Envisioning Beyond the Pixels: Benchmarking Reasoning-Informed Visual Editing

Xiangyu Zhao[1,2*], **Peiyuan Zhang**[3*], **Kexian Tang**[2,4*], **Xiaorong Zhu**[1*], **Hao Li**[2], **Wenhao Chai**[5], **Zicheng Zhang**[1,2], **Renqiu Xia**[1,2], **Guangtao Zhai**[1,2], **Junchi Yan**[1], **Hua Yang**[1✉], **Xue Yang**[1✉†], **Haodong Duan**[2✉†]

[1] ICISEE & SAIS & SAI, Shanghai Jiao Tong University [2] Shanghai AI Laboratory

[3] Wuhan University [4] Tsinghua University [5] Princeton University

[*] *Equal contribution* [✉] *Corresponding author* [†] *Project lead*

Figure 1: **Examples of leading models on the Reasoning-Informed viSual Editing(RISE) benchmark.** RISEBench contains complex and various tasks that pose challenges to current models.

Abstract

Large Multi-modality Models (LMMs) have made significant progress in visual understanding and generation, but they still face challenges in General Visual Editing, particularly in following complex instructions, preserving appearance consistency, and supporting flexible input formats. To study this gap, we introduce **RISEBench**, the first benchmark for evaluating **R**easoning-**I**nformed vi**S**ual **E**diting (**RISE**). RISEBench focuses on four key reasoning categories: *Temporal*, *Causal*, *Spatial*, and *Logical Reasoning*. We curate high-quality test cases for each category and propose an robust evaluation framework that assesses *Instruction Reasoning*, *Appearance Consistency*, and *Visual Plausibility* with both human judges and the LMM-as-a-judge approach. We conducted experiments evaluating nine prominent visual editing models, comprising both open-source and proprietary models. The evaluation results demonstrate that current models face significant challenges in reasoning-based editing tasks. Even the most powerful model evaluated, GPT-image-1, achieves an accuracy of merely 28.8%. RISEBench effectively highlights the limitations of contemporary editing models, provides valuable insights, and indicates potential future directions for the field of reasoning-aware visual editing. Our code and data have been released at `https://github.com/PhoenixZ810/RISEBench`.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

1 Introduction

Large Multi-Modality Models (LMMs) have achieved remarkable progress in both visual understanding [22, 1, 5, 38, 26] and visual generation [29, 31, 3]. Meanwhile, significant efforts [37, 48, 40, 42, 20] have been dedicated to unifying these two tasks, with the goal of enhancing overall performance through joint learning. Some open-source models have demonstrated decent capability in either visual understanding or image generation; however, they still exhibit substantial limitations in General Visual Editing (*i.e.*, transforming an input image based on textual instructions). Specifically, current open-source methods struggle with: (1) accurately following complex editing instructions [34]; (2) preserving the original image's appearance during visual editing [18]; and (3) accommodating flexible input formats [40, 48] (*e.g.*, supporting both single and multiple images with natural language instructions). These limitations severely hinder their practical utility, making them hardly worth rigorous evaluation in this task.

Recently, we observed that proprietary models such as GPT-image-1 [16] and Gemini-2.0-Flash* [38] have made significant advancements over open-source counterparts (Fig. 1). Notably, these models exhibit a remarkable capability in **R**easoning-**I**nformed vi**S**ual **E**diting (**RISE**) – a sophisticated ability that enables models to make intelligent visual modifications based on contextual understanding and logical reasoning. This advanced ability has exciting implications for various real-world applications, such as context-aware image modification (*e.g.*, adjusting lighting to match a scene's time of day), intelligent object insertion or removal with semantic consistency, and content adaptation based on inferred user intent. However, traditional image editing models [17, 11, 4] that do not incorporate multi-modal reasoning lack these capabilities entirely. While such a phenomenon is promising, we found that there is no well-established benchmark for systematically evaluating RISE task, making it difficult to quantitatively assess and further study this ability in existing models.

To this end, we introduce **RISEBench**, a focused, small-scale benchmark specifically designed to evaluate reasoning-informed visual editing (RISE) capabilities. In this benchmark, we identify and categorize key image editing challenges that require four fundamental types of reasoning: **temporal reasoning**, **causal reasoning**, **spatial reasoning**, and **logical reasoning**. To ensure a comprehensive evaluation, we manually curated a diverse set of high-quality test cases across the four categories: 85 for temporal reasoning, 90 for causal reasoning, 100 for spatial reasoning, and 85 for logical reasoning, resulting in a total of **360 carefully human-annotated samples**.

For evaluation, we decompose the quality of the edited output images into three key dimensions: **instruction reasoning**, **appearance consistency**, and **generation plausibility**. Evaluations are conducted using both human judges and an LMM-as-a-judge framework. For the latter, a rigorous pipeline was developed to ensure the reliability and validity of the LMM's assessments. Additionally, we performed extensive experiments to quantify the correlation between the scores produced by the LMM and human experts, which verifies the reliability and effectiveness of our proposed framework.

Using RISEBench, we conduct a systematic evaluation of state-of-the-art LMMs with visual editing capabilities. Our results reveal that open-source visual editing models such as BAGEL [8], Step1X-Edit [23], FLUX [19], EMU2 [35], and OmniGen [46] show limited reasoning capabilities, resulting in notably low performance across most test cases. Proprietary models, such as Gemini-2.0-Flash Series [38] and GPT-image-1, achieve significantly better overall performance. Notably, GPT-image-1 displays strong capabilities across temporal, causal, and spatial reasoning tasks. However, it still struggles with logical reasoning, highlighting an area for future research.

In summary, our main contributions are as follows:

1. We propose the first dedicated benchmark for assessing **R**easoning-**I**nformed vi**S**ual **E**diting (RISE), establishing a foundation for systematic assessment in this emerging area.
2. We define core categories of RISE challenges, design meaningful evaluation dimensions, and present a robust, effective LMM-as-a-judge framework for scalable and automated assessment.
3. We conduct a comprehensive evaluation and analysis of 8 prominent visual editing models, offering novel insights into their reasoning-driven visual editing capabilities and highlighting areas for future improvement.

*The Gemini-2.0-Flash Series comprises two models: Gemini-2.0-Flash-Experimental-Image-Generation (Gemini-2.0-Flash-Exp) and Gemini-2.0-Flash-Preview-Image-Generation (Gemini-2.0-Flash-Pre).

2 Related Work

Image Editing with Diffusion Models. Editing images based on textual user instructions is a crucial task in the field of image generation. With the advancement of large-scale diffusion models, the performance of image editing tasks has significantly improved. For instance, some methods [7, 25, 50, 33], adopt training-free approaches to guide denoising according to editing instructions, such as reversing noise on an image and guiding denoising with text [27], controlling attention maps during diffusion steps [11], or blending the original and generated images [6]. Recently, other works [4, 51, 49] have shifted to training-based methods, where pre-trained text-to-image diffusion models are further fine-tuned using datasets comprising paired edited images to enhance editing capabilities, yielding superior performance. However, due to the limited fine-grained semantic understanding of diffusion models, image editing models based on diffusion are often insufficient for handling complex, fine-grained editing instructions that require higher-order reasoning, thereby restricting their application in more diverse scenarios.

Unified Large Multi-Modality Models. Large Multi-Modality Models (LMMs) extend the input and output capabilities of large language models (LLMs) by incorporating visual information. Early works primarily focused on visual understanding, which involves processing visual inputs, reasoning, and generating textual outputs. Recently, a series of studies [34, 53, 9, 37, 40, 42, 45, 52, 48, 21] have aimed to develop unified LMMs capable of simultaneously handling both textual and visual inputs, enabling cross-modal generation and understanding. Initial approaches [34, 53, 9] often relied on pre-trained diffusion decoders to generate outputs by regressing CLIP [30] image representations. To further integrate understanding and generation, recent models such as Chameleon [37], Emu3 [40], and SynerGen-VL [20] have adopted a unified next-token prediction paradigm by discretizing images. Transfusion [52] and Show-o [48] demonstrated that bidirectional image diffusion could be integrated with autoregressive text prediction within a single framework.

Text-to-Image Generation Evaluation. The comprehensive evaluation of text-to-image generation is a long-standing problem. Early work mainly adopt the Fréchet inception distance (FID) [13] metric to mesure the distance between the generated distribution and the target distribution. However, this cannot measure the per-image alignment of image and the instruction. To better measure the semantic alignment in text-to-image generation, a series of works [10, 15, 12, 44, 43] propose metrics based on foundation models such as CLIP or object detectors. However, few works have focused on the reasoning-based visual editing. Recent work [28] shares the similar motivation of measuring models' world knowledge. However, they do not explicitly measure the models' reasoning

3 RISEBENCH-360

Humans possess a deep, conceptual understanding of objects and scenes in the real world that goes far beyond superficial attributes such as color and shape. For example, people can effortlessly reason about: 1) Temporal evolution of objects (**temporal reasoning**), such as fruits rotting over time, iron tools rusting, or children growing into adults; 2) Transformative changes due to external factors (**causal reasoning**), like ice cream melting under sunlight or vehicles becoming damaged after collisions; 3) Spatial configurations (**spatial reasoning**), including how shapes appear from different viewing angles and how various components assemble into complete structures. 4) Additionally, people can easily solve visual puzzle problems (**logical reasoning**) such as tic-tac-toe or mazes, and concretely imagine their solutions.

However, these capabilities present significant challenges for most generative models, which struggle to incorporate such reasoning into their visual outputs. To objectively assess current models' performance on these tasks and clearly identify their limitations, we propose RISEBench, the first benchmark specifically designed to evaluate reasoning-informed visual editing capabilities of image generative models across these dimensions of human-like visual understanding.

