R2I-Bench: Benchmarking Reasoning-Driven Text-to-Image Generation

Kaijie Chen★, Zihao Lin*♠, Zhiyang Xu*♣, Ying Shen*♦,
Yuguang Yao♡, Joy Rimchala◇, Jiaxin Zhang◇, Lifu Huang♠✉

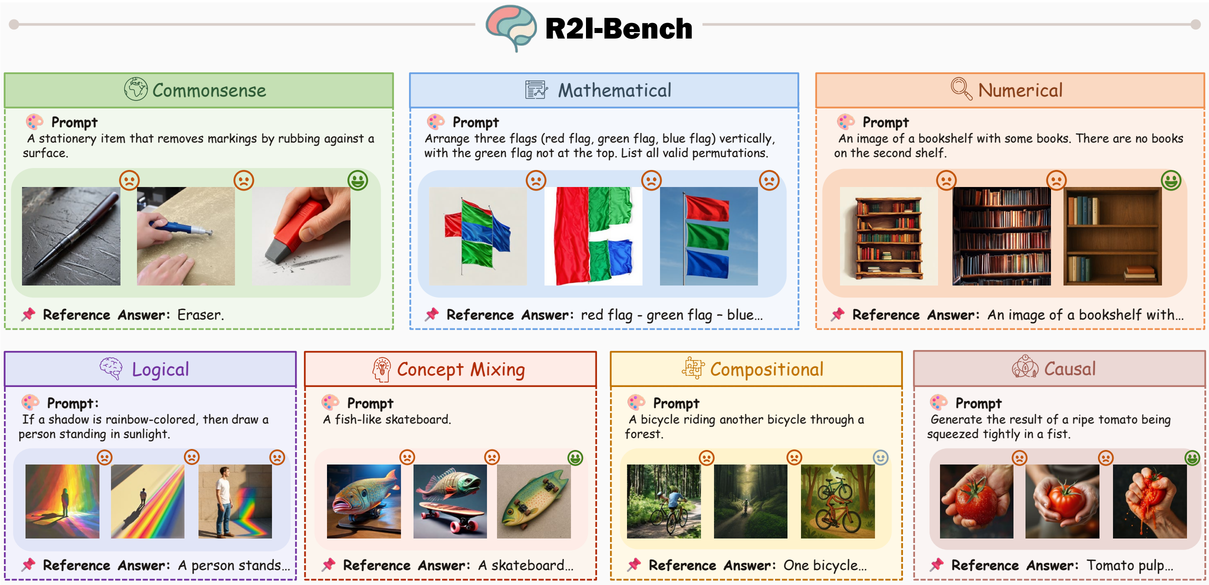

Figure 1: We introduce **R2I-Bench**, a comprehensive benchmark designed to assess the reasoning capabilities of text-to-image (T2I) generation models. It encompasses 7 primary reasoning categories, which are further subdivided into 32 fine-grained subcategories.

Abstract

Reasoning is a fundamental capability often required in real-world text-to-image (T2I) generation, e.g., generating "*a bitten apple that has been left in the air for more than a week*" necessitates understanding temporal decay and commonsense concepts. While recent T2I models have made impressive progress in producing photorealistic images, their reasoning capability remains underdeveloped and insufficiently evaluated. To bridge this gap, we introduce **R2I-Bench**, a comprehensive benchmark specifically designed to rigorously assess reasoning-driven T2I generation. **R2I-Bench** comprises $3,068$ meticulously curated data instances, spanning 7 core reasoning categories, including commonsense, mathematical, logical, compositional, numerical, causal, and concept mixing. To facilitate fine-grained evaluation, we design **R2I-Score**, a QA-style metric based on instance-specific, reasoning-oriented evaluation questions that assess three critical dimensions: *text-image alignment*, *reasoning accuracy*, and *image quality*. Extensive experiments with 16 representative T2I models, including a strong pipeline-based framework that decouples reasoning and generation using the state-of-the-art language and image generation models, demonstrate consistently limited reasoning performance, highlighting the need for more robust, reasoning-aware architectures in the next generation of T2I systems. Project page: `https://r2i-bench.github.io`.

1 Introduction

Reasoning is a fundamental capability underpinning most, if not all, human cognitive tasks, including text-to-image (T2I) generation. In real-world scenarios, prompts often require models to go beyond surface-level descriptions and engage in multi-step reasoning. For example, generating an image for "*a bitten apple that has been left in the air for more than a week*" requires understanding the concept of decay over time, inferring the visual appearance of a spoiled apple, composing that with contextual cues, and finally generating an image to depict "*a bitten and spoiled apple*".

However, despite recent advances, most existing T2I models, whether based on diffusion (Esser et al., 2024a; Xie et al., 2025a; Qin et al., 2025b; Yang et al., 2024),

* Equal Contribution ★ Tongji University ♠ University of California, Davis ♣ Virginia Tech ♦ UIUC ♡ Michigan State University ◇ Intuit AI Research

Benchmarks	Reasoning Capabilities Evaluated in Text-to-Image Generation							Human Annotation
	Commonsense	Compositional	Numerical	Mathematical	Concept Mixing	Logical	Causal	
OK-VQA (Marino et al., 2019)	✓	✗	✗	✗	✗	✗	✗	✗
Winoground (Thrush et al., 2022)	✗	✓	✗	✗	✗	✗	✗	✗
HEIM (Lee et al., 2023)	✗	✓	✗	✗	✗	✗	✗	✗
GeckoNum (Ghosh et al., 2023)	✗	✗	✓	✗	✗	✗	✗	✓
GenEval (Ghosh et al., 2023)	✗	✓	✓	✗	✗	✗	✗	✗
GenAI-Bench (Li et al., 2024)	✗	✗	✓	✗	✗	✗	✗	✗
ConceptMix (Wu et al., 2024)	✗	✓	✗	✗	✓	✗	✗	✗
Commonsense-T2I (Fu et al., 2024)	✓	✗	✗	✗	✗	✗	✗	✗
WISE (Niu et al., 2025)	✓	✗	✗	✗	✗	✗	✗	✗
R2I-Bench (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: **Comparison between R2I-Bench and existing text-to-image benchmarks.** **R2I-Bench** covers a broader spectrum of essential reasoning capabilities for text-to-image generation. In addition, **R2I-Bench** provides manually curated, high-quality evaluation criteria to support rigorous and consistent assessment.

autoregressive transformer (Sun et al., 2024a; Zhang et al., 2024; Chen et al., 2025; Wang et al., 2024; Chen et al., 2024), or unified architectures(Xiao et al., 2024; Xie et al., 2024; Zhou et al., 2024; Tong et al., 2024; Sun et al., 2023, 2024b), primarily focus on *semantic rendering*, where the prompt explicitly specifies what to generate and the model simply converts it into an image. Although recent work (Jiang et al., 2025; Guo et al., 2025; Liao et al., 2025) has begun to benchmark and enhance reasoning-driven T2I generation, they are often limited to narrow domains such as commonsense (Niu et al., 2025), numerical reasoning (Ghosh et al., 2023), or concept mixing (Wu et al., 2024). Furthermore, widely adopted evaluation metrics for T2I generation, such as CLIPScore (Hessel et al., 2021), VQAScore (Lin et al., 2024), and WIScore (Niu et al., 2025), mainly assess the semantic alignment between generated images and prompts or fail to generalize across diverse reasoning types, limiting meaningful development, comparison, and assessment of the underlying reasoning capabilities in T2I generation models.

To bridge these gaps, we introduce **R2I-Bench** (**R**easoning-**to**-**I**mage **Bench**mark), a comprehensive benchmark consisting of $3,068$ meticulously curated text prompts, specifically designed to evaluate the reasoning capabilities of T2I models. Each prompt is initially generated using a state-of-the-art large language model (i.e., GPT-4o) and subsequently validated and refined by domain experts to ensure the quality and reliability. As shown in Figure 1, **R2I-Bench** encompasses 7 core reasoning categories, including ***commonsense***, ***compositional***, ***logical***, ***mathematical***, ***causal***, ***numerical***, and ***concept mixing***, which are further subdivided into 32 fine-grained reasoning subcategories. In contrast to prior T2I evaluation datasets, **R2I-Bench** offers significantly broader and more systematic coverage of diverse reasoning skills, as summarized in Table 1.

To enable fine-grained evaluation of reasoning-driven T2I generation, each T2I prompt in **R2I-Bench** is paired with a set of instance-specific diagnostic questions and corresponding scoring criteria, all verified by human experts. These questions assess the quality of T2I generation along three critical aspects: (1) text-image alignment, (2) reasoning accuracy, and (3) image quality. Building on these evaluation questions and criteria, we introduce a QA-style metric, **R2I-Score**, which aggregates scores using a weighted scheme. **R2I-Score** demonstrates strong alignment with human judgments, offering a more faithful and interpretable performance measure of T2I models on **R2I-Bench**.

We systematically evaluate 16 representative T2I models on **R2I-Bench**, spanning diffusion-based, autoregressive, reasoning-enhanced, and closed-source models. To further explore the upper bound of reasoning-driven T2I generation, we also develop a strong pipeline-based framework that decouples reasoning and generation: a state-of-the-art LLM (GPT-4o) first performs reasoning over the prompt and rewrites it into a detailed description, which is then rendered by a high-performing T2I model (SD3-medium). Experimental results reveal several key insights: **(1)** All the open-source models achieve less than 45% accuracy, demonstrating limited reasoning capabilities in existing T2I models and underscoring the significance of **R2I-Bench** as a rigorous evaluation benchmark. Notably, these models tend to interpret prompts as bags of words, e.g., they generate both objects for the prompt "*either a spoon or a bowl*", disregarding the logical disjunction; **(2)** Mathematical reasoning remains a persistent challenge across all models, largely due to the lack of diverse, high-quality training data grounded in mathematical concepts and their visual representations; **(3)** Recent efforts to enhance reasoning through Chain of Thought (CoT) or Reinforcement Learning (RL) (Guo et al., 2025; Liao et al., 2025; Jiang et al., 2025) yield marginal improvements, highlighting the need for more robust, fundamentally reasoning-aware T2I models; and **(4)** While the pipeline-based framework improves performance, it still struggles with abstract mathematical reasoning and accurately interpreting specific linguistic constructs such as quantities, limiters, and quantifiers. Finally, we also conduct a comprehensive qualitative error analysis, categorizing model failures into three main categories, including reasoning errors, visual element errors, and image quality degradation, providing valuable insights to future research.

Our contributions are summarized as follows: (1) We introduce **R2I-Bench**, the first comprehensive benchmark specifically designed to evaluate reasoning-driven T2I generation. Covering a broad range of reasoning categories and meticulously curated through a rigorous human-in-the-loop process, **R2I-Bench** offers a valuable resource for benchmarking and advancing

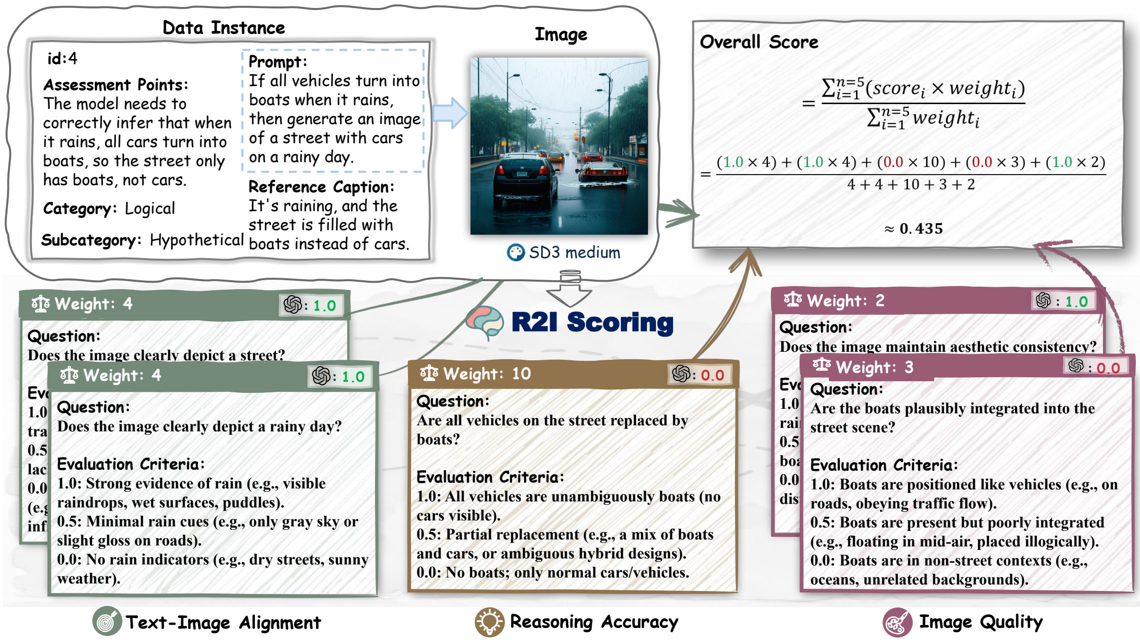

Figure 2: Example Illustration of **R2I-Bench** and **R2I-Score**.

