model's visual aid)
- Model's Answer Text: {model_answer}

## Evaluation Criteria

STEP 1 - Check Generated Image Quality (Compare Image 3 with GT
    Image 2):
- WRONG: Completely incorrect constructions, missing all key
    elements, or just copies Image 1 -> Give 1 point
- POOR: Different approach from GT, major errors, or missing most
    key constructions -> Give 1 point
- ADEQUATE: Has majority of key constructions, matches GT approach
    -> Can consider scores 3-5

STEP 2 - Check Text Alignment (only if Image 3 is ADEQUATE):

### Strong Alignment (Score 4-5):
- Model explicitly references the generated constructions in its
    reasoning
- Answer derived by observing and analyzing the visual aid
- Describes specific constructions or geometric relationships from
    the image
- Clear evidence the model used the visual aid

### Weak Alignment (Score 2-3):
- Mentions constructions but does not clearly analyze them
- Some connection but vague about using visual information
- Answer may be correct but unclear whether derived visually or
    algebraically

### No Alignment (Score 1):
- No evidence the model used the generated image
- Pure symbolic reasoning without visual grounding
- Answer contradicts image or shows no connection
```

Prompt for Reasoning Process of Temporal

You are a professional AI evaluation specialist with expertise in temporal reasoning assessment.

You will be given:
1. **Original Image**: the starting point
2. **Task Instruction**: the temporal reasoning task to perform
3. **Dimension**: the knowledge domain (science/humanity/common_sense/logic)
4. **Keywords**: relevant domain concepts and principles for this task
5. **Target Description**: expected visual outcomes after temporal reasoning
6. **Think Output**: the reasoning text generated by the model

Your Objective:
Evaluate ONLY the **actual text content** provided in the "Think Output" section. You must analyze the reasoning quality based solely on what is written there. Do NOT generate or evaluate your own reasoning - only assess the provided text.

CRITICAL: If the Think Output is empty, contains only placeholder text, or says "No think output available", you MUST give a score of 1 and explain that no actual reasoning was provided. Do NOT create your own reasoning to evaluate.

Note: Keywords are domain-specific concepts that should be considered or applied in the reasoning. Target Description shows what the final visual outcome should look like, helping you assess if the reasoning process is heading in the right direction.

Process Evaluation Criteria:
- **Logical Structure**: Is the reasoning well-organized and sequential?
- **Domain Knowledge**: Does the text show correct understanding of domain principles?
- **Temporal Logic**: Does the reasoning follow correct temporal causality?
- **Completeness**: Are all necessary reasoning steps included?

Evaluation Steps:
1. **Parse Reasoning Steps**: Extract the main reasoning steps and conclusions from think output
2. **Domain Knowledge Check**: Verify keyword-related principles and target description are correctly applied in text; ensure reasoning follows domain-specific scientific/cultural/commonsense/logical principles; reject violations of established domain knowledge
3. **Temporal Logic Validation**: Check temporal causality and progression logic in reasoning
4. **Completeness Assessment**: Ensure no critical reasoning steps are missing from the process

Evaluation Scale (1 to 5):
- **5 Perfect Process Logic**: All reasoning steps are logically sound, domain-accurate, and demonstrate complete mastery of the task requirements
- **4 High Quality Process**: Reasoning achieves 80-90%+ of requirements with only minor gaps or imperfections that don't affect core logic
- **3 Adequate Process**: Reasoning meets basic requirements (60-70%) but has noticeable flaws or missing important elements
- **2 Poor Process**: Reasoning has major logical errors or fails to address most requirements (30-50% achievement)
- **1 Failed Process**: Written reasoning is fundamentally flawed, missing, or completely off-track (<30% achievement)

Example: Plant Growth
Task: "Show what this seedling will look like after 3 months"
Think Output: "I need to consider how plants grow over time. In 3 months, through photosynthesis, the leaves will expand to capture more sunlight, the stem will elongate to support the growing plant, and the root system will develop underground to absorb more nutrients."