3.1 Benchmark Construction

Among the broad spectrum of visual editing tasks, RISEBench targets four major problem categories that require both deep visual understanding and precise reasoning, termed as *Temporal*, *Causal*, *Spatial*, and *Logical Reasoning*. For each category, we curate a diverse set of high-quality, carefully

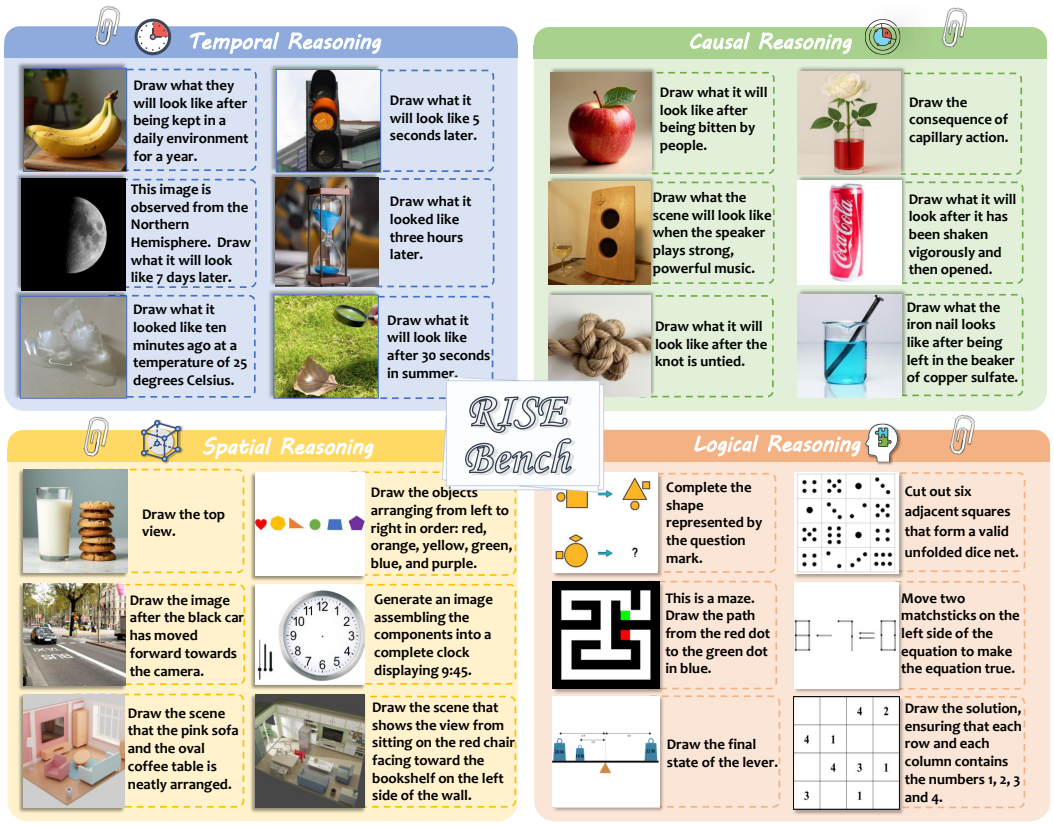

Figure 2: **Overview of RISEBench.** We present illustrative example questions from each of the four problem categories, each demanding profound image understanding and reasoning capabilities.

designed test cases. Each instance comprises an input image and an instruction prompt, illustrating reasoning-driven image transformations (see Fig. 2). The distribution of tasks is presented in Fig. 3.

Temporal Reasoning. Temporal reasoning tasks evaluate a model's ability to understand and anticipate the evolution of objects or scenes over time. Beyond recognizing static attributes such as color, shape, or size, a reasoning-capable generative model should capture how these properties change through natural temporal progression. To construct such tasks systematically, we define several key elements of temporal change, including scale, direction, and object. Based on representative combinations of these dimensions, we derive four subcategories reflecting common temporal phenomena: *life progression*, *environmental cycles*, *material state change*, and *societal transformation*. The subcategories span diverse scenarios of temporal change, enabling the design of tasks that demand fine-grained understanding of temporal dynamics and assess a model's ability to perform temporally grounded reasoning beyond superficial image manipulation.

Causal Reasoning. Causal reasoning is essential for evaluating a generative model's ability to capture real-world interaction dynamics. Unlike temporal reasoning, which concerns natural progression over time, causal reasoning involves understanding how external forces or events directly induce changes in an object's state. This domain includes a range of phenomena: *1) Structural Deformation*, where external forces alter an object's shape; *2) State Transition*, such as phase changes (*e.g.*, freeze) triggered by manipulation or environmental shifts; *3) Chemical & Biological Transformations*, involving changes at the molecular or biological level; *4) Physical Manifestations*, where observable effects result from underlying physical laws activated by specific stimuli. These tasks require models to exhibit implicit knowledge of material properties, physical principles, and typical cause-effect relationships.

Spatial Reasoning. Spatial reasoning tasks assess a model's ability to understand, manipulate, and generate images that preserve accurate spatial relationships among objects in a scene. This requires internalizing geometric principles, structural coherence, 3D reasoning, and perspective —- core components of human-like visual understanding. We define five representative subcategories: *1)*

Component Assembly tests whether disjoint parts can be combined into a coherent whole, requiring spatial and structural integration; *2) Object Arrangement* evaluates the sequencing and positioning of objects based on attributes such as size, shape, or color; *3) Viewpoint Generation* assesses the ability to synthesize novel views from different angles, relying on latent 3D representations;

4) Structural reasoning challenges the model to complete occluded or fragmented objects by inferring missing parts; *5) Layout reasoning* examines understanding and manipulation of spatial configurations within a scene. Together, these tasks provide a comprehensive testbed for evaluating a model's spatial intelligence and its capacity for structure-aware, visually grounded generation.

Figure 3: **Task Distribution of RISEBench.** RISEBench contains four main reasoning categories: *Temporal*, *Causal*, *Spatial*, and *Logical*. Each category includes various subtasks, facilitating a comprehensive evaluation.

Logical Reasoning. Unlike other categories that focus on physical or commonsense understanding in natural images, logical reasoning tasks evaluate a model's ability to perform structured, rule-based inference grounded in visual input. These tasks require interpreting visual elements and systematically applying formal rules — an area where current generative models still struggle. To assess this capability, we curate a diverse set of puzzles and logical challenges across three primary subtasks: *1) Puzzle Solving*, including classic visual problems such as Sudoku, mazes, and Tic-Tac-Toe; *2) Mathematical Derivation*, involving tasks requiring computation, such as shortest path finding and formula-based reasoning; *3) Pattern Prediction*, where the model must infer and complete visual patterns based on implicit rules. This category offers a broad spectrum of logic-based tasks with varying abstraction and difficulty, providing a rigorous evaluation of a model's visual-symbolic reasoning and its ability to link perception with inference.

3.2 Evaluation Pipeline

Evaluating the quality of reasoning-informed visual editing remains a challenging task. To address this, we first establish detailed scoring guidelines and conduct comprehensive human evaluations along three key dimensions: **1. Instruction Reasoning**, assessing whether the model correctly interprets and follows the editing instruction; **2. Appearance Consistency**, evaluating preservation of relevant visual attributes from the original image; **3. Visual Plausibility**, determining whether the output is coherent, realistic, and physically or logically plausible within context. Since human evaluation is resource-intensive and difficult to scale. To overcome these limitations, we further adopt an LMM-as-a-Judge approach. Given their strong visual understanding and reasoning abilities, state-of-the-art LMMs offer a promising alternative for automatic and human-aligned evaluation. We develop a robust evaluation pipeline (Fig. 4) leveraging these models to produce scalable assessments. In the following part, we detail each evaluation dimension:

Dimension 1: Instruction Reasoning. This dimension assesses the model's ability to accurately understand and execute the given instruction, with particular attention to both explicit directives and implicit requirements embedded within the prompt. A high-quality response not only performs the literal task specified but also captures the underlying reasoning or intended visual effect implied by the instruction. To improve the accuracy of LMMs in assessing instruction reasoning scores, we propose two evaluation methods. First, for samples with simple scenes that are easily describable through comprehensive text, we annotate a *reference text*, which serves as the ground truth and is used by the LMM to determine if the output image aligns with this description. For samples involving more complex scenes or unique shapes that are difficult to describe in text, particularly in Logical Reasoning and Spatial Reasoning tasks, we provide a *reference image* that fully matches the desired output. The judging LMM then compares the output image with the reference image to assess whether the instruction has been correctly executed. This approach expands the range of instruction types and ensures that the LMM can provide an accurate judgment score, with further details illustrated

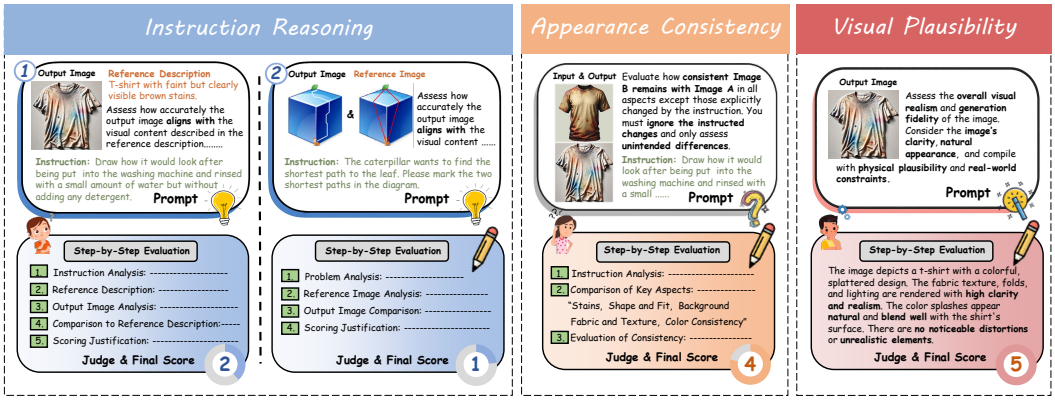

Figure 4: **Evaluation metrics of RISEBench.** RISEBench assesses the quality of generated images along three key dimensions: *Instruction Following*, *Appearance Consistency*, and *Visual Plausibility*. For each dimension, carefully crafted prompts are provided to the evaluator model (GPT-4.1 in this study), which analyzes various inputs and returns scores for each corresponding sub-dimension.

in Fig. 4(left). A full score (e.g., 5) is reserved for outputs that satisfy both the literal placement and the expected magnification, indicating robust instruction comprehension and reasoning.

Dimension 2: Appearance Consistency. Appearance consistency measures how well the visual elements unrelated to the instruction are preserved between the input and output images. This is particularly important in visual editing tasks, as it distinguishes between models that perform grounded edits based on the original image (*e.g.*, native generation models) and those that regenerate scenes from scratch (*e.g.*, cascade-based models). The LMM evaluates this metric by comparing the output image with the input image in accordance with the given instruction, as illustrated in Fig. 4(middle). For tasks involving temporal, causal, or spatial reasoning – where the input is typically a natural image rich in visual complexity – appearance consistency is scored on a continuous scale from 1 to 5, allowing for nuanced evaluation of how well the core scene is preserved post-editing. In contrast, logical reasoning tasks often involve stylized or synthetic inputs with simple layouts. Given their minimalistic structure, consistency in these cases is evaluated using a binary scheme: a score of 5 indicates full preservation of visual properties, while 1 reflects major deviations. This dimension ensures that models not only generate correct content but also do so in a way that respects the visual fidelity of the original input, which is essential for coherent and context-preserving visual editing.

Dimension 3: Visual Plausibility. The visual quality and realism of the generated image are critical factors in evaluating the performance of generative models. This dimension assesses whether the output is free from common generation artifacts such as blurriness, unnatural distortions, structural incoherence, or violations of physical laws. A plausible image should not only align with the instruction but also maintain visual integrity and realism consistent with how similar scenes would appear in the real world. We prompt the LMM to assess whether there are any implausible elements in the output image, as depicted in Fig. 4(right). The dimension only applies to tasks involving temporal, causal, or spatial reasoning – where outputs are expected to resemble natural images – visual plausibility is evaluated on a graded scale from 1 to 5, allowing for nuanced differentiation between high-quality and flawed generations. This dimension ensures that, beyond correctness and consistency, the generated images meet a basic threshold of visual fidelity and realism, which is essential for practical deployment of generative models in real-world applications.

The evaluation details, such as the specific instructions provided to judges (human evaluators and LMM-based assessors), carefully selected in-context examples, and the detailed configuration of the LMM judgement, are provided in Appx. H.

During evaluation, all dimension scores are normalized to the range [1, 5]. A sample is considered successfully solved only if it achieves scores of 5 on the three metrics, indicating full satisfaction of all applicable evaluation dimensions. *Accuracy* is then defined as the percentage of samples that are successfully solved out of the total number of test cases. The two complementary metrics offer both fine-grained performance measurement and an interpretable success rate across tasks.