T2I models. (2) To enable fine-grained evaluation of reasoning-driven T2I generation, we design a new metric, **R2I-Score**, built on human-validated evaluation questions and scoring criteria tailored to each data instance in **R2I-Bench**. **R2I-Score** assesses model performance across three critical dimensions, including text-image alignment, reasoning accuracy, and image quality. (3) Through extensive experiments and analysis, we identify several key limitations in all the existing T2I models and provide valuable insights for future research.

2 Related Work

Text-to-Image Generation Models Recent advances in text-to-image (T2I) generation have produced high-quality models across various architectures, including diffusion (Esser et al., 2024a; Gao et al., 2024; Xie et al., 2025a; Qin et al., 2025b; Yang et al., 2024), autoregressive (Sun et al., 2024a; Zhang et al., 2024; Chen et al., 2025; Wang et al., 2024), and unified frameworks (Xiao et al., 2024; Xie et al., 2024; Zhou et al., 2024; Tong et al., 2024; Sun et al., 2023, 2024b). More recently, reasoning-augmented models have incorporated chain-of-thought (CoT) reasoning (Liao et al., 2025) and reinforcement learning (Guo et al., 2025; Jiang et al., 2025) to better handle complex prompts. However, their reasoning capability remains underdeveloped and insufficiently evaluated.

Text-to-Image Evaluation Benchmarks and Metrics Existing T2I benchmarks evaluate isolated reasoning skills but lack comprehensive coverage. OK-VQA (Marino et al., 2019), WISE (Niu et al., 2025), and Commonsense T2I (Fu et al., 2024) emphasize shallow or knowledge-based reasoning, while GeckoNum (Kajić et al., 2024) focuses solely on numerical tasks. Benchmarks like Winoground (Thrush et al., 2022), GenEval (Ghosh et al., 2023), and GenAI-Bench (Li et al., 2024) target compositionally.

Despite progress, no existing benchmark offers a unified framework for evaluating the full spectrum of T2I reasoning abilities (see Table 1). Current evaluation metrics also lack reasoning sensitivity. CLIPScore (Hessel et al., 2021), DSGScore (Cho et al., 2023), and VQAScore (Lin et al., 2024) underperform on complex reasoning and struggle with compositional or numerical fidelity. LLM-based metrics such as LLMScore (Lu et al., 2023) and SemVarEffect (Zhu et al., 2024) overlook spatial or relational accuracy. While RIScore (Zhao et al., 2025) and WIScore (Niu et al., 2025) offer GPT-based scoring, they lack the granularity needed for fine-grained evaluation. Thus, a critical gap remains in metrics that rigorously assess reasoning in T2I generation.

3 R2I-Bench

Overview As shown in the top left part of Figure 2, each data instance in **R2I-Bench** consists of four elements: (1) a reasoning-based T2I prompt which serves as a textual input to the T2I models; (2) a reference caption that explicitly describes the content of the image that is supposed to be generated; (3) an explanation description, which explains the reasoning steps from the T2I prompt to the reference caption and is used to generate reasoning-driven evaluation questions; and (4) the category, the subcategory, and the index of the data instance. As illustrated in Appendix A.5 Figure 6, we adopt a human-in-the-loop pipeline to construct **R2I-Bench**, which comprises three main stages: (1) data collection, (2) data filtering, and (3) evaluation criteria generation.

Data Collection In the initial stage, a team of five human experts systematically reviews prior work relevant to text-to-image (T2I) reasoning tasks (Wu et al., 2024; Kajić et al., 2024; Thrush et al., 2022; Li et al., 2024; Liu et al., 2020; Liew et al., 2022; Fu et al., 2024; Chevalley et al., 2022; Niu et al., 2025; Lei et al., 2025; Lee et al., 2023). Based on this comprehensive analysis, they identify 7 core reasoning categories frequently required across diverse T2I scenarios: ***commonsense***, ***compositional***, ***logical***, ***concept-mixing***, ***numerical***, ***mathematical***, and ***causal reasoning***. These primary categories are further refined into 32 fine-grained subcategories, as illustrated in Figure 3. Detailed definitions for all the core and fine-grained reasoning categories are provided in Appendix A.3.

For each subcategory, we instruct GPT-4o to generate 100-120 T2I prompts designed to test the corresponding reasoning skill, accompanied by reference captions for subsequent evaluation. To ensure that the prompts emphasize reasoning and avoid direct visual descriptions, the generation instruction is constrained by two key guidelines: (1) prompts must not explicitly reveal the answer or directly describe visual features, and (2) the corresponding visual elements must be uniquely identifiable. In-context learning is used, where the model is conditioned on three positive and three negative exemplar prompts authored by human experts. For each generated instance, based on the T2I prompt and reference caption, we further instruct GPT-4o to generate an explanation description which will be leveraged to generate evaluation criteria in the later stage. The instructions for generating the T2I prompt and explanation description are shown in Appendix B.1.1 and B.1.2 respectively.

Data Filtering and Refinement To ensure the quality and validity of the collected data instances, we conduct manual filtering to exclude instances where the prompt fails to yield a renderable image or the associated visual elements are not uniquely identifiable. This filtering step yields approximately 800 high-quality instances from the initial set of 3,200. To expand the dataset while preserving both diversity and quality, we treat these 800 instances as seed T2I prompts. Human experts then engage in an iterative refinement process with GPT-4o, prompting the model to generate additional candidates. After each generation round, human experts evaluate the generated T2I prompts and provide targeted feedback to guide revisions, ensuring that each prompt adheres to the two guidelines. This iterative augmentation continues until each reasoning subcategory reaches approximately 100 validated instances.

Evaluation Criteria Generation and `R2I-Score` Existing T2I evaluation metrics often fail to adequately assess the reasoning abilities essential for high-quality image generation. Hence, we create an evaluation set (i.e., a set of evaluation questions and their corresponding scoring criteria) tailored to each data instance in **`R2I-Bench`**. The carefully designed evaluation questions assess the T2I models in three core dimensions: ① *Text-image alignment*: whether the generated image accurately contains all required elements, such as objects and attributes; ② *Reasoning accuracy*: whether the T2I model performs necessary reasoning over the input prompt to correctly generate the output image; ③ *Image quality*: measuring the clarity and quality (e.g., vagueness, distortions, and so on) of the generated images. Example questions for each evaluation dimension are provided in Figure 2.

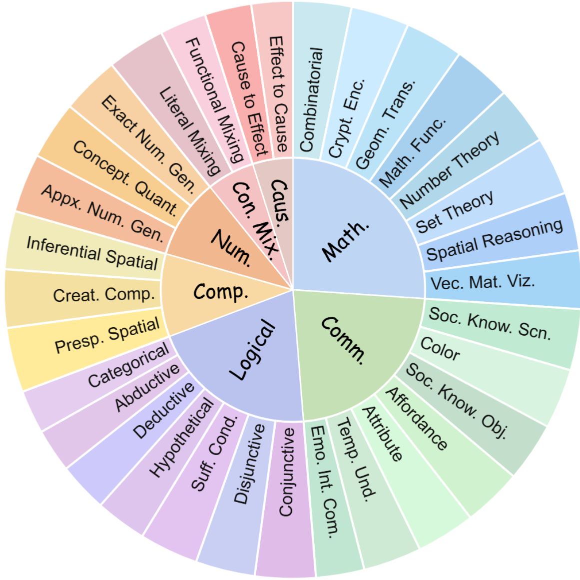

Figure 3: **Distribution of Diverse Reasoning Categories in `R2I-Bench`.** Caus.: Causal. Con. Mix.: Concept Mixing. Math.: Mathematical. Comm.: Commonsense. Num.: Numerical. Comp.: Compositional.

For efficient, we feed each previously generated T2I prompt, the corresponding reference caption, and explanation description to GPT-4o and ask it to generate a set of evaluation questions, each paired with an assigned evaluation dimension, an importance weight, and a scoring criterion. To further emphasize reasoning over surface-level features, we manually set a weight constraint range for each question based on its evaluation dimension: $[7, 10]$ for *reasoning accuracy*, $[4, 6]$ for *text-image alignment*, and $[1, 3]$ for *image quality*. This design reflects our goal of benchmarking reasoning-driven T2I generation, under the assumption that most modern T2I models already perform well in producing visually appealing images. To ensure reliability and consistency, all evaluation questions, scoring criteria, and importance weights are manually validated and refined by expert annotators. The complete instruction template in this process are provided in Appendix B.1.4.

Building on the evaluation set, we propose a new QA-style metric, **`R2I-Score`**. Given a generated image for a T2I prompt, we feed the image along with each evaluation question and its corresponding scoring criteria as input to GPT-4o, and ask it to select a score s_i based on the provided criteria. This scoring instruction template is detailed in Appendix B.1.3. We calculate

Model	Size	Overall	Commonsense	Compositional	Con.Mix	Logical	Numerical	Mathematical	Causal
Diffusion Models									
SD3-medium	2B	0.45	0.54	0.64	0.63	0.55	0.50	0.19	0.18
Lumina-Image 2.0	2.6B	0.42	0.49	0.65	0.54	0.56	0.43	0.13	0.40
Sana-1.5	4.8B	0.41	0.49	0.67	0.66	0.49	0.48	0.13	0.21
Lumina-T2I	5B	0.33	0.38	0.49	0.55	0.38	0.45	0.13	0.18
Omnigen	3.8B	0.40	0.43	0.60	0.43	0.51	0.47	0.18	0.34
LLM4GEN$_{SD1.5}$	0.86B	0.40	0.55	0.48	0.60	0.55	0.39	0.07	0.45
ELLA$_{SD1.5}$	0.07B	0.31	0.40	0.44	0.40	0.40	0.32	0.07	0.29
AutoRegressive Models									
EMU3	8.0B	0.41	0.44	0.59	0.62	0.55	0.61	0.09	0.41
Janus-Pro-7B	7B	0.38	0.45	0.60	0.64	0.46	0.46	0.07	0.36
LlamaGen	0.8B	0.29	0.38	0.39	0.49	0.38	0.35	0.07	0.12
Show-o	1.3B	0.36	0.42	0.59	0.56	0.42	0.57	0.12	0.30
Reasoning-Enhanced Models									
Show-o+ORM	1.3B	0.34	0.42	0.45	0.44	0.37	0.49	0.12	0.26
Show-o+DPO	1.3B	0.36	0.43	0.47	0.48	0.41	0.51	0.13	0.31
Show-o+PARM	1.3B	0.38	0.45	0.49	0.51	0.45	0.56	0.13	0.32
Close Source Models									
DALL-E-3	-	0.71	0.78	0.76	0.86	0.69	0.69	0.21	0.64
gpt-image-1	-	0.77	0.83	0.87	0.89	0.81	0.88	0.58	0.71
Prompt-Rewrite Pipeline									
gpt-4o+SD3-medium	2B	0.58$_{\uparrow 0.13}$	0.75$_{\uparrow 0.21}$	0.75$_{\uparrow 0.11}$	0.81$_{\uparrow 0.18}$	0.65$_{\uparrow 0.10}$	0.63$_{\uparrow 0.13}$	0.22$_{\uparrow 0.03}$	0.76$_{\uparrow 0.58}$

Table 2: **Evaluation on `R2I-Bench`.** The highest accuracy for closed-source and open-source text-to-image models are marked in red and blue respectively. Con.Mix.: Concept Mixing.

`R2I-Score` as follows:

$$\text{R2I-Score} = \frac{\sum_{i=1}^{n} w_i \cdot s_i}{\sum_{i=1}^{n} w_i} \tag{1}$$

where n is the total number of evaluation questions for a given instance, and w_i is the importance weight assigned to the ith evaluation question.

Dataset Statistics Finally, **`R2I-Bench`** comprises $3,068$ high-quality, reason-driven T2I prompts. Figure 1 includes an example T2I prompt for each core reasoning category. Figure 3 presents the distribution of these categories in **`R2I-Bench`**, and Table 4 in Appendix A.1 provides detailed statistics for **`R2I-Bench`**.