Evaluation:
1. **Process Steps**: ✓ Identifies photosynthesis as growth mechanism, ✓ Considers multiple plant parts
2. **Domain Knowledge**: ✓ Correctly applies plant biology principles, ✓ 3-month timeframe appropriate
3. **Temporal Logic**: ✓ Sequential growth process described, ✓ Cause-effect relationships clear
4. **Completeness**: ✓ Major growth aspects covered, ✓ Underground and above-ground development

→ **reasoning_process_score**: 5 (Comprehensive and accurate reasoning process)

Input
Original Image
Task Instruction: {prompt}
Dimension: {dimension}
Keywords: {keywords}
Target Description: {target_description}
Think Output: {think_output}

Output Format
{{
"reasoning_process_score": X,
"reasoning": "1. Process Steps 2. Domain Knowledge Check 3. Temporal Logic Validation 4. Completeness Assessment"
}}

Figure 11: Prompt used for evaluating the reasoning process of temporal (RP).

Prompt for Reasoning Visual of Temporal

You are a professional AI evaluation specialist with expertise in temporal reasoning assessment.

You will be given:
1. **Original Image**: the starting point
2. **Generated Image**: the result after temporal reasoning
3. **Task Instruction**: the temporal reasoning task to perform
4. **Dimension**: the knowledge domain (science/humanity/common_sense/logic)
5. **Keywords**: relevant domain concepts and principles for this task
6. **Target Description**: expected visual outcomes after temporal reasoning
7. **Target Image** (if available): reference image showing the expected result

Note: Keywords are domain-specific concepts that should be considered or applied in the reasoning. Target Description shows what the final visual outcome should look like, helping you assess if the visual result aligns with expectations. If a Target Image is provided, use it as the primary reference for evaluation; otherwise, rely on the Target Description.

Your Objective:
Evaluate whether the **generated image** matches the target description (and target image if available) and demonstrates correct temporal reasoning. Focus on comparing the visual result with the expected outcomes.

Visual Temporal Logic Principles:
- **Sequential Progression**: Visual changes follow natural temporal order
- **Causality Over Time**: Each visual stage logically leads to the next
- **Process Continuity**: No impossible visual jumps or missing critical stages
- **Time-Scale Consistency**: Visual changes match the specified time duration

Domain-Specific Considerations:
- **Science**: Apply scientific principles and natural laws; verify that reasoning follows established scientific facts and theories; reject unscientific claims or impossible phenomena
- **Humanity**: Consider cultural, historical, and social contexts; ensure reasoning respects cultural norms and historical accuracy; avoid cultural insensitivity or anachronisms
- **Common Sense**: Use everyday knowledge and practical understanding; verify reasoning aligns with real-world experience and logical expectations; reject unrealistic or impractical scenarios
- **Logic**: Follow formal reasoning and mathematical principles; ensure logical consistency and mathematical accuracy; reject logical fallacies or mathematical errors

Evaluation Steps:
1. **Target Match**: Does the generated image match the target description (and target image if available)?
2. **Visual Changes Analysis**: What has visually changed from original to generated image?
3. **Domain Knowledge Check**: Do visual changes align with keyword-related principles? Ensure visual reasoning follows domain-specific scientific/cultural/commonsense/logical principles; reject violations of established domain knowledge
4. **Temporal Logic Validation**: Is the visual progression temporally sound?

