T2I-ReasonBench: Benchmarking Reasoning-Informed Text-to-Image Generation

Kaiyue Sun[1] **Rongyao Fang**[2] **Chengqi Duan**[1] **Xian Liu**[2] **Aoxue Li**[3] **Xihui Liu**[1†]

[1]HKU [2]CUHK [3]Huawei Noah's Ark Lab

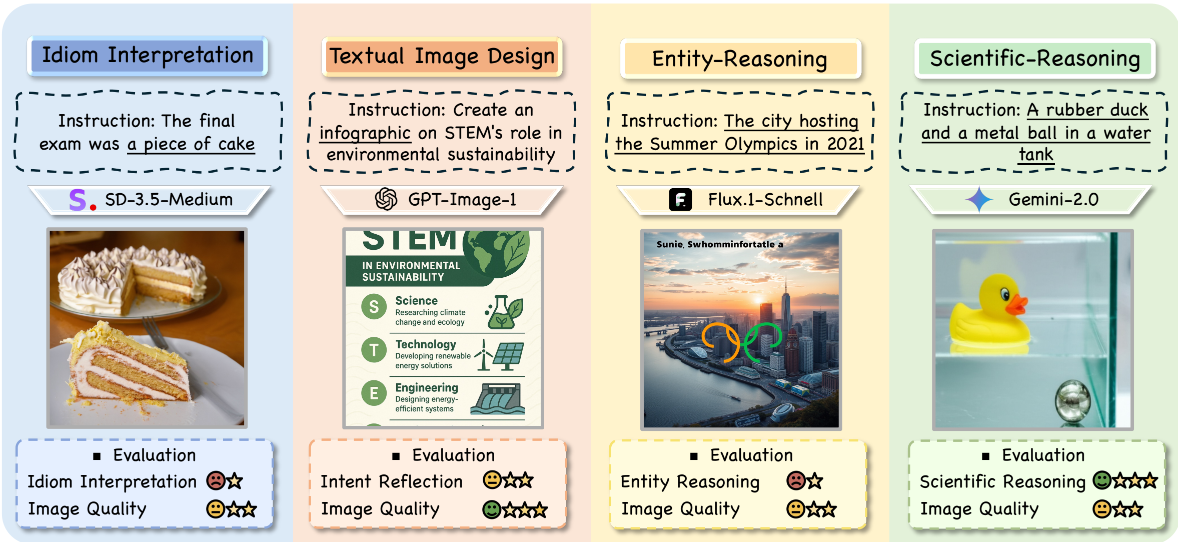


Figure 1: **Overview of T2I-ReasonBench.** We propose T2I-ReasonBench, a benchmark evaluating reasoning-informed generation of text-to-image (T2I) models. It consists of four dimensions: **Idiom Interpretation, Textual Image Design, Entity-Reasoning and Scientific-Reasoning**. We propose a two-stage evaluation protocol to measure T2I-ReasonScore, a metric that integrates reasoning accuracy, detail faithfulness and image quality. We benchmark various types T2I models, and provide comprehensive analysis on their reasoning and generation abilities.

Abstract

Text-to-image (T2I) generative models have achieved remarkable progress, demonstrating exceptional capability in synthesizing high-quality images from textual prompts. While existing research and benchmarks have extensively evaluated the ability of T2I models to follow the literal meaning of prompts, their ability to reason over prompts with domain knowledge to uncover implicit meaning and contextual nuances remains underexplored. To bridge this gap, we introduce T2I-ReasonBench, a novel benchmark designed to explore the knowledge-driven reasoning capabilities of T2I models. T2I-ReasonBench comprises 800 meticulously designed prompts organized into four dimensions: **(1) Idiom Interpretation**, **(2) Textual Image Design**, **(3) Entity Reasoning**, and **(4) Scientific Reasoning**. These dimensions challenge models to integrate domain knowledge, infer implicit meaning, and resolve contextual ambiguities. To quantify the performance, we introduce a two-stage evaluation framework: a large language model (LLM) generates prompt-specific question-criterion pairs that evaluate if the image includes the essential elements resulting from correct reasoning; a multimodal LLM (MLLM) then scores the generated image against these criteria. Our comprehensive study across 16 state-of-the-art diffusion and unified multimodal models (UMMs) reveal two primary bottlenecks. First, many models lack the foundational reasoning ability to fully comprehend complex prompts. Second, even models with stronger reasoning modules exhibit a persistent gap between their internal understanding and the final generated image. This highlights an urgent need for the next generation of T2I systems to not only improve their reasoning capability but also to enhance integration between reasoning and synthesis. Project page: https://github.com/KaiyueSun98/T2I-ReasonBench.

†Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2026, pages 8919–8944
July 2-7, 2026

1 Introduction

Recent advancements in T2I generative models have enabled the creation of visually appealing images from textual prompts. However, these models often struggle with generating complex scenes that demand reasoning. Current benchmarks (Yu et al., 2022; Hu et al., 2023a; Huang et al., 2023; Ghosh et al., 2023; Hu et al., 2024; Li et al., 2024a; Wu et al., 2024; Huang et al., 2025; Wei et al., 2025) primarily evaluate literal prompt-image alignment, focusing on object attributes (e.g., color, attribute, count) and relationships. While DPG-Bench (Hu et al., 2024) extends evaluation to long-text comprehension, it remains confined to multi-object composition tasks. These frameworks fail to test models' ability to reason beyond explicit instructions. For instance, generating an image of "A beach ball and a marble in a swimming pool" requires not only object composition but also reasoning about physical laws (e.g., inferring the ball floats while the marble sinks). Such reasoning necessitates understanding related scientific knowledge, such as material density and buoyancy, as well as integrating the reasoning process into T2I generation.

To address this gap, we propose **T2I-ReasonBench**, a novel benchmark designed to systematically evaluate the reasoning ability of T2I models in four dimensions: (1) Idiom Interpretation: deciphering the implicit meanings of idiomatic expressions with the context to generate appropriate images. (2) Textual Image Design: understanding the intention of design and effectively planning the integrated visual-textual contents. (3) Entity Reasoning: applying and integrating the knowledge about world entities in image generation, and (4) Scientific Reasoning: reasoning with scientific knowledge (e.g., physics, chemistry) to produce images adhering to the underlying scientific laws. T2I-ResonBench encompasses the above four dimensions with 800 meticulously designed prompts, all of which require deep reasoning.

To rigorously evaluate the performances of T2I models, we introduce a two-stage evaluation framework and propose **T2I-ReasonScore**, a quantitative metric for assessing the quality of reasoning-informed T2I generation. First, an LLM generates specific question-criterion pairs for each prompt. To evaluate the images, an MLLM then answers each question and assigns a score based on the paired criterion. By averaging these scores, we measure how faithfully the image reflects the implicit meaning of the prompt, capturing the effectiveness of model's reasoning. Our approach allows for fine-grained and interpretable evaluation of models' reasoning ability and addresses the limitation of previous benchmarks that focused solely on literal prompt following.

We evaluate 16 state-of-the-art T2I models, including 8 diffusion models, 5 UMMs, and 3 proprietary models. Based on the results, we present the first comprehensive study to decouple reasoning with generation and evaluate the gap in between. Our analysis reveals that reasoning is the primary bottleneck for current models. Furthermore, even though UMMs unify understanding and generation in a single framework, they still struggle to transfer their knowledge and reasoning effectively into the final image.

Our contributions are threefold: (1) We propose T2I-ReasonBench, a novel benchmark with meticulously designed tasks to explore the reasoning-informed T2I generation. (2) Our prompt-specific evaluation framework enables fine-grained and interpretable evaluation. (3) We evaluate a range of of T2I systems, conduct a thorough analysis on their reasoning and generation abilities, and provide insights into limitations of current models and future model design.