4 Experiments

4.1 Experimental Setup

To conduct evaluation on **`R2I-Bench`**, we carefully select 16 representative, high-performing T2I models with publicly available model checkpoints, spanning four distinct categories:

(1) *Diffusion Models*, featuring models including SD3-medium (Rombach et al., 2022), Lumina-Image 2.0 (Qin et al., 2025b), Sana-1.5 (Xie et al., 2025a), Lumina-T2I (Qin et al., 2025a), Omnigen (Xiao et al., 2024), LLM4GEN$_{SD1.5}$ (Liu et al., 2025), and ELLA$_{SD1.5}$ (Hu et al., 2024); (2) *Autoregressive Models*, including EMU3 (Wang et al., 2024), Janus-Pro-7B (Chen et al., 2025), LlamaGen (Sun et al., 2024a), and Show-o (Xie et al., 2024); (3) *Reasoning-Enhanced Models*, including Show-o+ORM, Show-o+DPO, and Show-o+PARM (Guo et al., 2025) ; and (4) *Closed-Source Models*, including DALL-E-3 (Ma et al., 2024) and gpt-image-1 (Hurst et al., 2024). Additional implementation details, such as model architectures, configurations, and inference parameters, are provided in Appendix B.3. For evaluation, we adopt the proposed **`R2I-Score`** metric.

Intuitively, reasoning-driven T2I generation could be more effectively addressed by decoupling reasoning from image generation—first leveraging a large language model to perform complex reasoning and generate a detailed textual description, and then using a powerful image generation model to render the final image (Niu et al., 2025). Motivated by this, we design a strong pipeline-based framework that explicitly separates the reasoning and generation stages.

The framework first employs a state-of-the-art LLM (`GPT-4o`) to interpret and reason over the original prompt, producing a detailed and structured image description. This rewritten prompt is then passed to a high-performing T2I model (`SD3-medium`) to generate the corresponding image. We name this pipelined framework as `gpt-4o+SD3-medium`.

4.2 Main Results

Table 2 presents the evaluation results of all T2I models across the core reasoning categories in **`R2I-Bench`**, with detailed subcategory-level results provided in Appendix C. The main findings are summarized as follows:

T2I Models Show Limited Capability in Reasoning-Driven Image Generation. Our evaluation reveals that most open-source models achieve a score lower than 45% based on **R2I-Score**, suggesting a notable gap in their ability to handle reasoning-driven T2I prompts. This limitation appears to stem from a shallow understanding of prompts, often interpreted as a bag of words rather than through compositional or logical reasoning. This hypothesis is further supported by our qualitative error analysis, illustrated in Appendix A.2, Figures 8 through 13, where the majority of models simply generate images that merely reflect the objects explicitly mentioned in the prompt without performing necessary inferential reasoning.

For instance, given the prompt "*a cat-like bed*" (Figure 11), most of the models, including EMU3, SD3-medium, ELLA, and PARM+Show-o, just naively depict a cat and a bed as distinct, unrelated objects. Similarly, in tasks involving logical operations or quantifiers such as the prompt "*either a spoon or a bowl*" (Figure 9), most models incorrectly render both objects, reflecting an inability to correctly interpret disjunctive semantics.

We hypothesize that these limitations are rooted in the bag-of-words encoding mechanism used by CLIP-based conditioning in diffusion models. A formal investigation of this hypothesis is left for future work.

Mathematical Reasoning Remains a Significant Bottleneck. Across all reasoning categories, T2I models exhibit profound limitations in addressing mathematical reasoning tasks. Most models achieve near-zero accuracy on this front. Notably, even the best-performing open-source model, SD3-medium, attains a score of merely 0.19, while others, including LlamaGen, Show-o, and $\text{ELLA}_{\text{SD1.5}}$, score below 0.10. As shown in Figure 10, prompts involving geometric transformations (e.g., "*rotate a square 90°*") frequently result in irrelevant outputs such as abstract art or clocks. Similarly, prompts grounded in number theory (e.g., "*visualize the twin prime pairs below 50*") yield outputs like mecha robots (EMU3) or glowing, non-descriptive artifacts (Show-o+PARM, Show-o). These observations indicate a severe lack of training data containing mathematical visual concepts, hindering the models' ability to perform reliable numerical or mathematical reasoning.

Marginal Improvements from Reasoning-Enhanced Architectures. Reasoning-enhanced models such as Show-o+PARM, Show-o+ORM, and Show-o+DPO demonstrate only incremental improvements over their respective base models. For example, the best-performing variant (i.e., Show-o+PARM) achieves an overall score of 0.38, compared to 0.36 achieved by the base model Show-o. Notably, these models continue to perform poorly on the most challenging categories, including mathematical reasoning ($\leq$ 0.13) and causal reasoning ($\leq$ 0.32), indicating that current methods, such as PARM (Potential Assessment Reward Model), ORM (Outcome Reward Model), and DPO (Direct Preference Optimization), offer limited improvements in reasoning-driven T2I generation. These results highlight the urgent need for more effective and targeted approaches for reasoning-driven T2I generation.

Closed-Source Models Set the Upper Bound for Current Reasoning Capabilities. Proprietary models such as DALLE-3 and gpt-image-1 significantly outperform their open-source counterparts, achieving 57.8% and 71.1% higher score than the best-performing open-source model (i.e., SD3-medium), respectively. Notably, gpt-image-1 consistently achieves the highest scores across all reasoning categories. This significant performance gap highlights the pressing need for open, reproducible benchmarks and the development of competitive open-source T2I models to bridge the capability gap with proprietary systems.

Pipeline-based T2I Framework Improves Commonsense, Causal Reasoning, but Yields Marginal Gains for Compositional, Numerical and Mathematical Reasoning. As shown in Table 2, the pipeline-based framework yields substantial gains in all reasoning categories by an average of 0.13, e.g., improvements ranging from 0.21 to 0.58 are observed in causal reasoning, commonsense reasoning. A detailed comparison across fine-grained reasoning subcategories is shown in Figure 21. Despite the general effectiveness of the pipeline-based framework, gains in *Compositional*, *Mathematical* categories remain modest ($\leq$ 0.13). As shown in Figure 4, many reasoning concepts remain challenging for T2I models to faithfully render, even when clearly articulated by the LLM. In *Compositional reasoning* (Example 1), despite GPT-4o correctly reasons that "*a surreal waterfall cascades upwards from the base*," SD3-medium still renders a downward-flowing waterfall. In *Numerical reasoning* (Example 2), although GPT-4o expands the original prompt "*Nineteen brown dogs and seventeen grey dogs*" with additional detail, the generated image fails to depict the correct number of dogs. For *Mathematical reasoning* (Example 3), the difficulty goes beyond language to abstract cognition: although GPT-4o specifies terms like "*display all three stages*" and "*regular pentagon*," the output remains visually inaccurate, with SD3-medium producing disorganized geometric shapes.

Success in this domain often requires models to grasp geometric structures such as points, lines, angles, and spatial transformations. We posit that overcoming these limitations will require not only more mathematically enriched training data but also the integration of architectural components or external modules capable of reasoning over structured symbolic knowledge.

4.3 Evaluation of R2I-Score

We further assess the effectiveness of our proposed **R2I-Score** by evaluating its alignment with human judgments. We conduct a human study involving a group of senior college students, where each participant compares the image outputs generated by two T2I models, Lumina-Image 2.0 (Qin et al., 2025b)

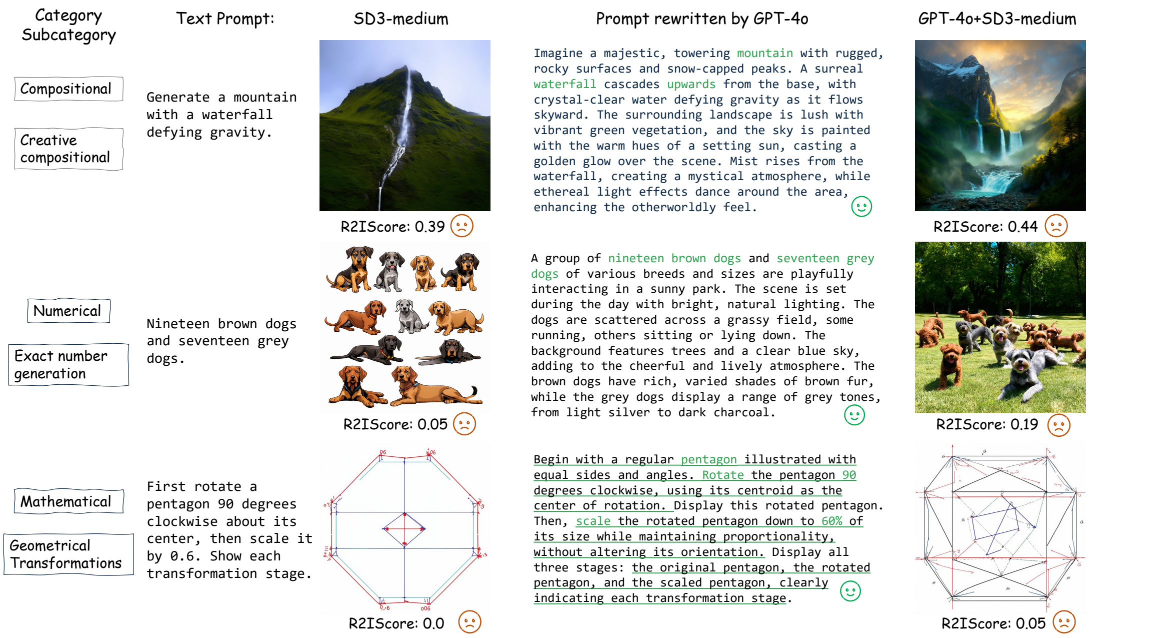

Figure 4: Failure Cases of the Pipeline-based Framework on Compositional/Numerical/Mathematical Reasoning.

and Sana-1.5 (Xie et al., 2025a), and selects the image that best aligns with the prompts or indicates if both are equally satisfactory or unsatisfactory. More details are provided in Appendix B.2. We also apply **`R2I-Score`** to evaluate the same set of image pairs and compute its judgements with those of human annotators, using three established evaluation metrics: *Pairwise Accuracy* (Deutsch et al., 2023), *Kendall's* τ (Jadhav and Ma, 2019), and *Spearman Correlation* (Tu et al., 2025). We compare **`R2I-Score`** against several widely adopted T2I generation evaluation metrics, including **DSGscore** (Cho et al., 2023), **VIEScore** (Ku et al., 2023), **CLIPScore** (Hessel et al., 2021), and **VQA score** (Lin et al., 2024). Since these existing metrics mainly focus on surface-level text-image alignment and image quality, **`R2I-Score`** consistently achieves superior alignment with human judgements across all alignment criteria, as shown in Table 3, demonstrating its effectiveness and robustness as an evaluation metric of reasoning-driven T2I generation. Further experimental details and additional results are provided in Appendix C.3.

Models	Pairwise Accuracy	Kendall's τ	Spearman Correlation
CLIPScore	0.631	0.263	0.310
DSGScore	0.520	0.220	0.254
VIEScore	0.694	0.494	0.451
VQAscore	0.629	0.463	0.563
`R2I-Score`	**0.713**	**0.747**	**0.694**

Table 3: Comparison of **`R2I-Score`** with other Evaluation Metrics for T2I Generation.

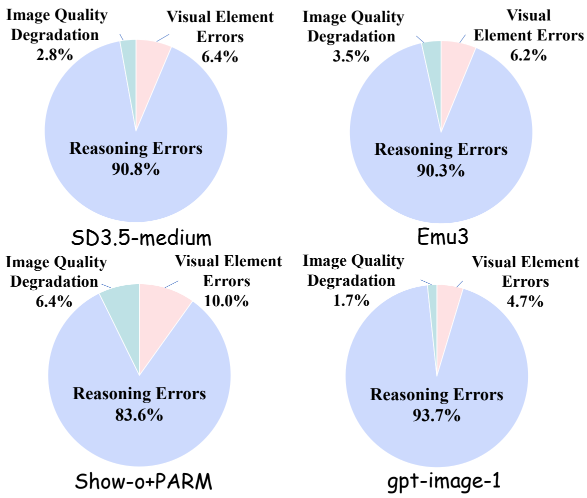

Figure 5: Distribution of Errors of `Emu3`, `SD3-medium`, `Show-o+PARM`, `gpt-image-1`.