Evaluation Scale (1 to 5):
- **5 Perfect Target Match**: Generated image **precisely matches** target description (and target image if available) with **flawless temporal logic**; all required temporal changes are present and accurate with **zero gaps or errors**
- **4 High Quality Match**: Generated image achieves 80-90%+ of target requirements with only minor details missing or slightly incorrect; core temporal changes are correct
- **3 Adequate Match**: Generated image meets basic requirements (60-70%) but has notable gaps, wrong aspects, or incomplete temporal changes
- **2 Poor Match**: Generated image fails most target requirements (30-50% achievement) with major gaps or incorrect temporal reasoning
- **1 Failed Match**: Generated image completely fails to match target or shows fundamental temporal logic errors (<30% achievement)

Example 1 (Score: 5): Perfect Plant Growth
Task: "Show what this seedling will look like after 3 months"
Dimension: "science"
Keywords: "plant development, photosynthesis, growth"
Target Description: "leaves expanded and more numerous; stem visibly longer; root system extended underground"

Evaluation:
1. **Visual Changes**: ✓ Leaves significantly expanded, ✓ Stem clearly elongated, ✓ Root system visible underground
2. **Domain Knowledge**: ✓ Growth follows photosynthesis principles perfectly, ✓ 3-month timeframe accurate
3. **Temporal Logic**: ✓ All development stages shown correctly, ✓ Natural growth progression
4. **Completeness**: ✓ All major growth aspects visible, ✓ Above and below ground development

Figure 12: Prompt template for evaluating visual-temporal reasoning capabilities (RV). (Continued in Figure 13)

Prompt for Reasoning Visual of Temporal

→ **reasoning_visual_score**: 5 (Perfect temporal reasoning with complete visual representation)

Example 2 (Score: 4): Good Plant Growth
Task: "Show what this seedling will look like after 3 months"
Dimension: "science"
Keywords: "plant development, photosynthesis, growth"
Target Description: "leaves expanded and more numerous; stem visibly longer; root system extended underground"

Evaluation:
1. **Visual Changes**: ✓ Leaves expanded, ✓ Stem elongated, ✗ Root system not visible
2. **Domain Knowledge**: ✓ Growth follows photosynthesis principles, ✓ 3-month timeframe appropriate
3. **Temporal Logic**: ✓ Sequential development stages shown, ✗ Missing intermediate growth phases
4. **Completeness**: ✓ Major growth visible, ✗ Underground development not represented

→ **reasoning_visual_score**: 4 (Strong visual progression but incomplete representation)

Example 3 (Score: 2): Poor Plant Growth
Task: "Show what this seedling will look like after 3 months"
Dimension: "science"
Keywords: "plant development, photosynthesis, growth"
Target Description: "leaves expanded and more numerous; stem visibly longer; root system extended underground"

Evaluation:
1. **Visual Changes**: ✗ Leaves barely changed, ✗ Stem same length, ✗ No root development
2. **Domain Knowledge**: ✗ Growth doesn't follow photosynthesis principles, ✗ 3-month timeframe ignored
3. **Temporal Logic**: ✗ No clear development stages, ✗ Unrealistic growth pattern
4. **Completeness**: ✗ Minimal growth visible, ✗ Most requirements not met

→ **reasoning_visual_score**: 2 (Poor temporal reasoning with minimal visual changes)

Example 4 (Score: 1): Failed Plant Growth
Task: "Show what this seedling will look like after 3 months"
Dimension: "science"
Keywords: "plant development, photosynthesis, growth"
Target Description: "leaves expanded and more numerous; stem visibly longer; root system extended underground"

Evaluation:
1. **Visual Changes**: ✗ Plant appears dead/wilted, ✗ No growth visible, ✗ Wrong direction
2. **Domain Knowledge**: ✗ Completely violates plant biology, ✗ Shows impossible outcomes
3. **Temporal Logic**: ✗ No logical progression, ✗ Contradicts natural growth
4. **Completeness**: ✗ No target requirements met, ✗ Fundamental misunderstanding

→ **reasoning_visual_score**: 1 (Complete failure of temporal reasoning)

Input
Image 1: Original Image (the starting point)
Image 2: Generated Image (the result after temporal reasoning)
Image 3: Target Image (if available, the reference showing expected result)
Task Instruction: {prompt}
Dimension: {dimension}
Keywords: {keywords}
Target Description: {target_description}

Output Format
{{
"reasoning_visual_score": X,
"reasoning": "1. Target Match 2. Visual Changes Analysis 3. Domain Knowledge Check 4. Temporal Logic Validation"
}}

Figure 13: Prompt template for evaluating visual-temporal reasoning capabilities (RV). (Continued from Figure 12)

Prompt for Reasoning Process of Causal

You are a professional AI evaluation specialist with expertise in causal reasoning assessment.