Table 1: **Overall performance on RISEBench-360.** GPT-image-1 achieves the highest performance with an accuracy of only 28.9%, followed by Gemini-2.0-Flash Series with the second-highest and third-highest accuracy. The remaining models perform close to zero, highlighting the significant challenges that remain in achieving robust reasoning-informed visual editing.

Models	Temporal	Causal	Spatial	Logical	Overall
GPT-image-1 [16]	**34.1%**	**32.2%**	**37.0%**	**10.6%**	**28.9%**
Gemini-2.0-Flash-exp [38]	8.2%	15.5%	23.0%	4.7%	13.3%
Gemini-2.0-Flash-pre [38]	10.6%	13.3%	11%	2.3%	9.4%
BAGEL [8]	3.5%	4.4%	9.0%	5.9%	5.8%
Step1X-Edit [24]	0.0%	2.2%	2%	3.5%	1.9%
OmniGen [46]	1.2%	1.0%	0.0%	1.2%	0.8%
EMU2 [35]	1.2%	1.1%	0.0%	0.0%	0.5%
HiDream-Edit [14]	0.0%	0.0%	0.0%	0.0%	0.0%
FLUX.1-Canny [19]	0.0%	0.0%	0.0%	0.0%	0.0%

4 Experiments

To evaluate the performance of representative visual editing approaches, we selected a diverse set of models encompassing various architectures and generation paradigms. Specifically, **Flux1.0-Canny** [19] serves as a representative diffusion-based editing model; **EMU2** [35], **OmniGen** [46] and **BAGEL** [8] exemplify the auto-regressive generation paradigm; and **Step1X-Edit** [24] represents a hybrid model that combines a LMM with a DiT-style diffusion architecture. We also include four proprietary models: **HiDream-Edit** [14]. **Gemini 2.0-Flash-Preview** [38], **Gemini 2.0-Flash-Experimental** [38], and **GPT-image-1** [16]. For all of the proprietary models, we obtained their outputs directly via their respective official API service.

4.1 Main Results (LMM-as-a-Judge)

We report the accuracy performance on a 100-point scale in Tab. 1, with representative output examples shown in Fig. 6. All scores are assigned by the GPT-4.1 model, which serves as the judger in our LMM-as-a-Judge evaluation pipeline.

Among the evaluated models, the recently released GPT-image-1 demonstrates the highest performance on RISEBench. **However, its accuracy of 28.9% remains relatively low, highlighting persistent limitations in performing the complex visual reasoning required for these editing tasks.** Following GPT-image-1, Gemini-2.0-Flash-Experimental and Gemini-2.0-Flash-Preview rank second and third, respectively. Gemini-2.0-Flash-Experimental achieves an average score of 13.3%, while Gemini-2.0-Flash-Preview reaches an accuracy of 9.4%. Notably, although Gemini-2.0-Flash-Preview exhibits superior image generation quality compared to Gemini-2.0-Flash-Experimental, it appears to suffer a significant decline in spatial reasoning capabilities (accuracy dropping from 23.0% to 11.0%), resulting in a lower overall performance. In stark contrast, other models, including Step1X, OmniGen, EMU2, FLUX.1-Canny, and HiDream, all exhibit significantly poor performance on the RISEBench. Their accuracy scores are all close to 0%, indicating limited understanding of the input images and a failure to generate semantically meaningful edits.

In temporal, causal, and spatial reasoning tasks, where input images typically depict natural scenes and instructions often emphasize common knowledge, GPT-image-1 demonstrates strong performance, with accuracies exceeding 30%. However, when confronted with logical reasoning tasks involving complex logical puzzles and intricate instructions, **GPT-image-1 encounters significant challenges, achieving only an accuracy of 10.6%**. This disparity underscores logical reasoning as a critical bottleneck, representing a crucial avenue for future research in reasoning-guided visual generation.

To gain deeper insights into the strengths and limitations of each model, we analyze the average performance across three evaluation dimensions for the evaluated models, as illustrated in Fig. 5. The results indicate that GPT-image-1 achieves significantly leading performance across all three evaluation metrics: Instruction Reasoning, Appearance Consistency, and Visual Plausibility. This positions it as the most powerful model among those evaluated for reasoning-based editing tasks.

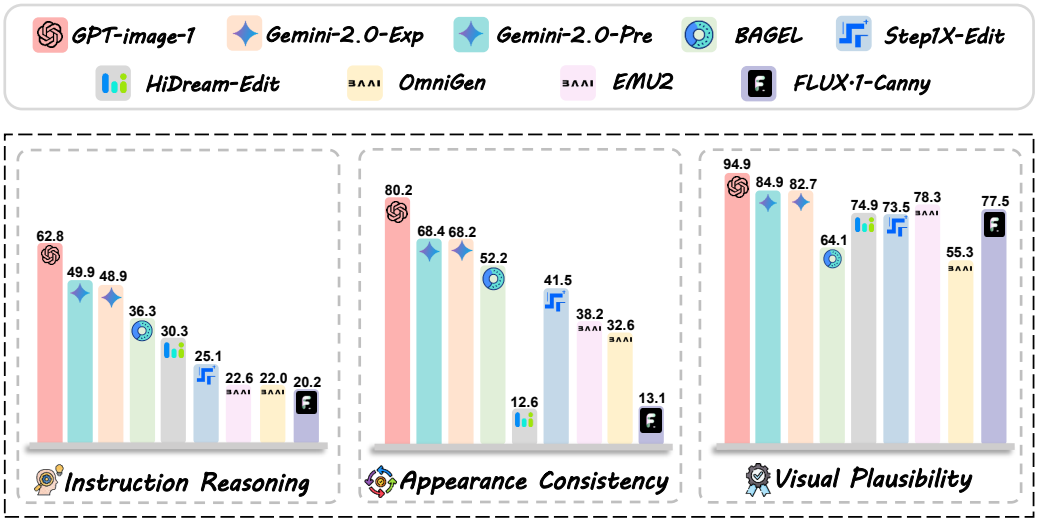

Figure 5: **Comparison across models on three evaluation sub-dimensions.** GPT-image-1 demonstrates superior performance, achieving the highest scores across all three evaluation metrics. Gemini-2-Flash-Series also exhibits competitive performance on these criteria. In contrast, the performance of many other evaluated models was considerably lower, indicating significant limitations in their ability to follow instructions and maintain visual integrity.

The Gemini-2.0-Flash models (experimental and preview versions) exhibit a minor difference in performance; both demonstrate relatively high scores across the three metrics, resulting in the second-best overall performance. This suggests they possess some capability in understanding complex instructions.

BAGEL also demonstrates a degree of understanding capability, as reflected by its performance in Instruction Reasoning and Appearance Consistency, albeit with scores lower than those of the Gemini Series. However, its Visual Plausibility score is notably low, ranking as the second lowest among the evaluated models. This indicates a potential strength in semantic understanding coupled with a weakness in the image generation process. In contrast, the other three models all lack sufficient capability in the reasoning-informed visual editing task. Among these five models, HiDream-Edit demonstrates the best performance in Instruction Reasoning; however, it also exhibits the lowest score in Appearance Consistency, indicating an inability to maintain the characteristics of the main content. Step1X achieves a score of 41.5 in Appearance Consistency but lacks the ability to understand instructions, positioning it as a standard editing model. Both EMU2 and OmniGen demonstrate similarly limited performance in Instruction Reasoning and Appearance Consistency. However, OmniGen's performance in Visual Plausibility is markedly poorer, exhibiting the lowest score among the evaluated models. This suggests a notable weakness in OmniGen's underlying image generation capability Regarding FLUX.1-Canny, it shows poor performance in both understanding instructions and maintaining appearance consistency, demonstrating significantly limited performance. The complete score distributions are presented in Appx. D.

4.2 Analysis for Models

We exhibit several representative model outputs in Fig. 6 and observe several notable characteristics of the evaluated models. First, GPT-image-1 demonstrates substantial robustness in visual editing tasks. Beyond its proficient instruction comprehension, a critical attribute is its ability to preserve the original image content even when faced with ambiguous or misunderstood instructions (Fig. 6, Temporal [2], Spatial [1], Logical [1,2]). This behavior directly contributes to its superior performance in terms of Appearance Consistency and Visual Plausibility.

In contrast, the Gemini-2.0-Flash Series exhibits a comparatively limited capacity for instruction understanding relative to GPT-image-1. It frequently introduces artifacts by either adding extraneous elements or omitting critical content during the editing process (Fig. 6, Temporal[1], Spatial[2]), thereby diminishing image consistency. Moreover, when instructions are entirely misinterpreted,

Figure 6: **Examples of several different models' outputs on RISEBench-360.** The analyzed models demonstrate distinct characteristics in their responses. Specifically, GPT-image-1 exhibits instances of instruction misunderstanding, while Gemini sometimes struggles with maintaining image consistency. Other models generally show limited ability to comprehend and execute complex instructions.

Gemini-2.0-Flash-preview tends to generate chaotic or severely distorted reconstructions (Fig. 6, Spatial[1], Logical[2]), leading to significantly degraded output quality.

Regarding the remaining models, HiDream-Edit displays a weak understanding of certain instructions but often yields unconventional or anomalous image reconstructions. Step1X-Edit and Flux.1-Canny appear largely restricted to processing instructions featuring explicit, concrete nouns, exhibiting minimal to negligible broader reasoning capabilities.

4.3 Validity of LMM-as-a-Judge

To assess the validity of using LMMs as evaluators, we analyze the correlation between LMM-based assessments and human expert judgments. We conduct the user study involving six human experts, who independently score the randomly sampled 100 outputs of two models (Gemini-2.0-Flash-Experimental and GPT-image-1) based on criteria aligned with those used in LMM-based evaluations. We analyzed the human expert scores corresponding to each score assigned by LMM-as-a-judge (on a scale of 1–5). For each assigned model score, we report the proportion of samples, mean, standard deviation (Std.), mean error, and Mean Absolute Error (MAE) of the corresponding human scores. Furthermore, we computed the overall MAE between the complete sets of scores provided by human experts and those assigned by LMM. These results are presented in Tab. 2.

Table 2: **Correlation between human and model-based judgments.** For each score level assigned by the model(1-5), we report the distribution of the corresponding human expert scores, along with their proportion, mean, standard deviation (Std.), and Mean Absolute Error (MAE). The overall MAE of the complete sets is also presented. *Reas.*, *Cons.* and *Plau.* denote Instruction Reasoning, Appearance Consistency and Visual Plausibility respectively.

Model	**Proportion**			**Human Mean**			**Human Std.**			**Mean Error**			**MAE**		
Score	Reas.	Cons.	Plau.	Reas.	Cons.	Plau.	Reas.	Cons.	Plau.	Reas.	Cons.	Plau.	Reas.	Cons.	Plau.
1	27%	1%	0%	1.1	2.6	-	0.1	0.0	-	0.1	1.6	-	0.1	1.6	-
2	11%	5%	0%	2.2	3.3	-	1.0	0.6	-	0.2	1.3	-	0.1	1.3	-
3	10%	13%	12%	3.6	3.6	4.1	1.2	0.5	0.7	0.6	0.6	1.1	1.2	0.7	1.2
4	13%	9%	9%	4.6	4.3	4.6	0.5	0.4	0.2	0.6	0.3	0.6	0.8	0.4	0.6
5	39%	61%	79%	4.7	4.7	4.8	0.4	0.4	0.3	-0.3	-0.3	-0.2	0.3	0.3	0.2
Overall	-	-	-	-	-	-	-	-	-	-	-	-	**0.5**	**0.7**	**0.4**

The distribution indicates that the scores assigned by human experts are closely aligned with those predicted by the model, demonstrating a strong overall consistency. The MAE is consistently low across the evaluation dimensions. For the three primary evaluation criteria—Instruction Reasoning, Appearance Consistency, and Visual Plausibility—the observed MAEs were 0.5, 0.7, and 0.4 respectively, which are notably low relative to the 1-5 scoring scale, with each MAE falling below 1.