4.4 Error Analysis

To better understand the limitations of current T2I models, we categorize and accordingly define three failure types: *basic element errors*, *reasoning errors*, and *visual quality issues*. For qualitative analysis, we examine representative models from each architectural category, including `Emu3`, `SD3-medium`, `Show-o+PARM`, and `gpt-image-1`. The relative distribution of these failure types is computed and visualized in Figure 5. As we can see, reasoning-related failures dominate the error distribution across all four models, accounting for over 80% of total errors. This observation highlights reasoning as the primary bottleneck in current T2I systems. Among the evaluated models, `Show-o+PARM` exhibits a relatively higher proportion of basic element errors, suggesting its limitation in accurately rendering basic visual components. In contrast, `gpt-image-1` demonstrates the lowest rates of both basic element and image quality errors, indicating its superior performance in both semantic fidelity and visual rendering.

5 Conclusion

This paper introduces **R2I-Bench**, a comprehensive benchmark designed to evaluate the reasoning capabilities of text-to-image (T2I) generation models across 7 core reasoning categories and 32 subcategories. Alongside **R2I-Bench**, we design **R2I-Score**, a QA-style evaluation metric specifically tailored for reasoning-driven T2I generation, with stronger correlation with human judgments compared to existing evaluation metrics. Our evaluation reveals consistently limited reasoning capabilities across all existing T2I models, highlighting the pressing need for more robust, reasoning-aware T2I generation architectures.

Limitations

Evaluation Method Constraints Despite our diligent efforts to design and refine evaluation questions and criteria for each data instance, aimed at enhancing reasoning-based evaluation, the current method is inherently constrained by the specific benchmark used in this study. As such, it cannot be directly generalized to other datasets without further adaptations. Although the manually crafted evaluation questions and criteria facilitate the use of vision language models for scoring, leading to more transparent and interpretable evaluations, the granularity of these evaluations remains relatively coarse compared to the detailed assessments conducted at the training level. Future work could focus on the development of a versatile reward model tailored for evaluating Text-to-Image (T2I) reasoning generation, which would also support reinforcement learning from Human Feedback (RLHF).

Language and Dataset Scope At present, our evaluation of T2I models is confined to **R2I-Bench**, which is based solely on English-language data. Consequently, the reasoning capabilities of models in non-English language contexts remain unexplored. Additionally, some models do not support symbolic inputs, such as emojis or complex mathematical notations. For the sake of ensuring the benchmark's general applicability, we have excluded data instances that feature such symbolic inputs. Besides, our benchmark is limited only to image generation. Extending to video/audio/3D generation can be another promising future direction.

Ethics Statement

Some instances in our dataset were generated using GPT-4o, a powerful language model that has been designed to simulate human-like text generation. Although this model produces high-quality outputs, it is important to note that the generated content reflects the biases and limitations inherent in the training data. We are aware of the ethical implications of using such models, especially in terms of the potential for reinforcing harmful stereotypes or generating inappropriate content. In this study, we have made efforts to mitigate these risks by carefully curating the dataset and implementing a manual review process. However, we acknowledge that there may still be residual biases present, and we encourage future work to focus on developing methods to reduce such biases, ensuring that generated content aligns with ethical guidelines and societal norms.

References

Liang Chen, Sinan Tan, Zefan Cai, Weichu Xie, Haozhe Zhao, Yichi Zhang, Junyang Lin, Jinze Bai, Tianyu Liu, and Baobao Chang. 2024. A spark of vision-language intelligence: 2-dimensional autoregressive transformer for efficient finegrained image generation. In *The Thirteenth International Conference on Learning Representations*.

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

Mathieu Chevalley, Yusuf Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. 2022. Causalbench: A large-scale benchmark for network inference from single-cell perturbation data. *arXiv preprint arXiv:2210.17283*.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2023. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *arXiv preprint arXiv:2310.18235*.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. *arXiv preprint arXiv:2305.14324*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024a. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024b. Scaling rectified flow transformers for high-resolution image synthesis.

Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. 2024. Commonsense-t2i challenge: Can text-to-image generation models understand commonsense? *arXiv preprint arXiv:2406.07546*.

Peng Gao, Le Zhuo, Dongyang Liu, Ruoyi Du, Xu Luo, Longtian Qiu, Yuhang Zhang, Chen Lin, Rongjie Huang, Shijie Geng, et al. 2024. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*.

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

Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. 2025. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*.

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

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*.

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

Sneha Jadhav and Shuangge Ma. 2019. Kendall's tau for functional data analysis. *arXiv preprint arXiv:1912.03725*.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. 2025. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*.

Ivana Kajić, Olivia Wiles, Isabela Albuquerque, Matthias Bauer, Su Wang, Jordi Pont-Tuset, and Aida Nematzadeh. 2024. Evaluating numerical reasoning in text-to-image models. *Advances in Neural Information Processing Systems*, 37:42211–42224.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2023. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2024. Viescore: Towards explainable metrics for conditional image synthesis evaluation.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. 2023. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36:69981–70011.

Jiayi Lei, Renrui Zhang, Xiangfei Hu, Weifeng Lin, Zhen Li, Wenjian Sun, Ruoyi Du, Le Zhuo, Zhongyu Li, Xinyue Li, et al. 2025. Imagine-e: Image generation intelligence evaluation of state-of-the-art text-to-image models. *arXiv preprint arXiv:2501.13920*.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. 2024. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*.

Jiaqi Liao, Zhengyuan Yang, Linjie Li, Dianqi Li, Kevin Lin, Yu Cheng, and Lijuan Wang. 2025. Imagegen-cot: Enhancing text-to-image in-context learning with chain-of-thought reasoning. *arXiv preprint arXiv:2503.19312*.

Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. 2022. Magicmix: Semantic mixing with diffusion models. *arXiv preprint arXiv:2210.16056*.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Mushui Liu, Yuhang Ma, Zhen Yang, Jun Dan, Yunlong Yu, Zeng Zhao, Zhipeng Hu, Bai Liu, and Changjie Fan. 2025. Llm4gen: Leveraging semantic representation of llms for text-to-image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5523–5531.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2023. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems*, 36:23075–23093.

Shuailei Ma, Kecheng Zheng, Ying Wei, Wei Wu, Fan Lu, Yifei Zhang, Chen-Wei Xie, Biao Gong, Jiapeng Zhu, and Yujun Shen. 2024. Learning visual generative priors without text. *arXiv preprint arXiv:2412.07767*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

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

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Xinyue Li, Dongyang Liu, Xiangyang Zhu, Will Beddow, Erwann Millon, Wenhai Wang Victor Perez, Yu Qiao, Bo Zhang, Xiaohong Liu, Hongsheng Li, Chang Xu, and Peng Gao. 2025a. Lumina-image 2.0: A unified and efficient image generative framework.

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, et al. 2025b. Lumina-image 2.0: A unified and efficient image generative framework. *arXiv preprint arXiv:2503.21758*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. 2024a. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024b. Generative multimodal models are in-context learners.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

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

Shengxin Tu, Chun Li, and Bryan E Shepherd. 2025. Between-and within-cluster spearman rank correlations. *Statistics in Medicine*, 44(3-4):e10326.

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

Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. 2024. Conceptmix: A compositional image generation benchmark with controllable difficulty. *arXiv preprint arXiv:2408.14339*.

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

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, and Song Han. 2025a. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer.

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, et al. 2025b. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*.

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

Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, and Bin Cui. 2024. Improving diffusion-based image synthesis with context prediction.

Qian Zhang, Xiangzi Dai, Ninghua Yang, Xiang An, Ziyong Feng, and Xingyu Ren. 2024. Var-clip: Text-to-image generator with visual auto-regressive modeling. *arXiv preprint arXiv:2408.01181*.

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

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

Xiangru Zhu, Penglei Sun, Yaoxian Song, Yanghua Xiao, Zhixu Li, Chengyu Wang, Jun Huang, Bei Yang, and Xiaoxiao Xu. 2024. Evaluating semantic variation in text-to-image synthesis: A causal perspective. *arXiv preprint arXiv:2410.10291*.

A More Details about R2I-Bench

A.1 Statistics of R2I-Bench

We list the statistics of **R2I-Bench** in Table 4.

Statistic	Number
Total data instances	3,068
- Commonsense	695 (22.65%)
- Compositional	311 (10.14%)
- Numerical	322 (10.50%)
- Causal	151 (4.92%)
- Mathematical	800 (26.08%)
- Logical	630 (20.53%)
- Concept Mixing	159 (5.18%)
Categories	7
Subcategories	32
Evaluation dimensions	3
Vocabulary size	7,184
Maximum prompt length	35
Maximum reference caption length	28
Maximum evaluation questions	18
Average prompt length	21.7
Average reference caption length	23.4
Average evaluation questions	12.2

Table 4: **Key Statistics of R2I-Bench.**

A.2 Image by Categories

This section presents examples of images from various categories in **R2I-Bench**. Figure 8 to 12 coresponding to images under the categories of *Commonsense Reasoning, Numerical Reasoning, Causal Reasoning, Logical Reasoning, Mathematical Reasoning, Concept Mixing Reasoning, Compositional Reasoning*, respectively.

A.3 Definition of Categories in R2I-Bench

The data instances in **R2I-Bench** encompass seven core categories: Commonsense Reasoning, Compositional Reasoning, Conceptual Mixing Reasoning, Numerical Reasoning, Logical Reasoning, Causal Reasoning, and Mathematical Reasoning. These categories are further subdivided into thirty-two more granular subcategories, providing a thorough evaluation of the reasoning capabilities of Text-to-Image (T2I) models.

Commonsense Reasoning Commonsense reasoning is a critical aspect of evaluating a model's understanding of general knowledge and contextual information. It involves utilizing external *knowledge resources*—such as *world knowledge*, *cultural context*, or *background information*—to reason about the content of an image, rather than simply replicating the image. This allows for a richer context in assessing the commonsense reasoning capabilities of *Text-to-Image* (T2I) models. In **R2I-Bench**, we categorize commonsense reasoning into seven distinct *subfields*, as shown in Figure 8, with detailed definitions provided in Table 11.

Compositional Reasoning Compositional reasoning refers to the ability to combine smaller, simpler *components* or pieces of *information* to form more complex *concepts*, *solutions*, or *conclusions*. It involves understanding the *relationships* between individual parts and how they contribute to the whole, enabling *logical reasoning* within structured, hierarchical, or layered systems. In **R2I-Bench**, we divide compositional reasoning into three *subfields*, as depicted in Figure 12, with their definitions outlined in Table 8.

Numerical Reasoning Numerical reasoning, in the context of T2I models, involves the ability of these models to accurately interpret, process, and generate *images* based on *numerical information* presented in *textual prompts*. In **R2I-Bench**, we categorize numerical reasoning into three *subfields*, as illustrated in Figure 14, with definitions provided in Table 10.

Concept Mixing Reasoning Concept-Mixing reasoning refers to the process of combining different *semantic elements* to create a new, unique *concept*. In **R2I-Bench**, we divide concept-mixing reasoning into three *subfields*, as shown in Figure 11, with their definitions in Table 6.

Logical Reasoning Logical reasoning involves using systematic, structured approaches to analyze *information*, draw *conclusions*, and solve *problems* based on given *premises* or *conditions*. In **R2I-Bench**, we break logical reasoning down into seven *subfields*, as illustrated in Figure 9, with definitions provided in Table 9.

Mathematical Reasoning Mathematical reasoning refers to the ability to represent, understand, and generate visual representations of abstract *mathematical concepts* and *symbols*. In **R2I-Bench**, we subdivide mathematical reasoning into eight *subfields*, as shown in Figure 10, with their definitions outlined in Table 5.

Causal Reasoning Causal reasoning is the ability to understand and explain *cause-and-effect relationships*. In **R2I-Bench**, we categorize causal reasoning into three *subfields*, as illustrated in Figure 13, with definitions provided in Table 7.

A.4 Definition of Subcategories in R2I-Bench

This section presents definitions of various subcategories under categories in **R2I-Bench**. Table 6 to 5 coresponding to subcategories under the categories of *Commonsense Reasoning, Numerical Reasoning, Causal Reasoning, Logical Reasoning, Mathematical Reasoning, Concept Mixing Reasoning, Compositional Reasoning*, respectively.

A.5 Data Generation Pipeline

We build a human-in-the-loop data generation pipeline as illustrated in Figure 6.