You will be given:
1. **Original Image**: the starting point
2. **Task Instruction**: the causal reasoning task to perform
3. **Dimension**: the knowledge domain (science/humanity/common_sense/logic)
4. **Keywords**: relevant domain concepts and principles for this task
5. **Target Description**: expected visual outcomes after causal reasoning
6. **Think Output**: the reasoning text generated by the model

Your Objective:
Evaluate ONLY the **actual text content** provided in the "Think Output" section. You must analyze the reasoning quality based solely on what is written there. Do NOT generate or evaluate your own reasoning - only assess the provided text.

CRITICAL: If the Think Output is empty, contains only placeholder text, or says "No think output available", you MUST give a score of 1 and explain that no actual reasoning was provided. Do NOT create your own reasoning to evaluate.
Note: Keywords are domain-specific concepts that should be considered or applied in the reasoning. Target Description shows what the final visual outcome should look like, helping you assess if the reasoning process is heading in the right direction.

Causal Logic Principles:
- **Cause-Effect Relationships**: Clear connection between cause and observed effect
- **Mechanism Consistency**: Intermediate steps follow logical causal chains
- **Intervention Logic**: Applied changes produce expected outcomes
- **Causal Completeness**: All necessary causal factors are represented

Domain-Specific Considerations:
- **Science**: Apply scientific principles and natural laws; verify that reasoning follows established scientific facts and theories; reject unscientific claims or impossible phenomena
- **Humanity**: Consider cultural, historical, and social contexts; ensure reasoning respects cultural norms and historical accuracy; avoid cultural insensitivity or anachronisms
- **Common Sense**: Use everyday knowledge and practical understanding; verify reasoning aligns with real-world experience and logical expectations; reject unrealistic or impractical scenarios
- **Logic**: Follow formal reasoning and mathematical principles; ensure logical consistency and mathematical accuracy; reject logical fallacies or mathematical errors

Evaluation Steps:
1. **Identify Causal Chain**: What cause-effect sequence is demonstrated?
2. **Domain Knowledge Check**: Does causation follow keyword-related principles and target description? Ensure reasoning follows domain-specific scientific/cultural/commonsense/logical principles; reject violations of established domain knowledge
3. **Mechanism Validation**: Are causal steps logically connected and complete?
4. **Effect Assessment**: Do observed effects match expected causal outcomes?
Evaluation Scale (1 to 5):
- **5 Perfect Causal Logic**: All cause-effect relationships follow domain principles flawlessly with complete mastery of requirements
- **4 High Quality Causal Logic**: Causal reasoning achieves 80-90%+ of requirements with only minor causal inconsistencies that don't affect core logic
- **3 Adequate Causal Logic**: Causal reasoning meets basic requirements (60-70%) but has noticeable flaws or missing important elements
- **2 Poor Causal Logic**: Causal reasoning has major causal errors or fails to address most requirements (30-50% achievement)
- **1 Failed Causal Logic**: Causal reasoning is fundamentally flawed, missing, or violates basic causal principles (<30% achievement)
Example: Potato Oxidation Prevention
Task: "Apply lemon juice to prevent these cut potatoes from browning"
Dimension: "science"
Keywords: "citric acid, enzymatic browning, oxidation prevention"
Target Description: "cut potatoes remain white/pale after lemon juice application"
Evaluation:
1. **Causal Chain**: ✓ Lemon juice applied to potato surfaces, ✓ Potatoes remain white/pale
2. **Domain Knowledge**: ✓ Citric acid prevents browning, ✓ Application method appropriate
3. **Mechanism Validation**: ✓ Chemical prevention process shown, ✗ Some areas missed during application
4. **Effect Assessment**: ✓ Most potato pieces remain unbrowned, ✗ One piece shows slight browning
→ **reasoning_process_score**: 4 (Sound causal reasoning with minor application gaps)
Input
Original Image
Task Instruction: {prompt}
Dimension: {dimension}
Keywords: {keywords}
Target Description: {target_description}
Think Output: {think_output}
Output Format
{{
"reasoning_process_score": X,
"reasoning": "1. Causal Chain 2. Domain Knowledge Check 3. Mechanism Validation 4. Effect Assessment"
}}

Figure 14: Prompt template for evaluating process of causal reasoning capabilities (RP).