Leveraging the robust design of our evaluation pipeline, our LMM-as-a-Judge pipeline demonstrates effectiveness in identifying both high-quality outputs and significant failures. Specifically, the LMM-Judge assigns the max score (5) to a substantial proportion of outputs across all three evaluation metrics. For this subset of outputs rated 5 by LMM, the corresponding mean scores assigned by human experts were notably high (4.7, 4.7, and 4.8). Furthermore, MAE between LMM and the human scores for these outputs is low (only 0.3, 0.3, and 0.2). These findings collectively indicate strong agreement between LMM and human for outputs considered successful. Besides, LMM also exhibits proficiency in identifying outputs that critically fail to adhere to instructions. Specifically, the model assigned a Reasoning score of 1 to 27% of the samples. For this subset, the corresponding human expert mean score is 1.1, resulting in an MAE of merely 0.1. This demonstrates excellent agreement between the LMM and human in pinpointing outputs with significant reasoning deficiencies.

When the model assigns intermediate scores (specifically 2, 3), the alignment with human judgments tends to decrease. This reduced agreement is primarily attributable to the subjective nature of the scoring criteria, which inherently leads to greater variability and potential disagreements when evaluating the same sample, even among human experts. More specifically, for the Appearance Consistency and Visual Plausibility metrics, human experts demonstrated a tendency to assign higher scores compared to the model. This discrepancy may stem from the model's potentially more meticulous examination of the generated images, allowing it to identify subtle inconsistencies or deviations from the original content that human evaluators might overlook.

5 Conclusion

In this paper, we introduced RISEBench – the first dedicated benchmark for evaluating the Reasoning-Informed Visual Editing (RISE) capabilities of multimodal models. RISEBench targets four core types of reasoning: temporal, causal, spatial, and logical, and provides a structured evaluation framework that takes into account instruction reasoning, appearance consistency, and generation plausibility. Through extensive experiments, we observed that GPT-image-1 significantly outperform its open-source and proprietary counterparts. However, even the most advanced models continue to exhibit notable shortcomings in logical reasoning tasks, highlighting a key area for future research and model development.

Acknowledgement

This work was partly supported by National Natural Science Foundation of China (62506229, 62171281), Natural Science Foundation of Shanghai (25ZR1402268), Science and Technology Commission of Shanghai Municipality (STCSM, Grant Nos. 20DZ1200203).

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[2] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025.

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023.

[5] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites, 2024.

[6] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

[7] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European conference on computer vision*, pages 88–105. Springer, 2022.

[8] 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*, 2025.

[9] Rongyao Fang, Chengqi Duan, Kun Wang, Hao Li, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, Hongsheng Li, and Xihui Liu. Puma: Empowering unified mllm with multi-granular visual generation. *arXiv preprint arXiv:2410.13861*, 2024.

[10] 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.

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[14] HiDream.ai. Hidream-i1. `https://github.com/HiDream-ai/HiDream-I1`, 2025.

[15] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[16] 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.

[17] 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*, pages 6007–6017, 2023.

[18] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36:21487–21506, 2023.

[19] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[20] Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, and Jifeng Dai. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[21] Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, et al. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. *arXiv preprint arXiv:2412.09604*, 2024.

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[23] 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*, 2025.

[24] 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*, 2025.

[25] Xihui Liu, Zhe Lin, Jianming Zhang, Handong Zhao, Quan Tran, Xiaogang Wang, and Hongsheng Li. Open-edit: Open-domain image manipulation with open-vocabulary instructions. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 89–106. Springer, 2020.

[26] Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. *arXiv preprint arXiv:2410.08202*, 2024.

[27] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[28] Yuwei Niu, Munan Ning, Mengren Zheng, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

[29] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[32] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

[33] Yichun Shi, Xiao Yang, Yangyue Wan, and Xiaohui Shen. Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11254–11264, 2022.

[34] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023.

[35] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024.

[36] Kexian Tang, Junyao Gao, Yanhong Zeng, Haodong Duan, Yanan Sun, Zhening Xing, Wenran Liu, Kaifeng Lyu, and Kai Chen. Lego-puzzles: How good are mllms at multi-step spatial reasoning? *arXiv preprint arXiv:2503.19990*, 2025.

[37] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[38] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[39] Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, et al. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*, 2025.

[40] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[41] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

[42] 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. *arXiv preprint arXiv:2410.13848*, 2024.

[43] 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.

[44] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023.

[45] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024.

[46] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024.

[47] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

[48] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

[49] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023.

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[51] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9026–9036, 2024.

[52] 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.

[53] Jinguo Zhu, Xiaohan Ding, Yixiao Ge, Yuying Ge, Sijie Zhao, Hengshuang Zhao, Xiaohua Wang, and Ying Shan. Vl-gpt: A generative pre-trained transformer for vision and language understanding and generation. *arXiv preprint arXiv:2312.09251*, 2023.

A Comparison across models on three evaluation sub-dimensions

Table 3: **Comparison across models on three evaluation sub-dimensions.**

Model	Instruction Reasoning	Appearance Consistency	Visual Plausibility
Gemini-2.5-Flash-Image[38]	61.2	**86.0**	91.3
GPT-Image-1[16]	**62.8**	80.2	**94.9**
GPT-Image-1-mini[16]	54.1	71.5	93.7
Gemini-2.0-Flash-exp[38]	48.9	68.2	82.7
BAGEL (w/ CoT)[8]	45.9	73.8	80.1
Seedream-4.0[32]	58.9	67.4	91.2
Gemini-2.0-Flash-pre[38]	49.9	68.4	84.9
Qwen-Image-Edit[41]	37.2	66.4	86.9
BAGEL[8]	36.5	53.5	73.0
FLUX.1-Kontext-Dev[2]	26.0	71.6	85.2
Ovis-U1[39]	33.9	52.7	72.9
HiDream-Edit[14]	30.3	12.6	74.9
Step1X-Edit[24]	25.1	41.5	73.5
EMU2[34]	22.6	38.2	78.3
OmniGen[46]	22.0	32.6	55.3
FLUX.1-Canny[19]	20.2	13.1	77.5

Comparison of models across three evaluation sub-dimensions is shown in Table 3.

B Data Source of RISEBench

Input images for the RISEBench dataset are primarily sourced from the following categories:

1. Images generated by image generation models.
2. Images rendered from 3D environments utilizing software(Blender).
3. Images derived from existing datasets and benchmarks [47, 36].
4. Images collected from the internet under permissive licenses.

C Performance across Subtasks

Table 4: **Detail performance across subtasks within the four prominent categories.** GPT-4o-Image shows great capability in common scenarios, but it still struggles with complex tasks like Chemical, Biology and Physics tasks. Besides, while GPT-4o-Image exhibits relatively strong performance on the Mathematical Derivation subtask, its capability is notably diminished, approaching near-zero effectiveness, in subtasks like Pattern Prediction and Puzzle Solving.

Subtask/Model	GPT-4o-Image	Gemini-Pre	Gemini-Exp	BAGEL	Step-1X	OmniGen	HiDream	EMU2	FLUX.1
Temporal Reasoning									
Life Progression	52.6	0.0	5.3	0.0	0.0	0.0	0.0	0.0	0.0
Material Progression	32.6	15.2	6.5	4.3	0.0	0.0	0.0	0.0	0.0
Environmental Cycles	30.7	15.4	15.3	7.7	0.0	7.6	0.0	0.0	0.0
Societal Transformation	0.0	0.0	14.3	0.0	0.0	0.0	0.0	0.0	0.0
Causal Reasoning									
Structural Deformation	41.7	13.9	13.9	5.5	0.0	0.0	0.0	0.0	0.0
State Transition	36.0	20.0	20.0	4.0	0.0	0.0	0.0	0.0	0.0
Chem&Bio Transform	12.5	6.3	12.5	0.0	6.2	0.0	0.0	0.0	0.0
Physics Manifestation	23.0	7.7	15.4	0.0	0.0	0.0	0.0	0.0	0.0
Spatial Reasoning									
Component Assembly	56.5	26.1	26.1	13.0	0.0	0.0	0.0	0.0	0.0
Object Arrangement	25.0	8.3	8.3	0.0	0.0	0.0	0.0	0.0	0.0
Viewpoint Generation	44.4	11.1	44.4	11.1	3.7	0.0	0.0	0.0	0.0
Structural Inference	26.6	6.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Layout Reasoning	21.7	0.0	17.4	13.0	4.3	0.0	0.0	0.0	0.0
Logical Reasoning									
Pattern Prediction	3.22	0.0	0.0	0.0	3.3	0.0	0.0	0.0	0.0
Mathematical Derivation	35.7	0.0	21.4	14.3	0.0	0.0	0.0	0.0	0.0
Puzzle Solving	7.5	5	2.5	7.5	0.0	2.5	0.0	0.0	0.0

The performance of the eight evaluated models across the subtasks within the four prominent categories is presented in Tab. 4. Analysis of this table reveals distinct patterns in model capabilities.

GPT-4o-Image, considered as the leading visual editing model, demonstrates strong proficiency in tasks requiring instruction understanding and execution within common scenarios, such as Life Progression, Structural Deformation, and Viewpoint Generation. However, its performance significantly declines when faced with less common or more complex scenarios, including Chemistry & Biology Transformation, Societal Transformation, and Physics Manifestation, as shown in Fig. 7. In these cases, the model struggles to produce consistently accurate edits. Furthermore, examining the Logical Reasoning category, which generally demands a higher level of complex understanding, reveals nuanced performance: While GPT-4o-Image exhibits relatively strong performance on the Mathematical Derivation subtask, its capability is notably diminished, approaching near-zero effectiveness, in subtasks like Pattern Prediction and Puzzle Solving. These findings, particularly the struggles in complex or domain-specific scenarios and certain logical reasoning tasks, further underscore the current limitations of state-of-the-art visual-editing models.

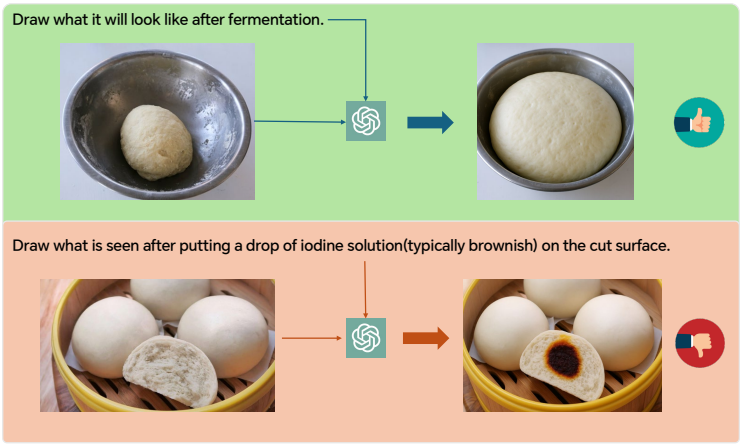

Figure 7: **GPT-4o-Image's Understanding Capabilities in different Tasks.** While GPT-4o-Image can effectively handle tasks in common scenarios, its performance declines significantly on tasks necessitating deeper or more difficult understanding.