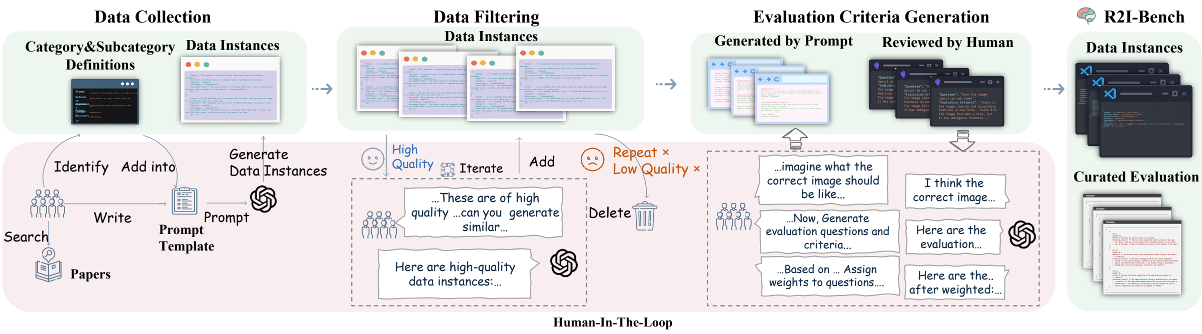

Figure 6: **Benchmark Curation Pipeline.** The pipeline starts with data collection, followed by data filtering, evaluation criteria generation, and ultimately results in **`R2I-Bench`**. To ensure data quality, human verification is performed at each key stage to eliminate low-quality data, annotations, and ambiguous evaluation questions.

Mathematical Reasoning	**Description**
Mathematical Function Visualization (12.50%)	Mathematical Function Visualization involves generating clear and informative images that depict *mathematical functions*, their *properties*, and the *relationships* between various *mathematical entities*, such as *variables* and *parameters*.
Vector & Matrix Visualization (12.50%)	Vector & Matrix Visualization involves understanding and illustrating *vectors*, *matrices*, and *transformations* in *geometrical space*.
Combinatorial Reasoning (12.50%)	Combinatorial Reasoning involves depicting *permutations*, *combinations*, or *arrangements* of *objects*, often within a *geometric* or *graphical context*.
Set Theory & Relations (12.50%)	Set Theory & Relations involves representing *sets*, *subsets*, and their *relations* in visual forms (e.g., using *Venn diagrams* or *set-builder notation*).
Cryptographic & Encoding Reasoning (12.50%)	Cryptographic & Encoding Reasoning involves rendering *encrypted texts*, *ciphers*, or *encoding schemes* (e.g., *Morse code*, *binary representations*).
Number Theory (12.50%)	Number Theory Visualization involves depicting *prime numbers*, *divisibility rules*, and other abstract *mathematical concepts*.
Geometrical Transformations (12.50%)	Geometrical Transformations involves illustrating *symmetry operations* like *rotations*, *reflections*, *translations*, or *dilations* in space.
Spatial Reasoning (12.50%)	Spatial Reasoning refers to the ability to reason and infer the correct *geometric configuration* of *objects*, such as *lines* and *shapes*, in a defined space, based on specified *spatial relationships*.

Table 5: Definitions and proportions of the eight subcategories in mathematical reasoning within **`R2I-Bench`**. The percentage indicates the proportion of each subcategory within the overall mathematical category.

Concept Mixing Reasoning	**Description**
Functional Mixing (44.44%)	Functional mixing includes creating new *concepts* that involve blending different *functional properties* of *objects*.
Literal Mixing (55.56%)	Literal Mixing Reasoning combines *elements* from different *concepts* in a *straightforward*, *literal* manner, such as merging *objects* or *creatures*.

Table 6: Definitions and proportions of the two subcategories in concept mixing reasoning within **`R2I-Bench`**. The percentage indicates the proportion of each subcategory within the overall concept mixing category.

Causal Reasoning	Description
Cause to Effect Reasoning (*52.98%*)	Given a *cause*, generate an *image* depicting the *effect*.
Effect to Cause Reasoning (*47.02%*)	Given an *effect*, generate an *image* depicting the possible *cause*.

Table 7: Definitions and proportions of the two subcategories in causal reasoning within **`R2I-Bench`**. The percentage indicates the proportion of each subcategory within the overall causal category.

Compositional Reasoning	Description
Creative Composition Reasoning (32.15%)	Creative compositional reasoning is the ability to combine different *ideas* or *objects* in *innovative* and *imaginative* ways to create *novel* and *unique scenes* that have not been seen before.
Inferential Spatial Reasoning (32.15%)	Inferential spatial reasoning refers to the ability to determine the *positions* or *size relationships* between *objects* without explicit descriptions.
Prescriptive Spatial Reasoning (35.69%)	Prescriptive Spatial Reasoning refers to the ability to follow clear *instructions* about where *objects* should be placed in a scene, ensuring the layout matches the described *relationships*. Understanding phrases like *"left of", "above", "inside"*.

Table 8: Definitions and proportions of the three subcategories in compositional reasoning within **`R2I-Bench`**. The percentage indicates the proportion of each subcategory within the overall compositional reasoning category.

Logical Reasoning	Description
Categorical Reasoning (11.90%)	Categorical reasoning involves determining whether a specific *concept* belongs to a particular *category*. This type of reasoning often involves *quantifiers* such as *"all,"*, *"everyone,"*, *"any,"*, *"no,"* and *"some."*
Hypothetical Reasoning (11.90%)	Hypothetical reasoning is the process of using a *systematic*, *structured* approach to analyze *information*, draw *conclusions*, and solve *problems* based on given *premises* or *conditions*.
Disjunctive Reasoning (16.51%)	Disjunctive reasoning involves *premises* in the form *"either ... or ..."*, where the *conclusion* holds as long as one *premise* is true.
Conjunctive Reasoning (16.51%)	Conjunctive reasoning involves *premises* in the form *"both ... and ..."*, where the *conclusion* holds only if all the *premises* is true.
Sufficient Conditional Reasoning (13.49%)	Sufficient Conditional Reasoning is based on *conditional statements* of the form *"If P, then Q"*, in which P is the *antecedent* and Q is the *consequent*.
Deductive Reasoning (13.97%)	Deductive reasoning focuses on deriving specific *conclusions* from general *principles* or *premises*, ensuring that *conclusions* logically follow if the *premises* are true.
Abductive Reasoning (16.03%)	Abductive reasoning, considered more *creative* and *open-ended*, involves forming *hypotheses* to explain *observations*, often generating the most *plausible explanation* rather than a *guaranteed conclusion*.

Table 9: Definitions and proportions of the seven subcategories in logical reasoning within **`R2I-Bench`**. The percentage indicates the proportion of each subcategory within the overall logical reasoning category.

Numerical Reasoning	**Description**
Exact Number Generation (31.06%)	Exact number generation examines the model's ability to correctly generate an *exact number* of *objects*.
Approximate Number Generation and Zero (31.37%)	Approximate number generation evaluates models on their ability to correctly depict *entities* with quantities expressed in *approximate terms* by means of *linguistic quantifiers(e.g., "many", "a few", or "more")*.
Conceptual Quantitative Reasoning (37.58%)	Conceptual quantitative reasoning evaluates models on prompts that require a *conceptual understanding* of *objects* and their *parts*.

Table 10: Definitions and proportions of the three subcategories in Numerical reasoning within **R2I-Bench**. The percentage indicates the proportion of each subcategory within the overall Numerical reasoning category.

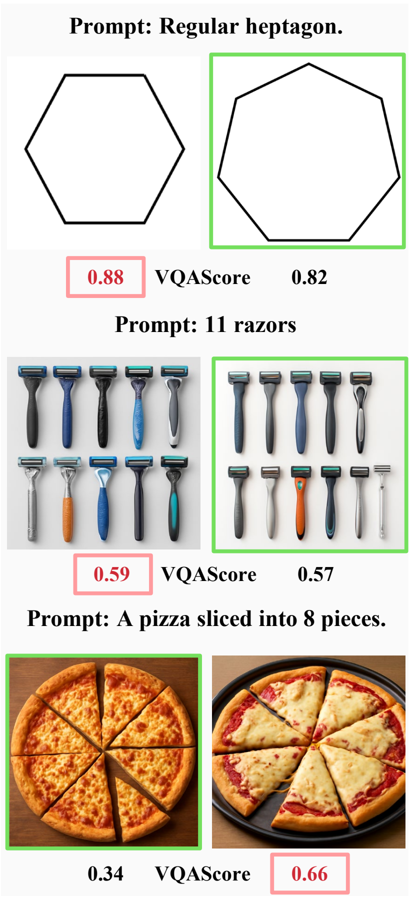

Figure 7: **Failure Cases of VQAScore.**

Figure 8: **Examples of Seven Subfields in Commonsense Reasoning**, spanning Affordance, Attribute, Color, Emotion Intention Commonsense, Social Cultural Knowledge Object and Scene and Temporal Understanding. We showcase the Text-lite version.

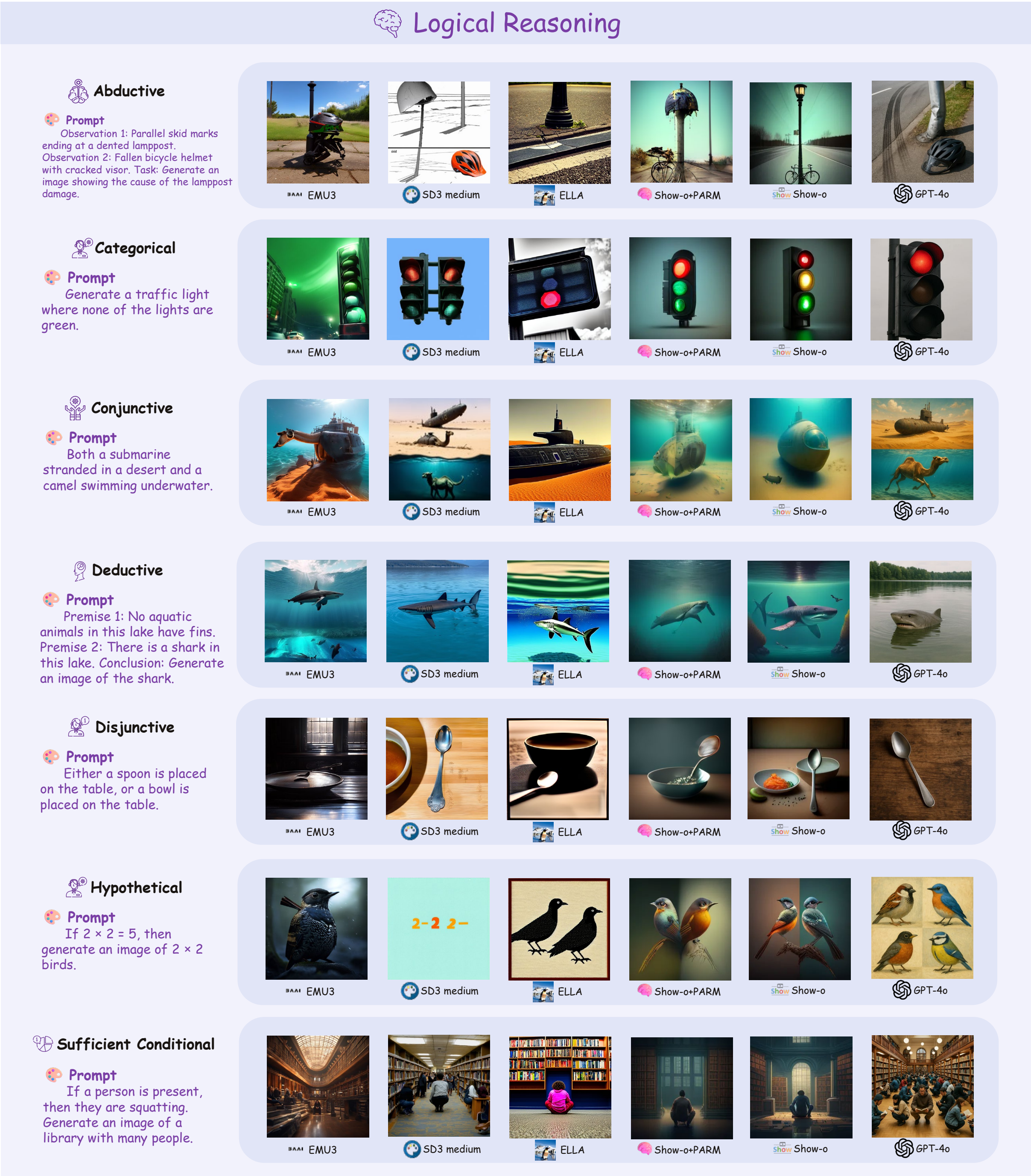

Figure 9: **Examples of Seven Subfields in Logical Reasoning**, spanning Abductive, Categorical, conjunctive, Deductive, Hypothetical, Sufficient Conditional.