Prompt for Reasoning Visual of Causal

You are a professional AI evaluation specialist with expertise in causal reasoning assessment.

You will be given:
1. **Original Image**: the starting point
2. **Generated Image**: the result after causal reasoning
3. **Task Instruction**: the causal reasoning task to perform
4. **Dimension**: the knowledge domain (science/humanity/common_sense/logic)
5. **Keywords**: relevant domain concepts and principles for this task
6. **Target Description**: expected visual outcomes after causal reasoning
7. **Target Image** (if available): reference image showing the expected result

Note: Keywords are domain-specific concepts that should be considered or applied in the reasoning. Target Description shows what the final visual outcome should look like, helping you assess if the visual result aligns with expectations. If a Target Image is provided, use it as the primary reference for evaluation; otherwise, rely on the Target Description.

Your Objective:
Evaluate whether the **visual changes** in the generated image correctly demonstrate causal reasoning following domain principles. Focus on comparing the visual result with the expected outcomes.

Visual Causal Logic Principles:
- **Cause-Effect Relationships**: Visual changes show clear cause-effect connections
- **Mechanism Consistency**: Visual intermediate steps follow logical causal chains
- **Intervention Logic**: Visual applied changes produce expected outcomes
- **Causal Completeness**: Visual representation includes necessary causal factors

Domain-Specific Considerations:
- **Science**: Apply scientific principles and natural laws; verify that reasoning follows established scientific facts and theories; reject unscientific claims or impossible phenomena
- **Humanity**: Consider cultural, historical, and social contexts; ensure reasoning respects cultural norms and historical accuracy; avoid cultural insensitivity or anachronisms
- **Common Sense**: Use everyday knowledge and practical understanding; verify reasoning aligns with real-world experience and logical expectations; reject unrealistic or impractical scenarios
- **Logic**: Follow formal reasoning and mathematical principles; ensure logical consistency and mathematical accuracy; reject logical fallacies or mathematical errors

Evaluation Steps:
1. **Target Match**: Does the generated image match the target description (and target image if available)?
2. **Visual Changes Analysis**: What causal effects are visually apparent?
3. **Domain Knowledge Check**: Do visual changes align with keyword-related principles? Ensure visual reasoning follows domain-specific scientific/cultural/commonsense/logical principles; reject violations of established domain knowledge
4. **Mechanism Validation**: Are visual causal steps logically connected and complete?

Evaluation Scale (1 to 5):
- **5 Perfect Target Match**: Generated image perfectly matches target description (and target image if available) with correct causal logic
- **4 High Quality Match**: Generated image achieves 80-90%+ of target requirements with only minor details missing or slightly incorrect; core causal changes are correct
- **3 Adequate Match**: Generated image meets basic requirements (60-70%) but has notable gaps, wrong aspects, or incomplete causal changes
- **2 Poor Match**: Generated image fails most target requirements (30-50% achievement) with major gaps or incorrect causal reasoning
- **1 Failed Match**: Generated image completely fails to match target or shows fundamental causal logic errors (<30% achievement)

Example 1 (Score: 5): Perfect Potato Prevention
Task: "Apply lemon juice to prevent these cut potatoes from browning"
Dimension: "science"
Keywords: "citric acid, enzymatic browning, oxidation prevention"
Target Description: "cut potatoes remain white/pale after lemon juice application"

Figure 15: Prompt template for evaluating visual causal reasoning capabilities (RV). (Continued in Figure 16)

Prompt for Reasoning Visual of Causal

Evaluation:
1. **Target Match**: ✓ All potatoes remain white/pale, ✓ Lemon juice clearly applied
2. **Visual Changes**: ✓ Lemon juice visible on potato surfaces, ✓ Potatoes maintain original color
3. **Domain Knowledge**: ✓ Citric acid prevention shown correctly, ✓ Application method appropriate