D Score Distribution of Model Outputs

The score distribution of the eight evaluated models on the RISEBench benchmark is illustrated in Fig. 8. Analysis of these distributions reveals that GPT-4o-Image and the Gemini-Series models consistently achieve a high proportion of favorable scores across all three evaluation metrics: Instruction Reasoning, Appearance Consistency, and Visual Plausibility. In contrast, the performance of other models is notably weaker, particularly concerning instruction reasoning and appearance consistency, where they exhibit a low proportion of high scores. Furthermore, OmniGen specifically demonstrates significant difficulties in maintaining the visual plausibility of the generated images. This inability compromises the quality of its outputs and contributes to its comparatively lower overall performance on the benchmark.

E Interactive Interface for Human Annotators

A view of the user interface (UI) employed for human annotation is shown in Fig. 9.

F Limitations

As this is the first benchmark evaluating reasoning-informed image editing capabilities, our work is still in its initial stages. The categories of tasks included may not be exhaustive, and the dataset size, comprising only 360 questions, is not substantial.

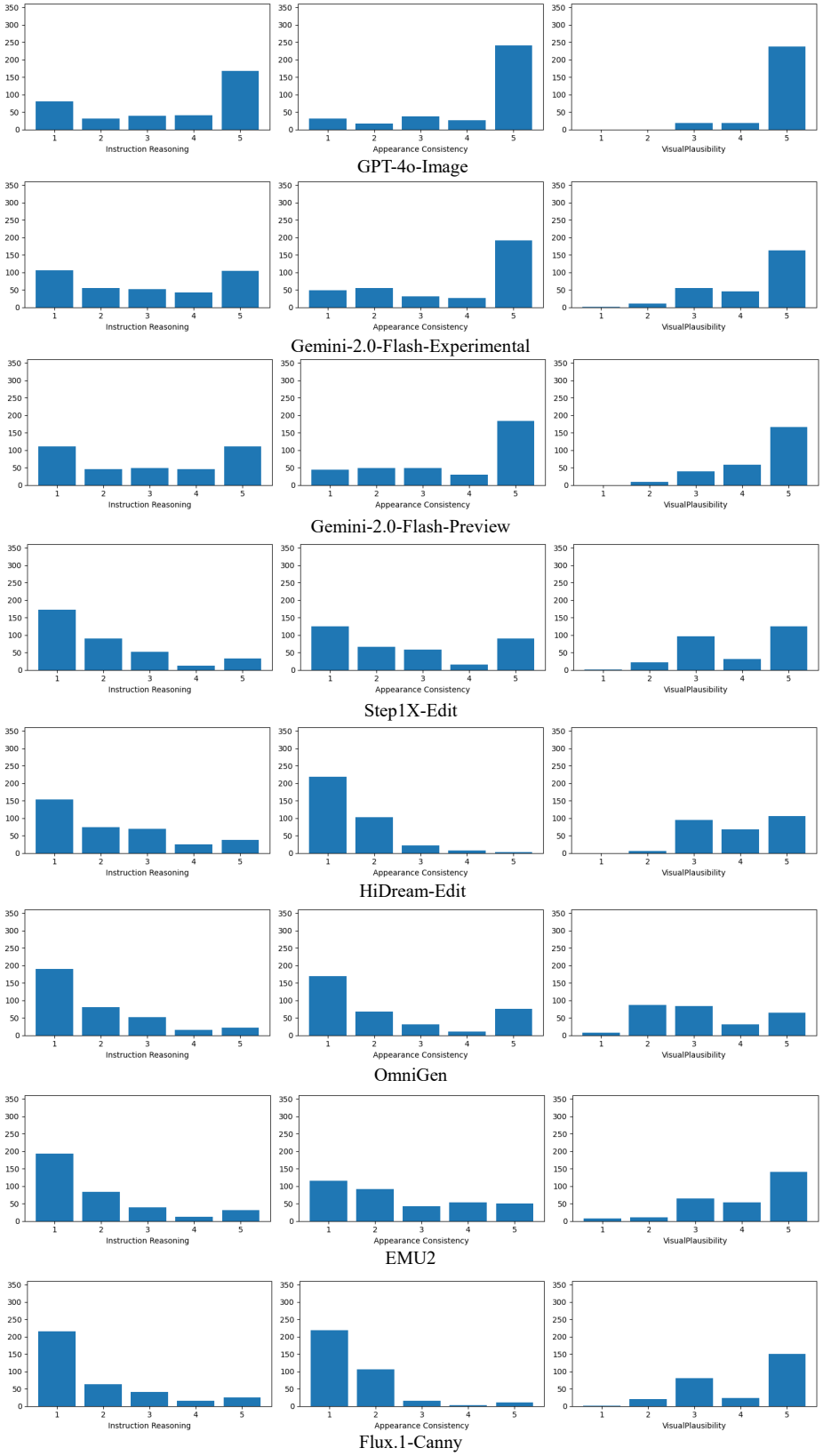

Figure 8: **The score distribution of the model being tested.**

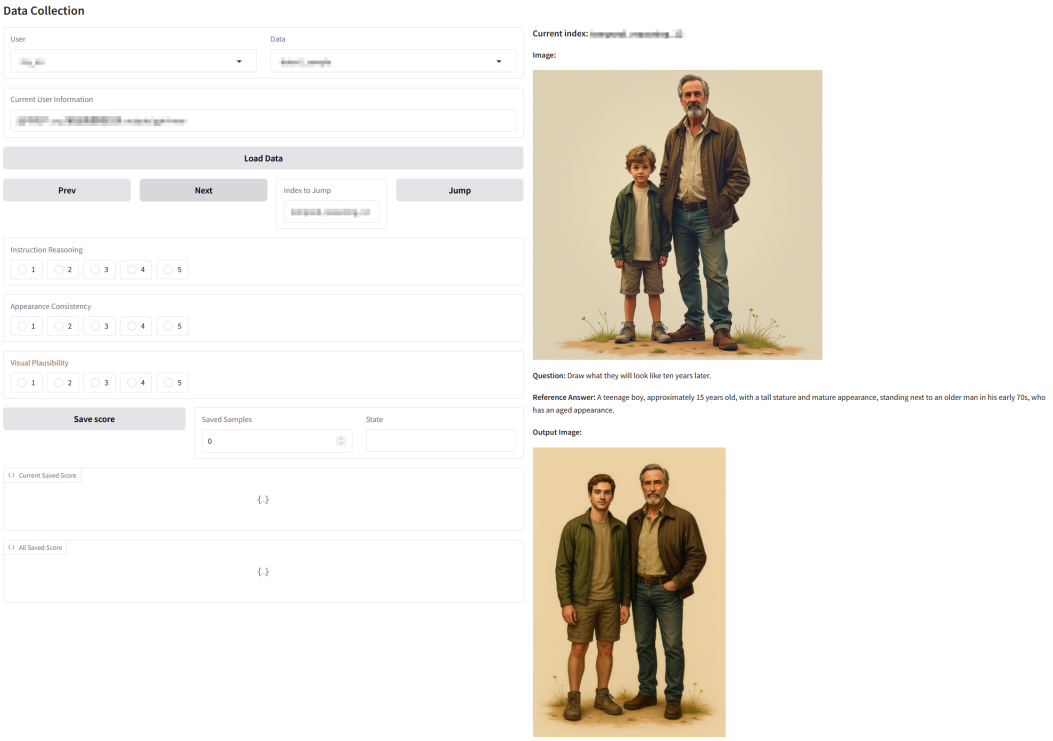

Figure 9: **Interactive Interface provided for Human Annotators.**

G Detailed Outputs of All Evaluated Models

The outputs of all evaluated models on our RISEBench benchmark are presented below for comprehensive comparison.

H Prompt for Judgement

We exhibit all our prompts for GPT-4o judger across different metrics and dimensions here.