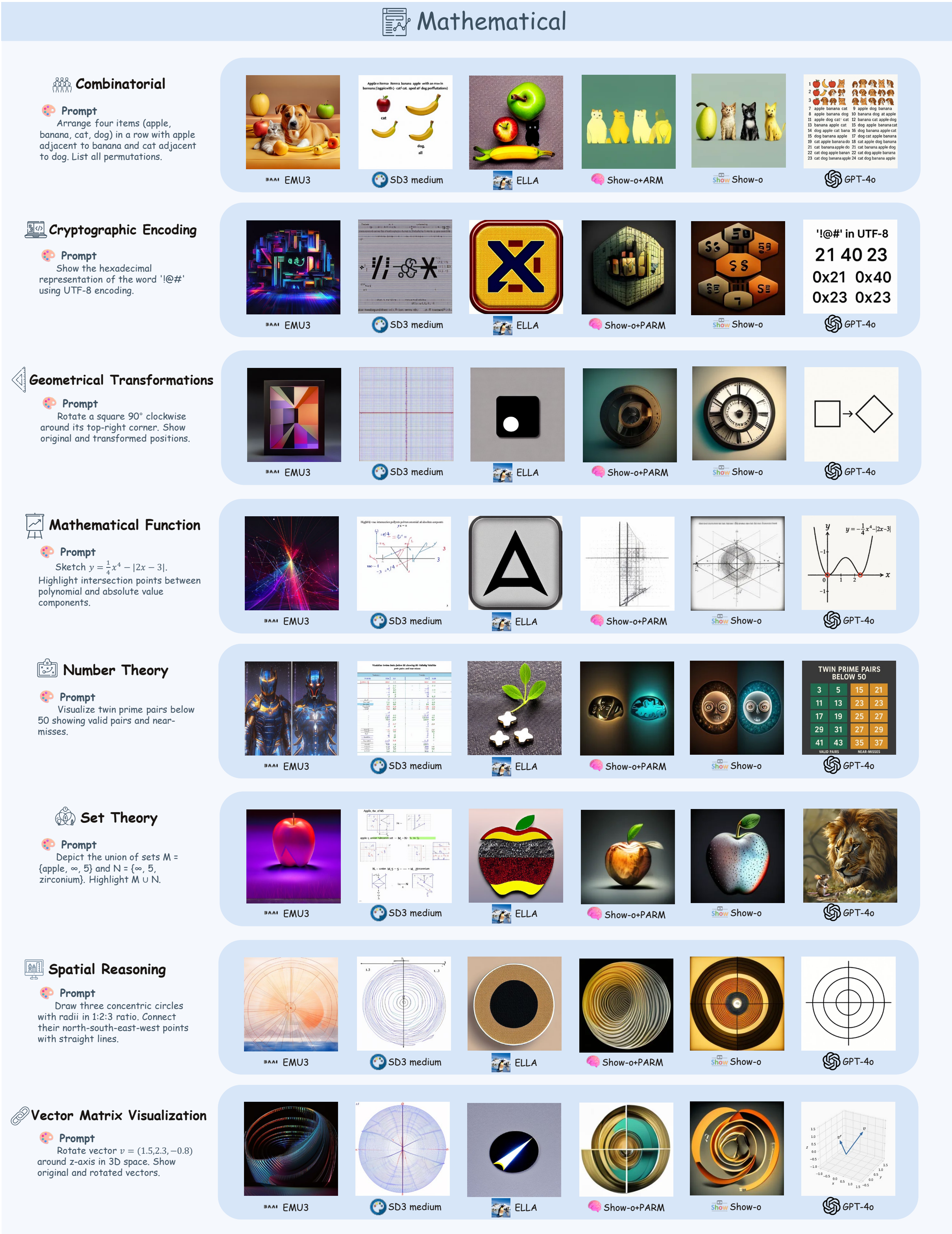

Figure 10: **Examples of Eight Subfields in Mathematical Reasoning**, spanning Combinatorial, Crypographic Encoding, Geometrical Transformations, Mathematical Function,spatial reasoning,et Theory, Spatial Reasoning and Vector Matrix Visualizations.

Commonsense Reasoning	Description
Affordance (14.53%)	Affordance commonsense reasoning involves providing a description of an object's potential *use* or *function*, requiring the model to generate an object based on that description.
Attribute (14.53%)	Attribute commonsense reasoning refers to the model's ability to infer or recognize the *properties* and *characteristics* of an object, utilizing both observable and unobservable information.
Color (14.82%)	Color commonsense reasoning pertains to the model's ability to infer the correct *color* of an object based on commonsense knowledge related to *color*.
Emotion Intention Commonsense (11.94%)	Emotion intention commonsense reasoning explores the model's ability to understand *emotional cues* and *intentions*, particularly in the context of *human-object interactions* in images. This subcategory evaluates how well the model can recognize and interpret *emotional states* and *intentions* from visual input.
Social & Cultural Knowledge (Object) (14.68%)	Social and cultural commonsense reasoning (Object) assesses the model's ability to leverage knowledge related to *social* and *cultural contexts* when generating a specific object.
Social & Cultural Knowledge (Scene) (15.11%)	Social and cultural commonsense reasoning (Scene) evaluates the model's ability to incorporate knowledge of *social* and *cultural contexts* when generating *scenes* or *environments* that accurately reflect specific *social* and *cultural* settings.
Temporal Understanding (14.39%)	Temporal understanding commonsense reasoning focuses on the model's ability to infer and apply knowledge related to *time-dependent changes* or *events*, including the ability to predict how *objects* or *scenes* may evolve over time based on contextual and temporal understanding.

Table 11: Definitions and proportions of the seven subcategories in commonsense reasoning within **R2I-Bench**. The percentage indicates the proportion of each subcategory within the overall commonsense reasoning category.

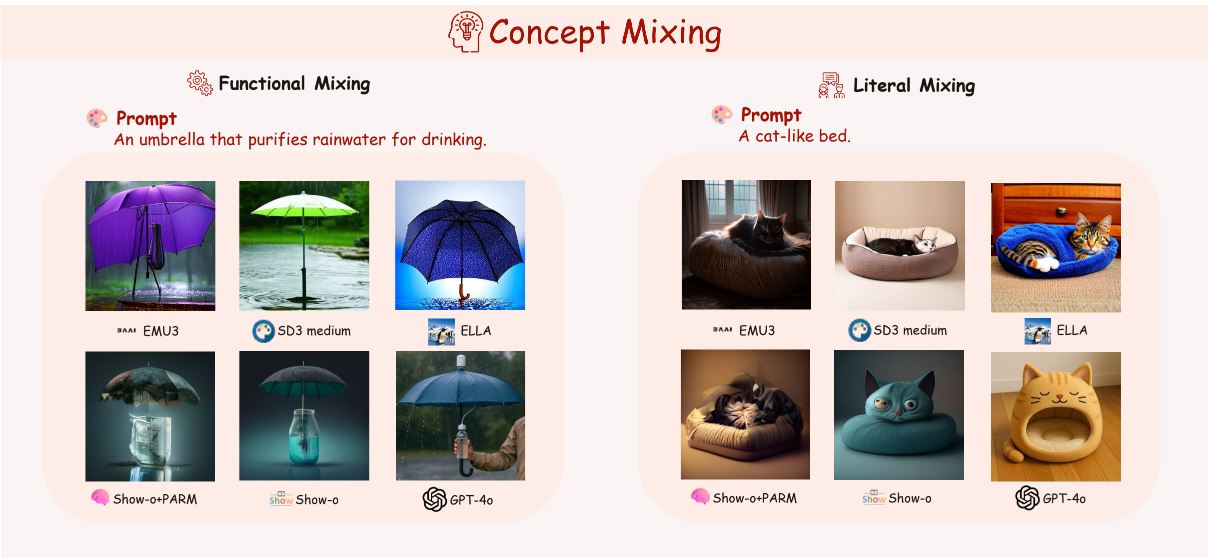

Figure 11: **Examples of Two Subfields in Concept Mixing**, including Functional Mixing and Literal Mixing.

Figure 12: **Examples of Three Subfields in Compositional Reasoning**, including Creative Compositional, Inferential Spatial, Color, Prescriptive Spatial.

Figure 13: **Examples of two Subfields in Causal Reasoning**, including Cause to Effect Reasoning and Cause to Effect Reasoning.

Figure 14: **Examples of Three Subfields in Numerical Reasoning**, including Approximate Number Generation, Conceptual Quantitative, Exact Number Generation.

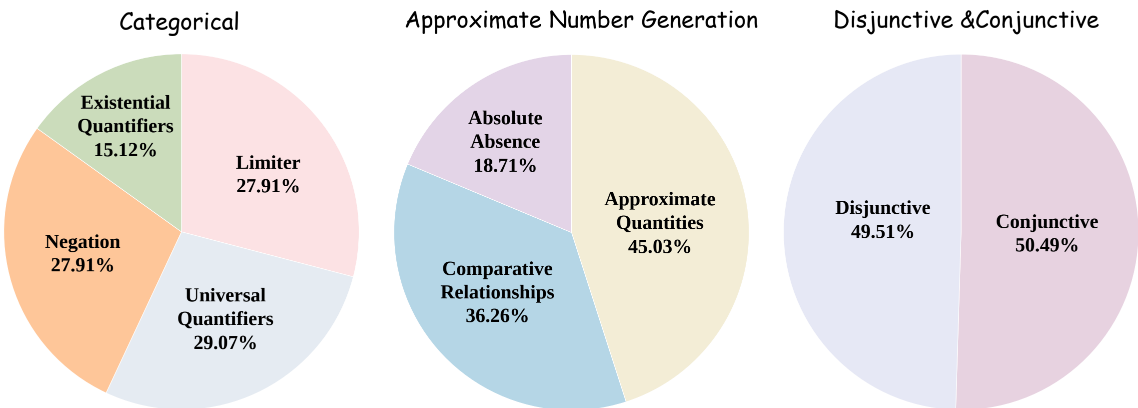

Figure 15: **Distribution of Quantifiers and Operations in Categorical, Approximate Number Generation, Disjunctive Reasoning, and Conjunctive Reasoning.**

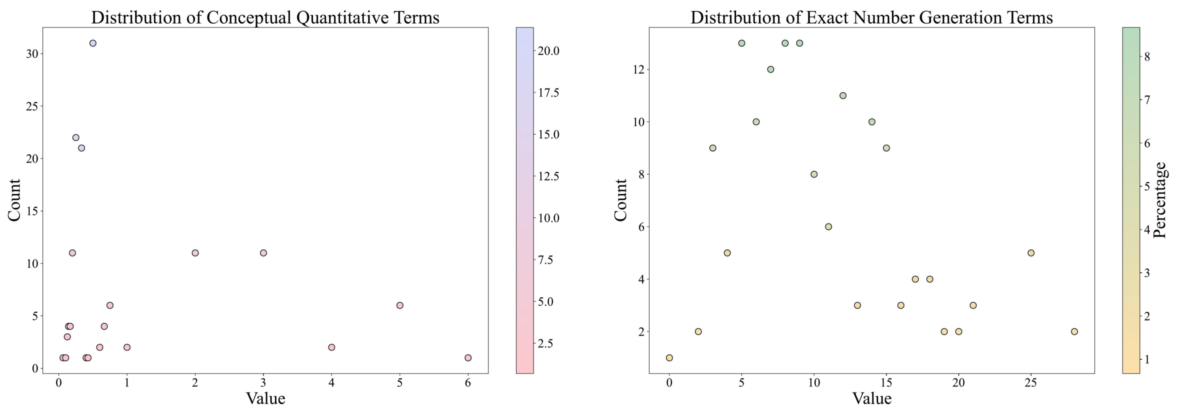

Figure 16: **Distribution of Numbers in Exact Number Generation and Conceptual Quantitative Reasoning.** Due to the current limitations of the best visual language models in numerical tasks, the numbers in Exact Number Generation are restricted to values within 30.

B Evaluation Details

All experiments with open-source models are performed on A-40 GPUs, whereas experiments involving closed-source models are conducted using the API key provided by the respective service. All experiments are conducted in a zero-shot setting to assess the generalization capabilities of text-to-image (T2I) generation models on reasoning tasks, without relying on few-shot prompting or additional fine-tuning.