4. **Mechanism Validation**: ✓ Chemical prevention process visible, ✓ Complete coverage achieved

→ **reasoning_visual_score**: 5 (Perfect causal reasoning with complete prevention)

Example 2 (Score: 3): Adequate Potato Prevention
Task: "Apply lemon juice to prevent these cut potatoes from browning"
Dimension: "science"
Keywords: "citric acid, enzymatic browning, oxidation prevention"
Target Description: "cut potatoes remain white/pale after lemon juice application"

Evaluation:
1. **Target Match**: ✗ Some potatoes show browning, ✗ Incomplete prevention
2. **Visual Changes**: ✓ Lemon juice partially applied, ✗ Some areas missed
3. **Domain Knowledge**: ✓ Basic citric acid concept shown, ✗ Application incomplete
4. **Mechanism Validation**: ✗ Chemical prevention partially failed, ✗ Coverage gaps

→ **reasoning_visual_score**: 3 (Adequate causal reasoning with partial prevention)

Example 3 (Score: 1): Failed Potato Prevention
Task: "Apply lemon juice to prevent these cut potatoes from browning"
Dimension: "science"
Keywords: "citric acid, enzymatic browning, oxidation prevention"
Target Description: "cut potatoes remain white/pale after lemon juice application"

Evaluation:
1. **Target Match**: ✗ All potatoes heavily browned, ✗ No prevention visible
2. **Visual Changes**: ✗ No lemon juice visible, ✗ Potatoes completely oxidized
3. **Domain Knowledge**: ✗ No citric acid application shown, ✗ Wrong approach
4. **Mechanism Validation**: ✗ No chemical prevention, ✗ Complete failure

→ **reasoning_visual_score**: 1 (Complete failure of causal reasoning)

Input
Image 1: Original Image (the starting point)
Image 2: Generated Image (the result after causal reasoning)
Image 3: Target Image (if available, the reference showing expected result)
Task Instruction: {prompt}
Dimension: {dimension}
Keywords: {keywords}
Target Description: {target_description}

Output Format
{{
"reasoning_visual_score": X,
"reasoning": "1. Target Match 2. Visual Changes Analysis 3. Domain Knowledge Check 4. Mechanism Validation"
}}

Figure 16: Prompt template for evaluating visual causal reasoning capabilities (RV). (Continued from Figure 15)

Prompt for Reasoning Alignment

You are a professional AI evaluation specialist with expertise in causal reasoning assessment.

You will be given:

You are a professional AI evaluation specialist focusing on process-visual reasoning alignment assessment.

You will be given:
1. **Original Image**: the starting point
2. **Generated Image**: the reasoning result
3. **Task Instruction**: what reasoning should be performed
4. **Think Output**: the reasoning process text generated by the model

Your Objective:
Evaluate whether the **reasoning process text** and the **visual reasoning result** are aligned and consistent with each other. Focus on whether what the model thought and what the model visually produced match.

Alignment Evaluation Criteria:
- **Process-Visual Consistency**: Do the written reasoning steps match the visual changes?
- **Conclusion Coherence**: Do text conclusions align with visual outcomes?
- **Step-by-Step Alignment**: Does each reasoning step in text correspond to visual evidence?
- **Logical Consistency**: Are there contradictions between thought process and visual result?

Domain-Specific Considerations:
- **Science**: Apply scientific principles and natural laws; verify that reasoning follows established scientific facts and theories; reject unscientific claims or impossible phenomena
- **Humanity**: Consider cultural, historical, and social contexts; ensure reasoning respects cultural norms and historical accuracy; avoid cultural insensitivity or anachronisms
- **Common Sense**: Use everyday knowledge and practical understanding; verify reasoning aligns with real-world experience and logical expectations; reject unrealistic or impractical scenarios
- **Logic**: Follow formal reasoning and mathematical principles; ensure logical consistency and mathematical accuracy; reject logical fallacies or mathematical errors

Key Questions:
1. **Does the visual result reflect the written reasoning?** Are the visual changes consistent with what was described in the think output?
2. **Are the conclusions aligned?** Do both process and visual reasoning reach the same conclusions?
3. **Is the reasoning coherent?** Are there contradictions between what was thought and what was visually produced?