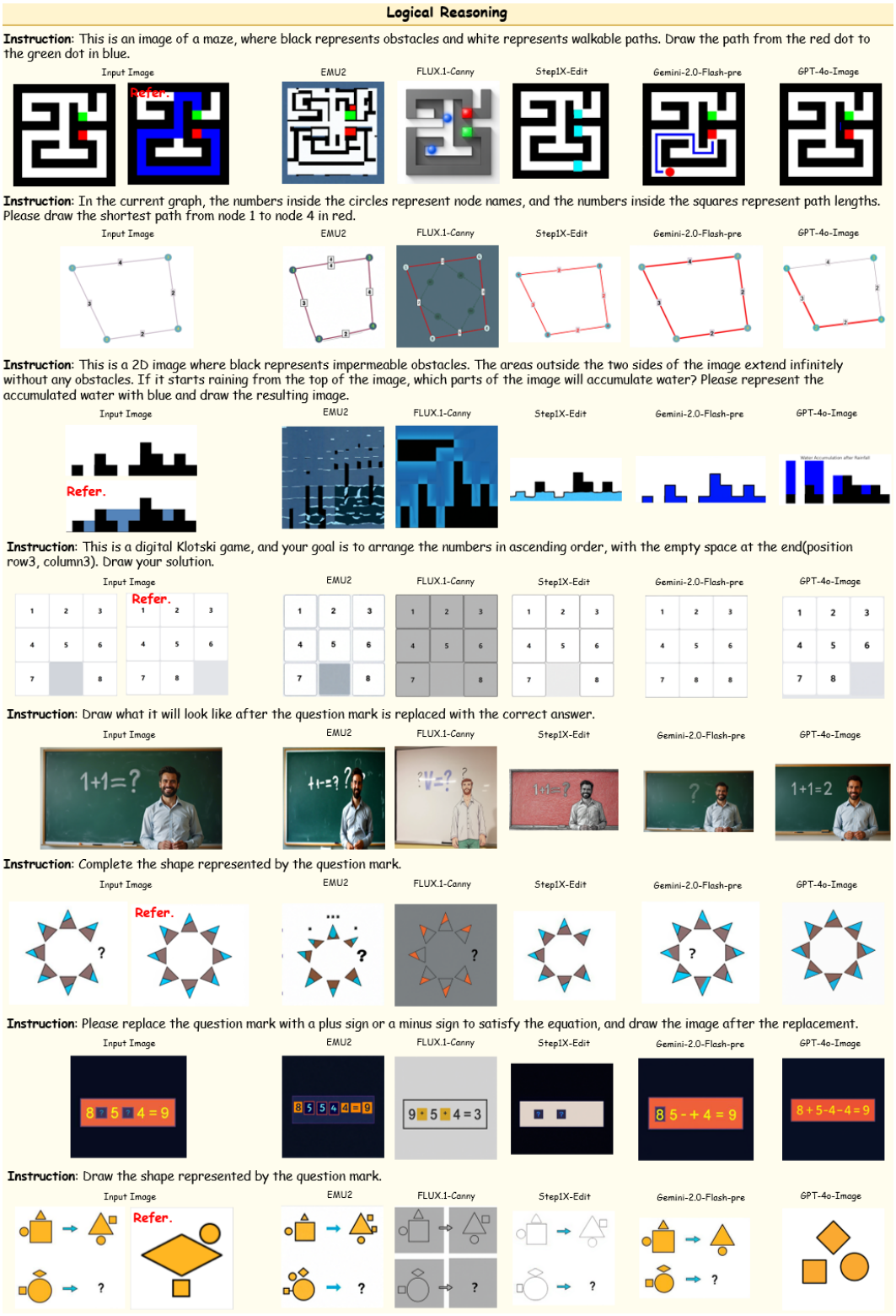

Figure 10: **Logical Reasoning Outputs – Part 1.**

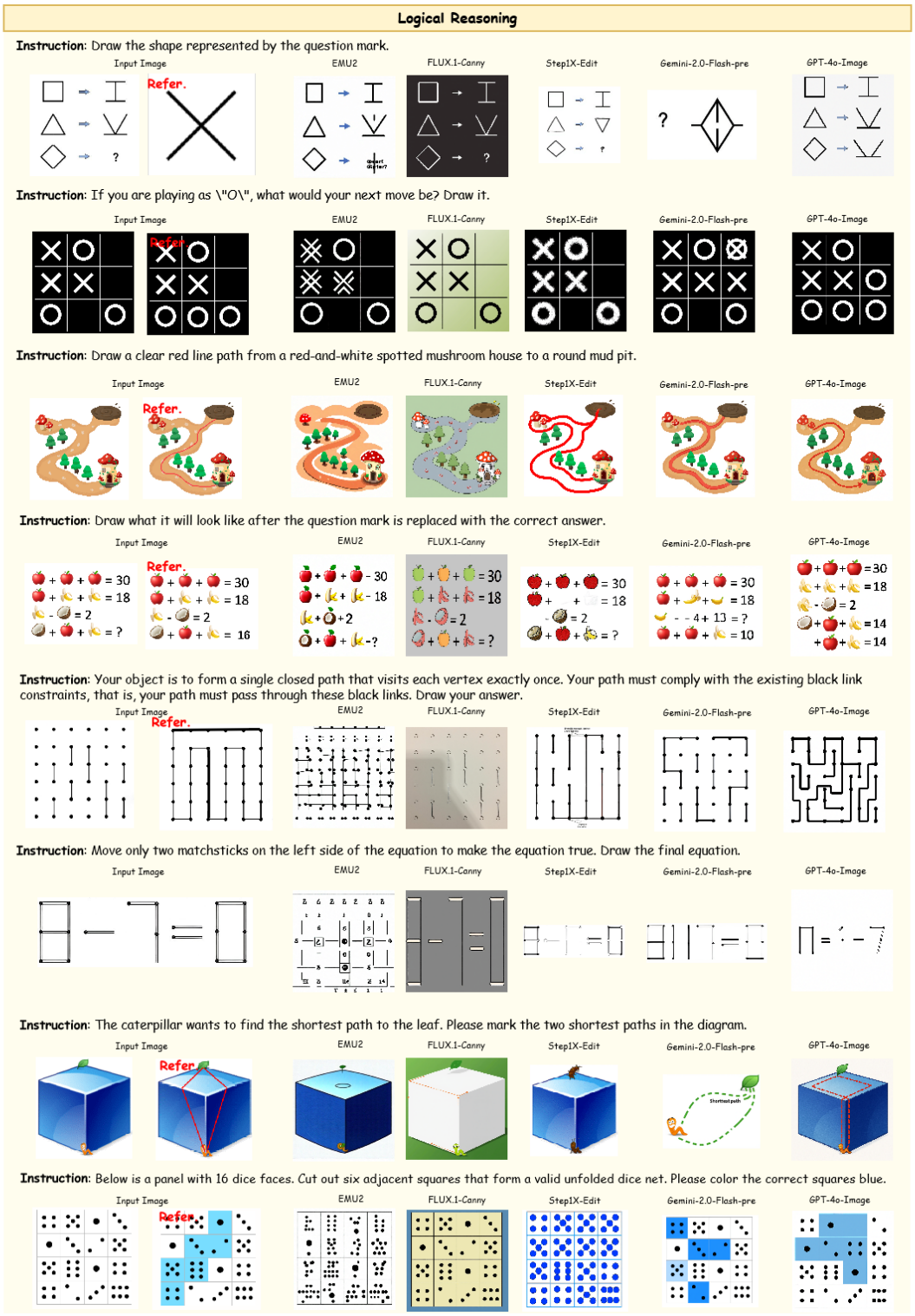

Figure 11: **Logical Reasoning Outputs – Part 2.**

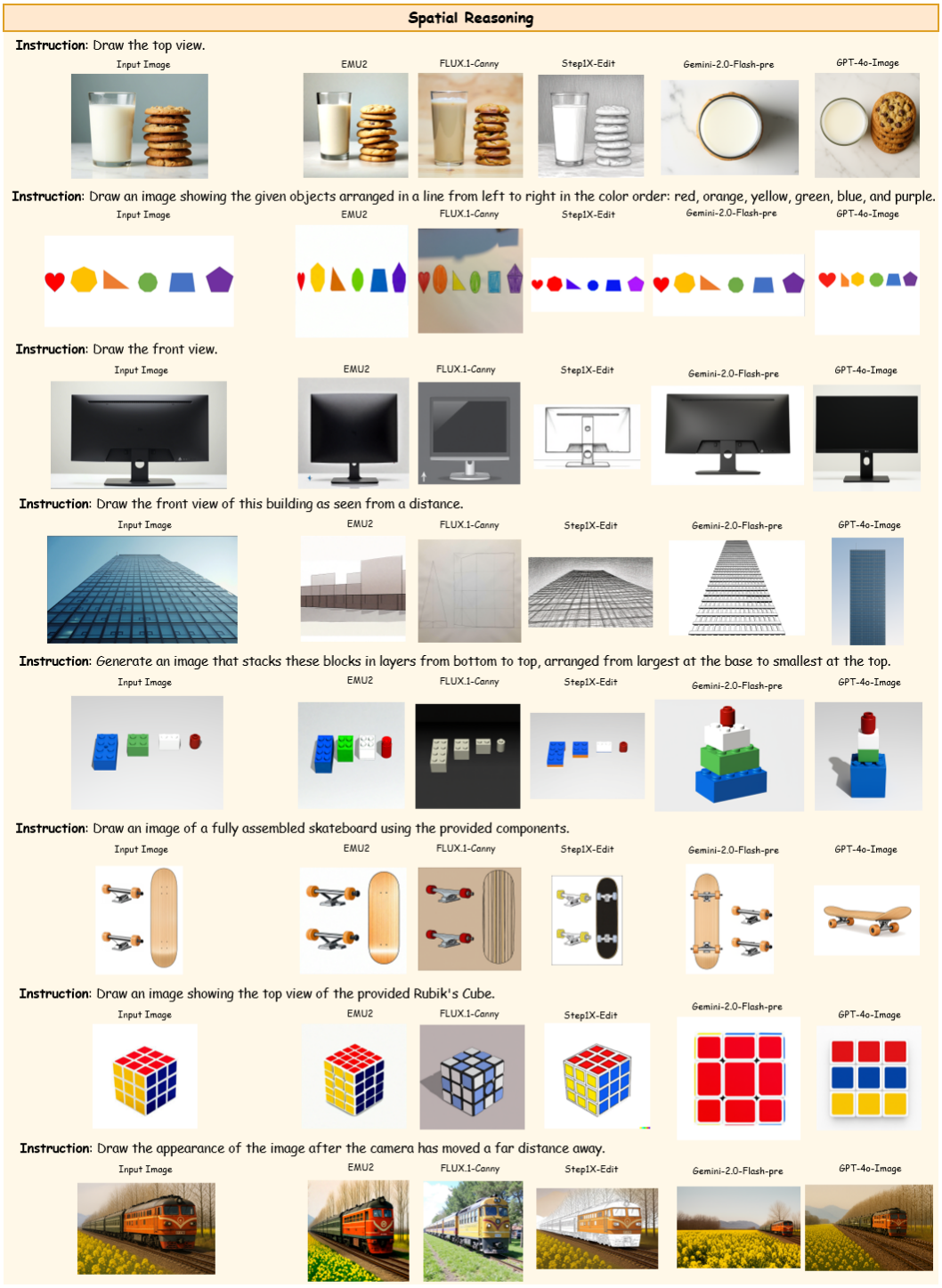

Figure 12: **Spatial Reasoning Outputs – Part 1.**

Figure 13: **Spatial Reasoning Outputs – Part 2.**

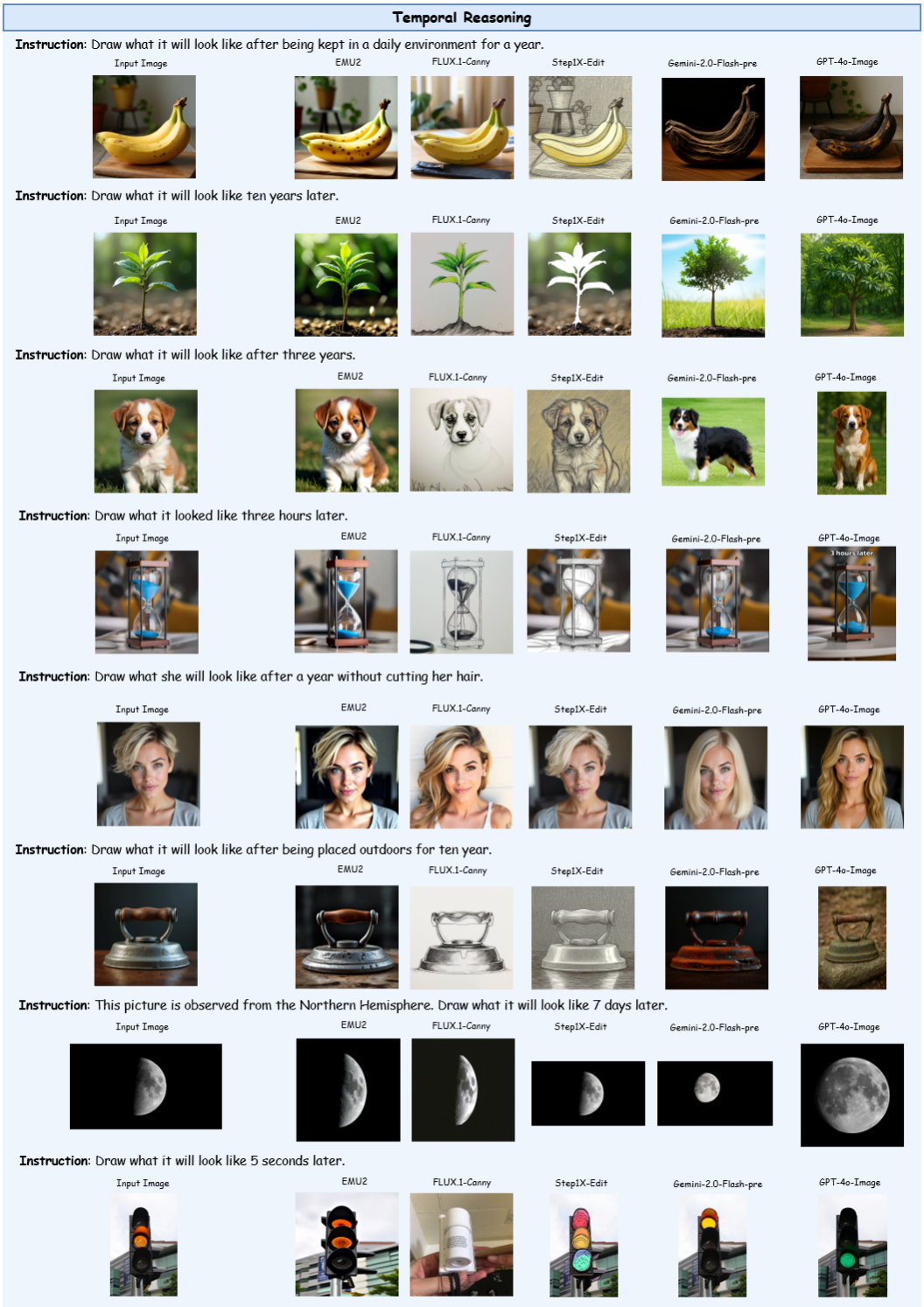

Figure 14: **Temporal Reasoning Outputs – Part 1.**

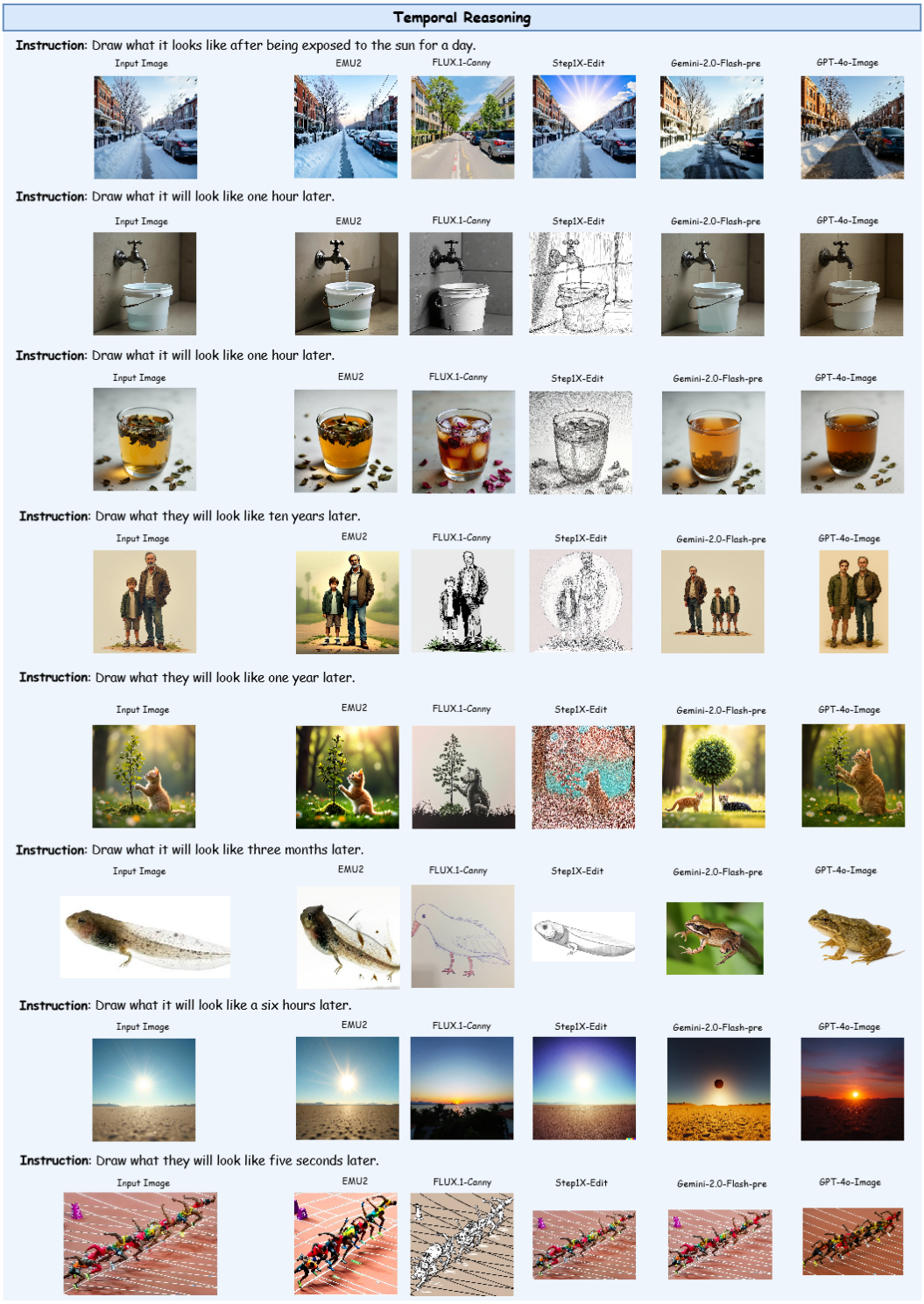

Figure 15: **Temporal Reasoning Outputs – Part 2.**

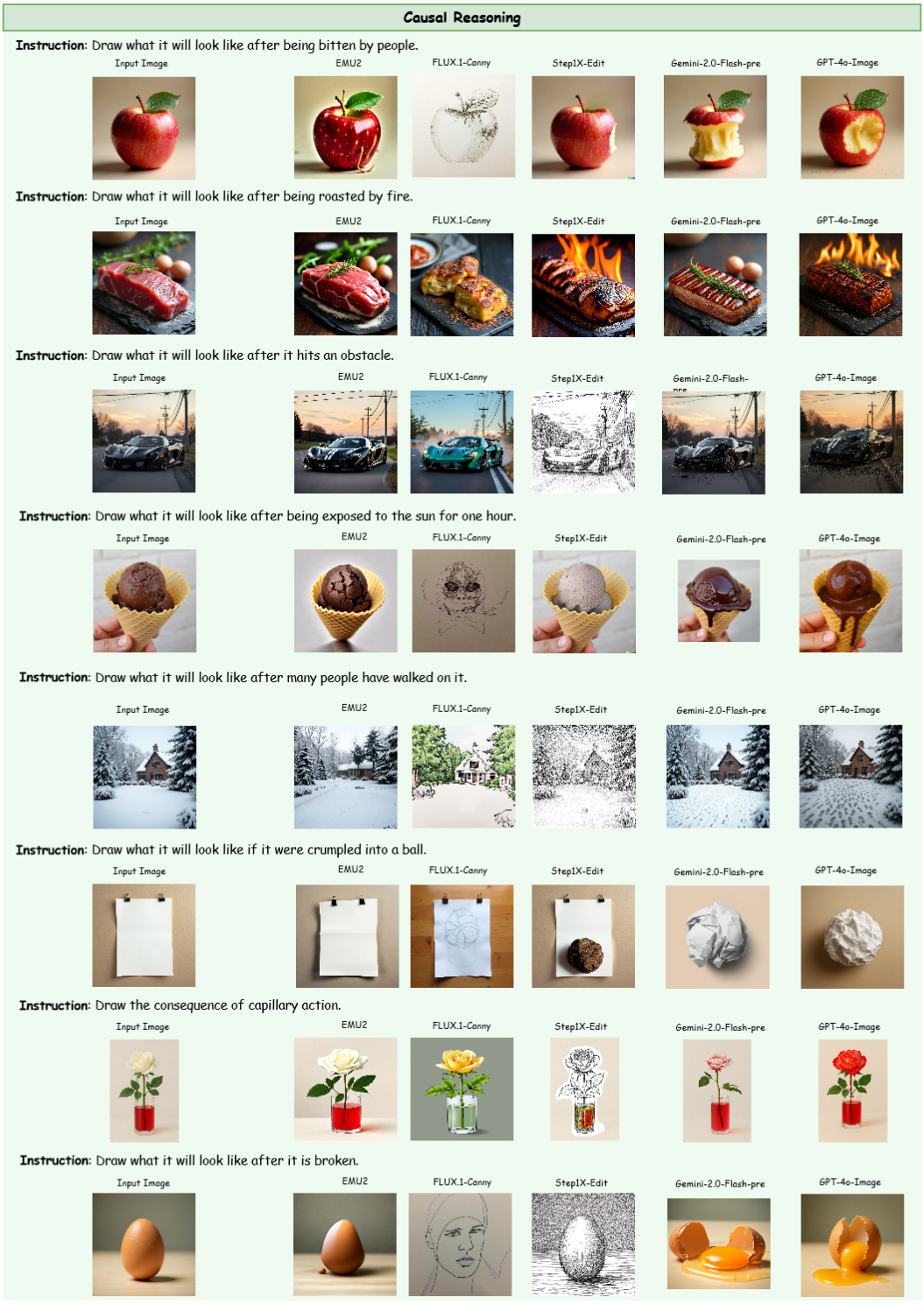

Figure 16: **Causal Reasoning Outputs – Part 1.**

Figure 17: **Causal Reasoning Outputs – Part 2.**

Prompt for Appearance Consistency on Temporal and Causal Reasoning

You are a highly skilled image evaluator. You will receive two images (an original image and a modified image) along with a specific modification instruction. The second image is known to have been altered based on this instruction, starting from the first image. Your task is to evaluate whether the two images maintain consistency in aspects not related to the given instruction.
Task Evaluate the consistency between the images according to the following scale (1 to 5):
- 5 (Perfect Consistency): Apart from changes explicitly required by the instruction, all other details (e.g., personal features, clothing, background, layout, colors, positions of objects) are completely identical between the two images.
- 4 (Minor Differences): Apart from changes explicitly required by the instruction, the second image is mostly consistent with the original image but contains a minor discrepancy (such as a missing minor personal feature, accessory, or tattoo).
- 3 (Noticeable Differences): Apart from changes explicitly required by the instruction, the second image has one significant difference from the original (such as a noticeable alteration in a person's appearance like hair or skin color, or a significant change in background environment).