2 Related Work

2.1 Text-to-image Generation.

Diffusion models. T2I generation has seen rapid advances in recent years, primarily driven by the emergence and refinement of diffusion models (Dhariwal and Nichol, 2021; Ho et al., 2020; Nichol et al., 2021; Saharia et al., 2022). By formulating image synthesis as a progressive denoising process, these models pushed the boundaries of quality and controllability of T2I generation, and established the backbone for modern T2I systems like the Stable Diffusion series (Esser et al., 2024a; Rombach et al., 2022), and the Flux series (Labs, 2024). Recent models like HiDream (hidream, 2024) and Qwen-Image (Wu et al., 2025) further extend this paradigm, achieving fine-grained, photorealistic T2I generation, solidifying diffusion as the backbone of modern T2I systems.

Unified multimodal models. To achieve better token-level alignment between text and image modalities, recent research has shifted toward LLM-based architectures. This includes both autoregressive models, which synthesize images by directly predicting sequences of visual tokens(Ramesh et al., 2021; Ding et al., 2021; Sun et al., 2024; Liu et al., 2024), and unified multimodal models (Team, 2024; Xie et al., 2024; Chen et al., 2025d; Deng et al., 2025; Chen et al., 2025b; Fang et al., 2025; Duan et al., 2025). These unified systems typically combine an autoregressive language model with a diffusion module to integrate understanding and generation. For instance, GoT (Fang et al., 2025) uses an MLLM for semantic-spatial reasoning before diffusion-based synthesis, and Bagel (Deng et al., 2025) unifies an LLM and a diffusion model within a single transformer to generate reasoning chains prior to image creation.

2.2 Text-to-image Benchmarks and Evaluation Metrics.

Benchmarks. Current T2I benchmarks (Yu et al., 2022; Hu et al., 2023a; Huang et al., 2023; Ghosh et al., 2023; Hu et al., 2024; Li et al., 2024a; Wu et al., 2024; Huang et al., 2025; Wei et al., 2025) primarily evaluate literal prompt-image alignment. For example, GenEval (Ghosh et al., 2023) utilizes object detection techniques to test whether generated images correctly capture object co-occurrence, position, count, and color described in the prompts. Recent benchmarks have shifted focus from literal alignment to reasoning capabilities of T2I models. For instance, Commonsense-T2I (Fu et al., 2024) tests everyday logic through adversarial prompt pairs; PhyBench (Meng et al., 2024) evaluates physical common sense; WISE (Niu et al., 2025) assesses broader world knowledge; and R2I-Bench (Chen et al., 2025c) includes both composition and reasoning categories. Our benchmark encompasses a comprehensive set of reasoning dimensions and provides a deeper, more insightful analysis of models' reasoning and generation abilities.

Evaluation Metrics. Conventional text-image alignment metrics like CLIPscore (Hessel et al., 2021) and VQAscore (Lin et al., 2024) work as bag-of-words models, lacking the expertise needed to evaluate specific composition and reasoning generations. To address this, many works adopt a more targeted, disentangled question-answering framework. Leveraging powerful LLMs and MLLMs, this method first generates specific diagnostic questions and then uses VQA models to answer them by inspecting the image. This approach has been successfully applied across studies evaluating text-image alignment (Hu et al., 2023b; Cho et al., 2023a; Yarom et al., 2023; Cho et al., 2023b), composition (Wu et al., 2024), reasoning (Chen et al., 2025c), and factual correctness (Lim et al., 2025). Given the complexity of our benchmark, this targeted, disentangled approach is more reliable than metrics with identical instructions.

3 Benchmark Construction

While modern T2I models are good at explicit prompt-to-image translation, their capacity for reasoning-informed generation remains underexplored. Existing benchmarks focus predominantly on literal text-image alignment (e.g., object existence, spatial arrangements) but fail to evaluate whether models possess reasoning abilities to uncover the deeper meaning behind the text and generate logically coherent visual content. To this end, we identify four dimensions that challenge T2I models to reason about the instructions with domain knowledge before visualizing them:

Idiom Interpretation. An idiom is a phrase or combination of words with a figurative meaning that differs from its literal meaning. Idioms are common in everyday language, and their meanings usually cannot be deduced by analyzing individual words. For T2I models, prompts containing idioms demand reasoning to obtain the latent meaning before generating correct visual content. This process requires leveraging linguistic knowledge and effectively analyzing context.