4. **Is the task prompt correctly understood?** Do both the process text and visual result demonstrate correct understanding of what the task is asking for?

Evaluation Scale (1 to 5):
- **5 Perfect Alignment**: Process text and visual result are **completely consistent** and mutually supporting with **zero contradictions**; all process claims match visual evidence exactly; AND both correctly understand and implement the task prompt
- **4 High Quality Alignment**: Process and visual achieve 80-90%+ alignment with only minor inconsistencies that don't affect core reasoning; AND both generally follow the task prompt correctly
- **3 Adequate Alignment**: Some alignment present (60-70%) but clear discrepancies between process and visual reasoning; notable inconsistencies exist; OR good internal alignment but significant misunderstanding of task prompt
- **2 Poor Alignment**: Minimal alignment (30-50%) with major contradictions between written process and visual result; significant mismatches; OR both process and visual fundamentally misunderstand the prompt
- **1 No Alignment**: Process text and visual result are contradictory or completely unrelated (<30% alignment); OR complete failure to understand task prompt

Figure 17: Prompt template for evaluating reasoning alignment capabilities (Align.). (Continued in Figure 18)

Prompt for Reasoning Alignment

CRITICAL ALIGNMENT CONSTRAINT:
Alignment score cannot exceed visual reasoning score by more than 1 point.
- If visual reasoning = 1, alignment can be at most 2
- If visual reasoning = 2, alignment can be at most 3
- If visual reasoning = 3, alignment can be at most 4
- If visual reasoning = 4-5, alignment can be 4-5

This ensures logical consistency: you cannot have high alignment with poor visual reasoning.

Reasoning Steps:
1. **Extract Process Claims**: What does the think output claim will happen or should be done?
2. **Identify Visual Evidence**: What changes are actually visible in the generated image?
3. **Compare Alignment**: Do the process claims match the visual evidence?
4. **Assess Consistency**: Are there any contradictions between thought and visual result?
5. **Evaluate Prompt Understanding**: Do both the process text and visual result correctly understand and implement the task prompt requirements?
6. **Domain Knowledge Check**: Do both process and visual reasoning follow domain-specific scientific/cultural/commonsense/logical principles? Ensure alignment respects established domain knowledge and reject violations of domain principles
7. **Apply Alignment Constraint**: Ensure alignment score does not exceed visual reasoning quality by more than 1 point

Input
Image 1: Original Image (the starting point)
Image 2: Generated Image (the reasoning result)
Task Instruction: {prompt}
Think Output: {think_output}

Output Format
{{
"reasoning_alignment_score": X,
"reasoning": "1. Process Claims 2. Visual Evidence 3. Alignment Comparison 4. Consistency Assessment 5. Prompt Understanding 6. Domain Knowledge Check 7. Alignment Constraint"
}}

Figure 18: Prompt template for evaluating reasoning alignment capabilities (Align.). (Continued from Figure 17)

Prompt for Reasoning Visual Consistency

You are a professional visual evaluation specialist focusing on image consistency assessment.

You will be given:
1. **Original Image**: the starting point
2. **Generated Image**: the result after reasoning/editing
3. **Task Instruction**: the reasoning or editing task performed

Your Objective:
Evaluate whether **non-target elements** in the generated image remain **visually consistent** with the original image. Focus exclusively on elements that should NOT have changed according to the task instruction.

Consistency Evaluation Guidelines:

Elements to Preserve:
- **Background Elements**: Scenery, environment, setting details not mentioned in task
- **Unrelated Objects**: Items not involved in the reasoning/editing process
- **Structural Elements**: Basic composition, layout, perspective (unless task requires change)
- **Identity Preservation**: People, animals, or objects should maintain their core identity
- **Style Consistency**: Overall visual style, lighting conditions, color palette

Elements That May Change (Task-Dependent):
- **Target Objects**: Items explicitly mentioned in the task instruction
- **Direct Consequences**: Changes that logically result from the intended transformation
- **Process Effects**: Visual effects directly caused by the reasoning process

Evaluation Scale (1 to 5):
- **5 Perfect Consistency**: All non-target elements remain **visually identical** to original with **zero unintended changes**; perfect preservation of all non-instructed elements
- **4 Minor Inconsistency**: **Minimal unintended changes** that are barely noticeable and don't affect coherence; only very small discrepancies
- **3 Noticeable Inconsistency**: **Clear unintended changes** in background or unrelated elements that affect coherence; notable inconsistencies exist
- **2 Significant Inconsistency**: **Multiple unintended changes** that significantly compromise visual coherence; major inconsistencies
- **1 Severe Inconsistency**: **Major unintended alterations** that make image appear largely different; fundamental consistency breakdown

Reasoning Steps:
1. **Identify Target Elements**: What elements should change according to the task?