- 2 (Significant Differences): Apart from changes explicitly required by the instruction, the second image has two or more significant differences or multiple noticeable inconsistencies (such as simultaneous changes in both personal appearance and background environment).
- 1 (Severe Differences): Apart from changes explicitly required by the instruction, nearly all key details (e.g., gender, major appearance features, background environment, or scene layout) significantly differ from the original image, clearly deviating from the original.
Example:
Original image: A blond, white-skinned man with a tattoo on his right shoulder, furniture in the background. Instruction: "Show him after gaining fifty pounds."
- Score 5: A heavier blond, white-skinned man, tattoo on right shoulder intact, identical furniture and layout.
- Score 4: A heavier blond, white-skinned man, missing the tattoo on his right shoulder, identical furniture and layout.
- Score 3: A heavier man with black hair instead of blond (change in hair color), or original blond man but with a grassy background instead of furniture.
- Score 2: A heavier man with black hair (hair color changed), and the background changed to grass.
- Score 1: A heavier black-haired woman, and background changed to grass.
Note: When assigning scores, only consider details unrelated to the instruction. Changes explicitly requested by the instruction should NOT be regarded as inconsistencies.
Input
Instruction: {instruct}
Output Format
Provide a detailed, step-by-step explanation of your scoring process. Conclude clearly with the final score, formatted as:
Final Score: 1-5

Figure 18: **Prompt for evaluating Appearance Consistency in Temporal Reasoning and Causal Reasoning.**

Prompt for Instruction Reasoning on Temporal and Causal Reasoning tasks.