B.1 Prompts details

Prompt for data instances creation

```
DATA_CREATION_PROMPT="""
You are an expert in text-to-image generation. I am
↪  currently creating a benchmark to test the
↪  reasoning ability of text-to-image generation
↪  models.

### Definition of [CATEGORY]:
[CATEGORY DEFINITIN]

### Definition of [SUBCATEGORY]:
[SUBCATEGORY DEFINITIN]
### Task:
Can you generate 10 test cases for [SUBCATEGORY] to
↪  evaluate the text-to-image model's reasoning
↪  ability?

### Notices:
1. The prompt should not explicitly mention the
↪  aspects that require reasoning, in order to avoid
↪  leaking the reasoning process.
2. Ensure that your ground truth is **uniquely
↪  determinable**.

### Good Cases
[SUBCATEGORY GOOD CASES]

### Incorrect Implementation Example
Here are some WRONG IMPLEMENTATION you will make:

[SUBCATEGORY BAD CASES]
### Output Format:
```json
{
  "Prompt": "xxx",
  "Reference Caption": "xxx"
}
"""
```

Figure 17: Prompt for data instances creation

B.1.1 Prompt for Data Instances Creation

We initially generate data instance using `gpt-4o` with the prompt in Figure 17 and deleted lots of duplicated and low-quality data instances in data filtering stage.

B.1.2 Prompt for Explanation Generation

We generate explanations using `GPT-4o` with the prompt in Figure 18.

B.1.3 Prompt for Image Evaluation

After refining the evaluation questions and scoring criteria, we use these prompts with the state-of-the-art visual language model, GPT-4o, with a fixed temperature of 0.1, to obtain scores for each image relative to its corresponding text.

Prompt for Explanation Generation

```
EXPLANATION_GENERATION_PROMPT = """
Following is a prompt focused on testing the reasoning
↪  capabilities of a text-to-image generation model.
Prompt: [PROMPT]
Reference Caption: [REFERENCE CAPTION]
Your task is to explain why the correct answer
↪  corresponding to the Prompt is the Reference
↪  Caption.
### Output Format:
```json
{
  "Explanation": "xxx"
}
"""
```

Figure 18: Prompt for explanation generation

Prompt for Image Evaluation

```
IMAGE_EVALUATION_PROMPT = """
# Text-to-Image Quality Evaluation Protocol

## System Instructions
You are an AI quality auditor for text-to-image
↪  generation. Your task is to evaluate images with
↪  ABSOLUTE RUTHLESSNESS. Only images that meet the
↪  HIGHEST standards should receive the top scores.

## Task Overview
The image is generated based on the following prompt:
[PROMPT]

## Evaluation Criteria
[QUESTION LIST]

## Output Format
You may provide an analysis in your output, but ensure
↪  that the final line is formatted as shown below:

## Important Enforcement
[IMPORTANT ENFORCEMENT]

```json
{
    "id": score,
    ...
}
```
"""
```

Figure 19: Prompt for image evaluation

Prompt for Evaluation Criteria Generation

```
IMAGINE_IMAGE = """
This test case is designed to evaluate the image
↪  generation model. What do you think the correct
↪  image should look like based on this prompt?
Prompt: [PROMPT]
Expected: [EXPECTED]
"""

DESIGN_EVALUATION_QUESTIONS = """
Now, create a set of evaluation questions to determine
↪  whether the image is accurate.
For each question, define the criteria for different
↪  levels of performance, with the rating scale
↪  ranging from [0, 1].

Prompt: [PROMPT]
Expected: [EXPECTED]
"""

WEIGHT = """
# Image Generation Model Assessment

This prompt is designed to evaluate the performance of
↪  an image generation model.

Prompt: [PROMPT]
Reference Answer: [EXPECTED]
Assessment Points: [ASSESSMENT POINTS]

---

### Weight Assignment Instructions

- Based on key evaluation criteria, assign weights
↪  (1-10) to each evaluation question.
- Higher weights should be assigned to critical
↪  factors related to **core reasoning points**.
- Mid-range weights should be assigned to aspects that
↪  are not related to reasoning but are still
↪  relevant to the image.
- Lower weights should be assigned to aspects like
↪  **image quality**, **realism**, and **clarity**.

---

### Evaluation Questions

Please provide the **complete list of evaluation
↪  questions** without any omissions.

```json
[
    {
        "id": "number",
        "weight": "weight",
        "question": "...",
        "evaluation_criteria": "..."
    }
]
"""
```

Figure 20: Prompt for evaluation criteria generation

B.1.4 Prompt for Evaluation Criteria Generation

We generate evaluation questions and scoring criteria initially using the GPT-4o API with a fixed temperature of 0.1 to ensure score stability. Due to performance degradation in GPT-4o when handling long contexts, we have separated the prompts for generating questions and assigning weights. This approach ensures that GPT-4o can fully adhere to all the key points specified in each prompt. These outputs are then carefully reviewed and refined by human annotators to ensure they align with human judgment. The prompt we used is in Figure 20

B.2 Human Annotators

To incorporate human judgment and validate the effectiveness of our evaluation approach, we organize a group of senior college students. Each participant is tasked with comparing the image outputs generated by two similarly performing models, `Lumina-Image 2.0` and `Sana-1.5`, selecting the image they find most aligned with the prompt or indicating if both outputs are equally satisfactory or unsatisfactory.

B.3 Model Details

Model Sources. For different T2I models, we select their latest models and best-performing configurations for evaluation to fully **`R2I-Bench`** their reasoning ability. Table 12 presents the release time and model sources of MLLMs used in **`R2I-Bench`**.

C Detailed Experimental Results

C.1 Main Results across 33 Subcategories

Table 13 to 17 are the main results of the models across subcategories in *Mathematical Reasoning, Logical Reasoning, Commonsense Reasoning, Concept Mixing Reasoning, Causal Reasoning, Numerical Reasoning and Compositional Reasoning*.

C.2 Comparison of Subcategory Performance: Standard T2I Model vs. Pipeline-based Framework

Figure 21 presents detailed performance comparison: standard T2I model vs. pipeline-based framework

C.3 Results of Our Evaluation Methods and Additional Metrics in the Benchmark

In this section, we present the results of the evaluation methods employed, along with other metrics. The detailed evaluation results are provided in Table 18.
```

Table 12: **The Release Time and Model Source of T2I Models Evaluated in `R2I-Bench`.**

Model	Release Time	Source	URL
EMU3 (Wang et al., 2024)	2024-09	local checkpoint	https://github.com/baaivision/Emu3
Janus-Pro-7B (Chen et al., 2025)	2025-01	local checkpoint	https://github.com/deepseek-ai/Janus/
LlamaGen (Sun et al., 2024a)	2024-06	local checkpoint	https://huggingface.co/FoundationVision/LlamaGen
SD3-medium (Esser et al., 2024b)	2024-10	local checkpoint	https://huggingface.co/stabilityai/stable-diffusion-3.5-medium
Lumina-Image-2.0 (Qin et al., 2025b)	2025-03	local checkpoint	https://github.com/Lumina-Image2.0
Sana-1.5 (Xie et al., 2025b)	2025-03	local checkpoint	https://github.com/NVlabs/Sana
Lumina-T2I (Qin et al., 2025b)	2024-05	local checkpoint	https://huggingface.co/Alpha-VLLM/Lumina-Next-SFT-diffusers
LLM4GEN$_{SD1.5}$ (Liu et al., 2025)	2024-07	local checkpoint	https://github.com/YUHANG-Ma/LLM4GEN
ELLA$_{SD1.5}$ (Hu et al., 2024)	2024-03	local checkpoint	https://github.com/TencentQQGYLab/ELLA
Show-o+PARM (Guo et al., 2025)	2025-01	local checkpoint	https://huggingface.co/ZiyuG/Image-Generation-CoT
Show-o+DPO (Guo et al., 2025)	2025-01	local checkpoint	https://huggingface.co/ZiyuG/Image-Generation-CoT
Show-o+ORM (Guo et al., 2025)	2025-01	local checkpoint	https://huggingface.co/ZiyuG/Image-Generation-CoT
gpt-image-1 (Hurst et al., 2024)	2025-04	API	https://platform.openai.com/

Method	Overall	Comb.	Crypt. Enc.	Geo. Trans.	Math Func.	Num. Th.	Spatial Reas.	Vec/Mat. Vis.	Set Th.
Diffusion Models									
SD3-medium (Esser et al., 2024a)	0.19	0.07	0.10	0.37	0.01	0.23	0.24	0.13	0.13
Lumina-Image 2.0 (Qin et al., 2025b)	0.13	0.09	0.09	0.18	0.03	0.06	0.28	0.01	0.16
Sana-1.5 (Xie et al., 2025b)	0.13	0.10	0.06	0.32	0.02	0.08	0.16	0.06	0.16
Lumina-T2I (Qin et al., 2025b)	0.13	0.04	0.01	0.17	0.03	0.07	0.16	0.05	0.07
OminGen (Xiao et al., 2024)	0.18	0.19	0.05	0.27	0.06	0.33	0.26	0.08	0.21
LLM4GEN$_{SD1.5}$ (Liu et al., 2025)	0.07	0.03	0.01	0.16	0.01	0.01	0.01	0.01	0.09
ELLA$_{SD1.5}$ (Hu et al., 2024)	0.07	0.01	0.03	0.14	0.01	0.01	0.11	0.07	0.07
AutoRegressive Models									
EMU3 (Wang et al., 2024)	0.09	0.05	0.01	0.18	0.03	0.08	0.14	0.05	0.08
Janus-Pro-7B (Chen et al., 2025)	0.07	0.02	0.01	0.16	0.02	0.06	0.12	0.01	0.06
LlamaGen (Sun et al., 2024a)	0.07	0.04	0.01	0.24	0.01	0.01	0.01	0.05	0.10
Show-o (Xie et al., 2024)	0.12	0.12	0.02	0.20	0.01	0.26	0.19	0.03	0.14
Reasoning-Enhanced Models									
Show-o+ORM (Guo et al., 2025)	0.12	0.12	0.02	0.19	0.02	0.24	0.18	0.04	0.13
Show-o+DPO (Guo et al., 2025)	0.13	0.14	0.04	0.20	0.03	0.23	0.20	0.06	0.14
Show-o+PARM (Guo et al., 2025)	0.13	0.13	0.03	0.21	0.02	0.27	0.20	0.04	0.15
Close Source Models									
DALL-E3 (Ma et al., 2024)	0.21	0.07	0.14	0.32	0.05	0.18	0.42	0.12	0.36
GPT-4o (Hurst et al., 2024)	0.58	0.43	0.43	0.73	0.59	0.49	0.74	0.46	0.75

Table 13: Evaluation of mathematical capabilities in generative models. Comb.: Combinatorial, Crypt. Enc.: Cryptographic Encoding, Geo. Trans.: Geometrical Transformations, Math Func.: Mathematical Function, Num. Th.: Number Theory, Spatial Reas.: Spatial Reasoning, Vec/Mat. Vis.: Vector & Matrix Visualization, Set Th.: Set Theory.

Method	Overall	Abduc.	Cat.	Conj.	Ded.	Disj.	Hypo.	Suff.
Diffusion Models								
SD3-medium (Esser et al., 2024a)	0.55	0.44	0.61	0.85	0.45	0.43	0.48	0.56
Lumina-Image 2.0 (Qin et al., 2025b)	0.56	0.44	0.57	0.87	0.38	0.51	0.52	0.54
Sana-1.5 (Xie et al., 2025b)	0.49	0.46	0.56	0.89	0.50	0.48	0.59	0.56
Lumina-T2I (Qin et al., 2025b)	0.38	0.33	0.50	0.69	0.54	0.43	0.55	0.53
OminGen (Xiao et al., 2024)	0.51	0.42	0.64	0.69	0.39	0.52	0.41	0.47
LLM4GEN$_{SD1.5}$ (Liu et al., 2025)	0.55	0.33	0.48	0.70	0.40	0.53	0.55	0.49
ELLA$_{SD1.5}$ (Hu et al., 2024)	0.40	0.29	0.41	0.64	0.26	0.59	0.40	0.39
AutoRegressive Models								
EMU3 (Wang et al., 2024)	0.55	0.38	0.53	0.71	0.44	0.64	0.52	0.58
Janus-Pro-7B (Chen et al., 2025)	0.46	0.25	0.64	0.85	0.13	0.57	0.52	0.22
LlamaGen (Sun et al., 2024a)	0.38	0.15	0.48	0.55	0.17	0.59	0.29	0.35
Show-o (Xie et al., 2024)	0.42	0.40	0.62	0.71	0.35	0.42	0.32	0.38
Reasoning-Enhanced Models								
Show-o+ORM (Guo et al., 2025)	0.37	0.33	0.48	0.47	0.35	0.41	0.37	0.17
Show-o+DPO (Guo et al., 2025)	0.41	0.29	0.44	0.44	0.36	0.43	0.34	0.18
Show-o+PARM (Guo et al., 2025)	0.45	0.38	0.53	0.76	0.33	0.45	0.31	0.36
Close Source Models								
DALLE (Ma et al., 2024)	0.69	0.56	0.67	0.87	0.70	0.46	0.79	0.78
gpt-image-1 (Ma et al., 2024)	0.81	0.79	0.88	0.95	0.79	0.76	0.79	0.73