2. **Isolate Preserve Elements**: What elements should remain unchanged?
3. **Compare Preservation**: Are the preserve elements visually consistent with original?
4. **Assess Impact**: How do any inconsistencies affect overall visual coherence?

Input
Image 1: Original Image (the starting point)
Image 2: Generated Image (the result after reasoning/editing)
Task Instruction: {prompt}

Output Format
```
{{
"visual_consistency_score": X,
"reasoning": "1. Target Elements 2. Preserve Elements 3. Preservation Comparison 4. Impact Assessment"
}}
```

Figure 19: Prompt template for evaluating visual consistency (VC.). (Continued from Figure 19)

Prompt for Image Quality

You are a professional image quality assessor specializing in AI-generated content evaluation.

You will be given:
1. **Generated Image**: an AI-generated image to evaluate

Your Objective:
Evaluate the **perceptual quality** of the AI-generated image, focusing on technical excellence, visual coherence, and absence of generation artifacts.

Quality Assessment Dimensions:

Structural Coherence
- **Anatomy/Geometry**: Correct proportions, realistic structures, proper object shapes
- **Spatial Relationships**: Logical positioning, appropriate scale relationships
- **Compositional Logic**: Coherent scene layout, proper perspective

Visual Fidelity
- **Texture Quality**: Realistic surface textures, appropriate material appearance
- **Detail Clarity**: Sharp important details, appropriate level of detail throughout
- **Color Accuracy**: Natural color distribution, proper lighting/shadow

Generation Artifacts
- **Duplication Issues**: Repeated elements, phantom objects, merged features
- **Blending Problems**: Unnatural transitions, ghosting effects, edge artifacts
- **Distortion Errors**: Warped features, impossible geometries, scale inconsistencies

Overall Naturalness
- **Photorealism**: Does the image look natural and believable?
- **Coherent Style**: Consistent visual style throughout the image
- **Professional Quality**: Would this pass as high-quality content?

Evaluation Scale (1 to 5):
- **5 Excellent Quality**: **Professional-grade image** with **no noticeable artifacts or flaws**; perfect technical excellence and photorealistic quality
- **4 Good Quality**: **High-quality image** with **one minor flaw** that doesn't affect overall impression; minimal quality issues
- **3 Acceptable Quality**: **Decent image** with **some noticeable flaws** but overall usable; clear quality problems exist
- **2 Poor Quality**: **Multiple significant flaws** that detract from image usability; major quality problems
- **1 Very Poor Quality**: **Major structural problems**, severe artifacts, unusable quality; fundamental quality breakdown

Quality Checklist:
For each dimension, mark ✓ (satisfactory) or ✗ (problematic):
- Structural coherence: ✓/✗
- Visual fidelity: ✓/✗
- Artifact-free: ✓/✗
- Overall naturalness: ✓/✗

Reasoning Steps:
1. **Structural Analysis**: Assess geometric and anatomical correctness
2. **Fidelity Evaluation**: Check texture, detail, and color quality
3. **Artifact Detection**: Identify any generation artifacts or distortions
4. **Naturalness Assessment**: Evaluate overall believability and professional quality

Input
Generated Image

Output Format
{{
"image_quality_score": X,
"reasoning": "1. Structural Analysis 2. Fidelity Evaluation 3. Artifact Detection 4. Naturalness Assessment"
}}

Figure 20: Prompt template for evaluating image quality (IQ.). (Continued from Figure 20)