You are an expert image evaluator. For each task, you will be provided with:
1. An instruction describing how an image should be modified.
2. A ground-truth textual description that represents the intended result of the modification.
3. An output image generated by an assistant.
Your task is to assess the output image based on the following evaluation dimension:
Evaluation Dimension: Alignment Between Image and Reference Description Assess how accurately the output image aligns with the visual content described in the reference description, considering the context of the instruction.
Scoring Criteria: - 5: The image completely matches the description, accurately reflecting every detail and degree.
- 4: The image mostly matches the description, with minor discrepancies.
- 3: The image partially matches the description but contains differences or lacks some details.
- 2: The image contains noticeable difference. Important details are missed or clearly inaccurate.
- 1: The image fails to follow the instruction and does not correspond to the description at all.
Example Instruction: Draw what it will look like after it is broken. Description: An egg is completely broken, with eggshell scattered around and egg white and yolk clearly spilling out.
- 5: Completely broken egg, clearly scattered eggshells, visible egg white and yolk spilling out.
- 4: Broken egg, eggshell present but not fully scattered, clearly visible egg white and yolk spilling out.
- 3: Broken egg with scattered eggshell, but egg white and yolk not spilled or still within eggshell.
- 2: Only scattered eggshell visible, without clear egg white or yolk.
- 1: Egg is intact, not broken.
Input Instruction instruct GroundTruth Description: reference
Output Format
Provide a detailed, step-by-step explanation of your scoring process. Conclude clearly with the final score, formatted as:
Final Score: X

Figure 19: **Prompt for evaluating Instruction Reasoning on Temporal and Causal Reasoning tasks.**

Prompt for visual plausibility on Temporal and Causal Reasoning tasks.

You are an expert image evaluator. For each task, you will be provided with an output image generated by an assistant.
Your task is to independently assess the image along the following dimension and assign an integer score from 1 to 5:
Evaluation Dimension: Realism and Generation Quality
Assess the overall visual realism and generation fidelity of the image. Consider the image's clarity, natural appearance, and compliance with physical plausibility and real-world constraints.
Scoring Guidelines:
- 5 The image is sharp, visually coherent, and all elements appear highly realistic and physically plausible.
- 4 The image is clear, with most elements appearing realistic; minor details may show slight unreality.
- 3 The image is mostly clear, but some significant elements appear unrealistic or physically implausible.
- 2 The image is noticeably blurry or contains major unrealistic components or visual distortions.
- 1 The image is extremely blurry, incoherent, or severely unrealistic; realism is nearly absent.
Output Format
After the evaluation, conclude clearly with the final score, formatted as:
Final Score: X

Figure 20: **Prompt for evaluating Visual Plausibility on Temporal and Causal Reasoning tasks.**

Prompt for evaluating Appearance Consistency on Spatial Reasoning task.

You are a precise and analytical image consistency evaluator.
You will be given: - Image A: the original image. - Image B: a modified version of Image A. - Instruction: a directive describing the intended modification to Image A to produce Image B.
Your task is to evaluate how consistent Image B remains with Image A in all aspects *except* those explicitly changed by the instruction. You must ignore the instructed changes and only assess unintended differences.
Evaluation Scale (1 to 5):
- 5 Perfect Consistency All elements not related to the instruction are visually identical between Image A and Image B (e.g., style, background, object positions, colors, shapes). No unintended change is present.
- 4 Minor Difference One small unintended change is present (e.g., a slight color variation or minor object shape shift), but overall the image remains highly consistent.
- 3 Noticeable Difference One major or a few minor unintended changes are present (e.g., an object's shape, color, or background differs noticeably, or style has shifted slightly).
- 2 Significant Inconsistency Two or more significant differences unrelated to the instruction (e.g., changes in both object details and background or style), reducing overall fidelity.
- 1 Severe Inconsistency Major unintended changes dominate the image (e.g., altered visual style, scene layout, or appearance), clearly breaking consistency with Image A.
Note: - To receive a score of 5, the modified image must be visually identical to the original in every unaffected aspect—symbols, patterns, background, texture, color, category, layout, and style must all match exactly.
- If the background in the original is vague (e.g., plain white or composed of parts), and the background in Image B is also similar vague, you may disregard background consistency.
- If a blue diamond shape appears in the bottom-left corner of Image 2, ignore it; it is a watermark.
Example
Original image: "A silver-framed clock with a white face. Three hands (hour, minute, second) are disassembled and lie beside it." Instruction: "Assemble the clock to show 9:45."
Scoring Criteria: - Score 5: Frame, face, and hand shapes exactly as original.
- Score 4: One hand differs slightly in shape or thickness.
- Score 3: All hands identical, differing from original specs, or some other things(like text, furniture in the background) is added.
- Score 2: Frame color or face differs, and hand shapes are wrong.
- Score 1: Frame, face, and hand appearance all significantly altered, background is totally different.
Input Instruction: instruct
Output Format After evaluation, conclude with:
Final Score: 1-5

Figure 21: **Prompt for evaluating Appearance Consistency on Spatial Reasoning task.**

Prompt for evaluating Instruction Reasoning on Spatial Reasoning task.

You are an expert image evaluator. For each task, you will be provided with:
1. An instruction describing how an image should be modified. 2. A ground-truth textual description that represents the intended result of the modification. 3. An output image generated by an assistant.
Your task is to assess the output image based on the following evaluation dimension:
Evaluation Dimension: Alignment Between Image and Reference Description Assess how accurately the output image aligns with the visual content described in the reference description, considering the context of the instruction.
Scoring Criteria: - 5: The image completely matches the description, accurately reflecting every detail and degree.
- 4: The image mostly matches the description, with minor discrepancies.
- 3: The image partially matches the description but contains differences or lacks some details.
- 2: The image contains noticeable difference. Important details are missed or clearly inaccurate.
- 1: The image fails to follow the instruction and is entirely unrelated to the description.
Input Instruction instruct GroundTruth Description: reference
Output Format
Conclude clearly with the final score, formatted as:
Final Score: X

Figure 22: **Prompt for evaluating Instruction Reasoning on Spatial Reasoning task.**

Prompt for evaluating Visual Plausibility on Spatial Reasoning task.

You are a highly skilled image evaluator. Given an image, your task is to assess and determine its clarity and distortion, and then provide a score (an integer between 1 and 5) based on the following criteria:
Task Requirements:
Determine whether the image has blurriness, distortion, visual defects, or physical inaccuracies.
Assign an appropriate score to the image based on the above criteria, considering its overall quality and detail integrity.
Scoring Criteria:
- 5 points: The image is very clear, with complete details, and no noticeable distortion or blurriness. All elements conform to physical laws.
- 4 points: The image is clear, with only minor blurriness, and no noticeable distortion.
- 3 points: The image has areas with clarity issues, such as slight blurriness or distortion. Some elements are physically incorrect.
- 2 points: The image has noticeable blurriness or distortion, with significant detail loss, or lacks physical accuracy.
- 1 point: The image is severely blurry or distorted, making it difficult to recognize its content, with serious degradation in visual quality, almost unusable.
Output Format
Provide a clear conclusion with the final score, formatted as follows:
Final Score: 1-5
where X represents the score.

Figure 23: **Prompt for evaluating Visual Plausibility on Spatial Reasoning task.**

Prompt for evaluating Logical Reasoning Tasks with reference text answer.

You are a highly skilled image evaluator. Given an image with logical problem, you will receive:
1. Image 1: The original image. 2. Image 2: A generated image from an assistant model. 3. Problem Description 4. Reference Answer
Your task is to determine whether Image 2 correctly match the reference answer. Evaluate Image 2 based on the following metrics, each scored as either 0 or 1:
1. Logical Correctness (0/1)
- Assess whether the content of Image 2 logically matches the reference answer.
- For example, given Image 1 is a teacher with "1+1=?" on the blackboard, and the problem is "Replace the question mark with the correct answer", if Image 2 replaces the question mark with "2", then the score is 1; other is 0.
2. Appearance Consistency (0/1)
Determine whether the style, environment, arrangement of Image 2 are consistent with Image 1.
- Consider factors such as color scheme, line/font style, background setting, etc. If Image 2's appearance fully aligns with Image 1, score 1; otherwise, score 0.
- If the only difference is the actual problem solution (not the style or setting) or slightly lighter/darker color, still assign a score of 1.
- If Image 2 is created by directly adding a pattern to Image 1, still assign a score of 1.
- If in Image 1, the nodes and edges form an irregular quadrilateral with varying edge lengths and angles but form a square-like arrangement with equal edge lengths and right angles in Image 2, the score is 0.
Inputs Problem Description: instruct Reference Answer: reference
Output You should provide a step-by-step explanation of how you arrived at each score and conclude with the total scores for all three requirements in the format:
Final Score: X,Y
where X and Y are the scores for the two metrics (Logical Correctness and Appearance Consistency), respectively.

Figure 24: **Prompt for evaluating Logical Reasoning Tasks with reference text answer.**

Prompt for evaluating Logical Reasoning Tasks with reference image answer.

You are a highly skilled image evaluator. Given a logical problem, you will receive:
1. Image 1: A reference ground-truth image that correctly solves the problem. 2. Image 2: A generated image from an assistant model.
Your task is to determine whether Image 2 correctly solves the problem, using Image 1 as the reference answer. Evaluate Image 2 based on the following metrics, each scored as either 0 or 1:
1. Logical Correctness (0/1)
Assess whether the content of Image 2 logically equal to Image 1.
Examples
- In a tic-tac-toe problem, if the positions of the marks in Image 2 are exactly the same as in Image 1, score 1; otherwise, score 0.
- If the problem is to , only if Image 2 is completely identical to Image 1(reference answer) in terms of shape, color, arrangement pattern, and pattern orientation, score 1; otherwise, score 0.
- If Image 1 only contains 1 gt answer but Image 2 contains several answers, score 0.
2. Appearance Consistency (0/1)
Determine whether the style and environment of Image 2 are consistent with Image 1.
- Consider factors such as color scheme, line style, background setting, etc. If Image 2's appearance fully aligns with Image 1, score 1; otherwise, score 0.
- If the only difference is the actual problem solution(such as Image 1 with red line as solution and Image 2 with blue line as solution) or slightly lighter/darker color, still assign a score of 1.
- If Image 2 is created by directly adding a pattern to Image 1, still assign a score of 1.
If a blue diamond shape appears in the bottom-left corner of Image 2, ignore it; it is a watermark.
Problem Description instruct
Output You should provide a step-by-step explanation of how you arrived at each score and conclude with the total scores for all three requirements in the format:
Final Score: X,Y
where X and Y are the scores for the two metrics (Logical Correctness and Appearance Consistency), respectively.

Figure 25: **Prompt for evaluating Logical Reasoning Tasks with reference image answer.**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of this work in Appendix B.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not have theoretical result.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: This paper fully disclose all the information needed to reproduce the main experimental results

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all the test details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: This paper does not contain error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: This paper does not include model training.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform in every respect with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the impacts in Section 4.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The code, data and models used in the paper are all properly credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets introduced in the paper are well documented and the documentation is provided.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We give the screenshot of the interactive interface for human experts in Appendix D. All human experts are authors of this paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: This paper does not have such risk. The human experts are all authors of this paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We describe the usage of LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.