Table 14: Evaluation of text-to-image generation on Logical Reasoning in **R2I-Bench**. Abduc.: Abductive, Cat.: Categorical, Conj.: Conjunctive, Ded.: Deductive, Disj.: Disjunctive, Hypo.: Hypothetical, Suff.: Sufficient Conditional

Method	Overall	Afford.	Attribute	Color	Emotion	Object	Scene	Temp.
Diffusion Models								
SD3-medium (Esser et al., 2024a)	0.54	0.56	0.53	0.55	0.63	0.44	0.55	0.52
Lumina-Image 2.0 (Qin et al., 2025b)	0.49	0.46	0.53	0.51	0.65	0.34	0.53	0.46
Sana-1.5 (Xie et al., 2025b)	0.49	0.42	0.60	0.51	0.64	0.33	0.53	0.51
Lumina-T2I (Qin et al., 2025b)	0.38	0.36	0.47	0.40	0.57	0.33	0.46	0.39
OminGen (Xiao et al., 2024)	0.43	0.41	0.51	0.39	0.54	0.30	0.47	0.41
LLM4GEN$_{SD1.5}$ (Liu et al., 2025)	0.55	0.37	0.47	0.44	0.66	0.36	0.56	0.51
ELLA$_{SD1.5}$ (Hu et al., 2024)	0.40	0.33	0.40	0.34	0.37	0.28	0.36	0.32
AutoRegressive Models								
EMU3 (Wang et al., 2024)	0.46	0.40	0.50	0.43	0.58	0.39	0.52	0.42
Janus-Pro-7B (Chen et al., 2025)	0.45	0.38	0.57	0.45	0.58	0.32	0.49	0.40
LlamaGen (Sun et al., 2024a)	0.38	0.38	0.42	0.39	0.40	0.29	0.38	0.38
Show-o (Xie et al., 2024)	0.42	0.44	0.48	0.41	0.44	0.32	0.44	0.36
Reasoning-Enhanced Models								
Show-o+ORM (Guo et al., 2025)	0.42	0.42	0.49	0.40	0.47	0.35	0.47	0.38
Show-o+DPO (Guo et al., 2025)	0.43	0.43	0.52	0.44	0.45	0.36	0.47	0.36
Showo-o+PARM (Guo et al., 2025)	0.45	0.45	0.48	0.46	0.55	0.40	0.49	0.47
Close Source Models								
DALLE3 (Ma et al., 2024)	0.78	0.70	0.80	0.86	0.81	0.81	0.77	0.72
gpt-iamge-1 (Ma et al., 2024)	0.83	0.89	0.79	0.80	0.89	0.85	0.87	0.75

Table 15: Evaluation Results of text-to-image generation on Commonsense Reasoning in `R2I-Bench`. Afford.: Affordance. Temp.: Temporal Understanding. Emotion: Emotion Intention Commonsense Reasoning. Object: Social Cultural Knowledge (Object). Scene: Social Cultural Knowledge (Scene).

Method	Overall	Numerical			Overall	Causal Reasoning	
		Approx.	**Conceptual.**	**Exact.**		**C2E**	**E2C**
Diffusion Models							
SD3-medium (Esser et al., 2024a)	0.50	0.53	0.49	0.48	0.18	0.20	0.16
Lumina-Image 2.0 (Qin et al., 2025b)	0.43	0.54	0.40	0.35	0.40	0.37	0.44
Sana-1.5 (Xie et al., 2025b)	0.47	0.58	0.37	0.47	0.21	0.23	0.19
Lumina-T2I (Qin et al., 2025b)	0.45	0.53	0.45	0.38	0.18	0.18	0.18
OminGen (Xiao et al., 2024)	0.47	0.59	0.40	0.42	0.34	0.26	0.41
LLM4GEN$_{SD1.5}$ (Liu et al., 2025)	0.39	0.44	0.36	0.36	0.45	0.46	0.44
ELLA$_{SD1.5}$ (Hu et al., 2024)	0.32	0.41	0.25	0.30	0.29	0.22	0.38
AutoRegressive Models							
EMU3 (Wang et al., 2024)	0.61	0.73	0.54	0.56	0.41	0.36	0.47
Janus-Pro-7B (Chen et al., 2025)	0.46	0.53	0.38	0.48	0.36	0.34	0.39
LlamaGen (Sun et al., 2024a)	0.35	0.43	0.31	0.30	0.12	0.12	0.12
Show-o (Xie et al., 2024)	0.57	0.68	0.50	0.53	0.30	0.23	0.38
Reasoning-Enhanced Models							
Show-o+ORM (Guo et al., 2025)	0.49	0.52	0.46	0.49	0.26	0.30	0.23
Show-o+DPO (Guo et al., 2025)	0.51	0.58	0.46	0.50	0.31	0.35	0.28
Show-o+PARM (Guo et al., 2025)	0.56	0.65	0.49	0.53	0.32	0.36	0.27
Close Source Models							
DALLE (Ma et al., 2024)	0.69	0.71	0.64	0.72	0.64	0.69	0.59
GPT-4o (Hurst et al., 2024)	0.88	0.90	0.81	0.92	0.71	0.85	0.56

Table 16: Evaluation of text-to-image generation on Numerical Reasoning and Causal Reasoning in **R2I-Bench**. Approx.: Approximate Number Generation. Conceptual: Conceptual Quantitative Reasoning. Exact: Exact Number Generation. C2E: Cause to Effect Reasoning. E2C: Effect to Cause Reasoning.

Method	Overall	Concept Mixing		Overall	Compositional		
		Functional	Literal		Creative	Inferential	Prescriptive
Diffusion Models							
SD3-medium (Esser et al., 2024a)	0.63	0.49	0.75	0.64	0.46	0.73	0.72
Lumina-Image 2.0 (Qin et al., 2025b)	0.54	0.52	0.56	0.65	0.50	0.72	0.73
Sana-1.5 (Xie et al., 2025b)	0.66	0.55	0.75	0.67	0.59	0.79	0.63
Lumina-T2I (Qin et al., 2025b)	0.55	0.47	0.62	0.49	0.42	0.56	0.49
Omnigen (Xiao et al., 2024)	0.43	0.27	0.58	0.60	0.46	0.80	0.54
LLM4GEN$_{SD1.5}$ (Liu et al., 2025)	0.60	0.48	0.70	0.48	0.44	0.61	0.39
ELLA$_{SD1.5}$ (Hu et al., 2024)	0.40	0.33	0.46	0.44	0.34	0.55	0.43
AutoRegressive Models							
EMU3 (Wang et al., 2024)	0.62	0.51	0.70	0.59	0.50	0.68	0.59
Janus-Pro-7B (Chen et al., 2025)	0.64	0.55	0.71	0.60	0.56	0.73	0.52
LlamaGen (Sun et al., 2024a)	0.49	0.45	0.53	0.39	0.42	0.50	0.27
Show-o (Xie et al., 2024)	0.56	0.42	0.68	0.55	0.41	0.65	0.60
Reasoning-Enhanced Models							
ORM (Guo et al., 2025)	0.44	0.30	0.56	0.45	0.35	0.54	0.45
DPO (Guo et al., 2025)	0.48	0.35	0.61	0.47	0.38	0.56	0.47
PARM (Guo et al., 2025)	0.51	0.37	0.63	0.49	0.39	0.58	0.51
Close Source Models							
DALLE3 (Ma et al., 2024)	0.86	0.82	0.90	0.76	0.73	0.82	0.72
GPT-4o (Hurst et al., 2024)	0.89	0.88	0.90	0.87	0.81	0.84	0.95

Table 17: Evaluation of text-to-image generation on Concept Mixing and Compositional Reasoning in `R2I-Bench`. Functional: Functional Mixing Reasoning. Literal: Literal Mixing Reasoning. Creative: Creative Compositional Reasoning. Inferential: Inferential Spatial Reasoning. Prescriptive: Prescriptive Spatial Reasoning

Table 18: Comparison of our evaluation methods and other image-text alignment metrics across different models and categories.

Category	Models	Pairwise Accuracy ↑	Kendall's τ ↑	Spearman's Rank Correlation ↑
Commonsense	CLIPScore (Hessel et al., 2021)	0.61	0.22	0.42
	DSGScore (Cho et al., 2023)	0.54	0.10	0.30
	VIEScore (Ku et al., 2024)	0.70	0.45	0.34
	VQAscore (Lin et al., 2024)	0.60	0.22	0.39
	Ours	0.64	0.60	0.62
Compositional	CLIPScore (Hessel et al., 2021)	0.71	0.42	0.39
	DSGScore (Cho et al., 2023)	0.50	0.38	0.26
	VIEScore (Ku et al., 2024)	0.58	0.40	0.32
	VQAscore (Lin et al., 2024)	0.64	0.48	0.45
	Ours	0.73	0.76	0.61
Logical	CLIPScore (Hessel et al., 2021)	0.61	0.22	0.30
	DSGScore (Cho et al., 2023)	0.63	0.15	0.25
	VIEScore (Ku et al., 2024)	0.78	0.63	0.40
	VQAscore (Lin et al., 2024)	0.76	0.72	0.68
	Ours	0.76	0.72	0.63
Causal	CLIPScore (Hessel et al., 2021)	0.54	0.18	0.21
	DSGScore (Cho et al., 2023)	0.51	0.22	0.28
	VIEScore (Ku et al., 2024)	0.69	0.64	0.68
	VQAscore (Lin et al., 2024)	0.62	0.33	0.70
	Ours	0.69	0.64	0.64
Concept Mixing	CLIPScore (Hessel et al., 2021)	0.62	0.24	0.25
	DSGScore (Cho et al., 2023)	0.42	0.25	0.18
	VIEScore (Ku et al., 2024)	0.52	0.16	0.28
	VQAscore (Lin et al., 2024)	0.67	0.52	0.48
	Ours	0.83	0.91	0.87
Numerical	CLIPScore (Hessel et al., 2021)	0.61	0.22	0.16
	DSGScore (Cho et al., 2023)	0.47	0.21	0.29
	VIEScore (Ku et al., 2024)	0.87	0.74	0.68
	VQAscore (Lin et al., 2024)	0.78	0.64	0.57
	Ours	0.65	0.67	0.62
Mathematical	CLIPScore (Hessel et al., 2021)	0.72	0.44	0.54
	DSGScore (Cho et al., 2023)	0.60	0.45	0.43
	VIEScore (Ku et al., 2024)	0.72	0.44	0.46
	VQAscore (Lin et al., 2024)	0.63	0.33	0.67
	Ours	0.69	0.93	0.87
Average	CLIPScore (Hessel et al., 2021)	0.631	0.263	0.310
	DSGScore (Cho et al., 2023)	0.520	0.220	0.254
	VIEScore (Ku et al., 2024)	0.694	0.494	0.451
	VQAscore (Lin et al., 2024)	0.629	0.463	0.563
	Ours	0.713	0.747	0.694

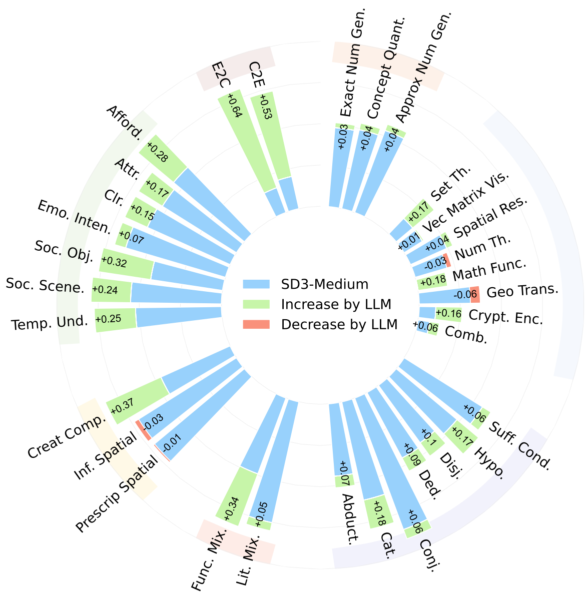

Figure 21: **Detailed Performance Comparison: Standard T2I Model vs. Pipeline-based Framework**. We denote the results of standard T2I models in blue pillars and highlight the increase and decrease magnitude with the pipeline-based framework by green and red colors, respectively.