By sourcing from a book (idi, 2023) and the internet, we collect 200 idioms that are commonly used in daily life but challenging for T2I models. We then use an LLM to generate sentences containing the idioms but without explicitly revealing their meanings. These idioms span diverse topics such as social interactions, lifestyle, and emotions. For example, the sentence "He told a funny joke to break the ice at the start of the meeting" uses the idiom "break the ice", which means to ease tension, rather than literally destroying the ice.

Textual Image Design. Rich-text images combine visuals and text harmoniously. These images are used to serve specific communicative goals, such as education, marketing, and promotion. Generating such content requires T2I models reasoning about the purpose behind the image and applying goal-oriented design skills like text-visual layout planning, information structuring and rendering.

In this dimension, we first collected 200 real-world images featuring rich text from different datasets. Using an MLLM, we then extracted the underlying

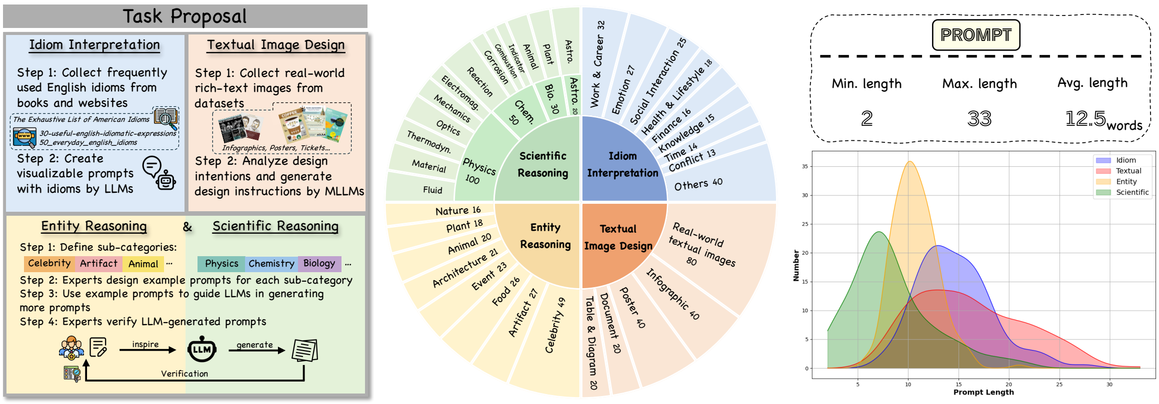


Figure 2: ***Left: Prompt Generation Pipeline.*** Idiom Interpretation: collect idioms from established sources and use LLMs to create prompts. Textual Image Design: source real-world rich-text images, use MLLMs to analyze design intentions and generate corresponding instructions. Entity & Scientific Reasoning: A Human-in-the-loop process: define subcategories, expert prompt design, LLM expansion into larger sets, and final expert verification to ensure accuracy. ***Middle: Distribution of Subcategories.*** The inner ring represents the four dimensions, the outer rings display subcategories and their sample counts. ***Right: Prompt Suite Statistics.*** Top: Statistics about prompt length. Bottom: The distribution of prompt lengths across the four dimensions.

design intentions from these images, resulting in 200 design prompts. Each focuses on the functional purpose of the image rather than describing visual details. For example, "Design an infographic on online risks for children aged 9-16". Based on the image sources, the prompts span categories like infographics, posters, documents, and diagrams.

Entity Reasoning. In everyday life, people often forget specific entity names but remember related details. For example, the prompt "Generate an image of the team lifting the trophy at the 2022 FIFA World Cup" requires the T2I model to reason about the context and then retrieve relevant knowledge to generate the entities not explicitly stated.

In Entity Reasoning, we begin by defining subdomains for various entities, such as celebrities, artifacts, and architectures. We manually create several example prompts to guide an LLM in generating more prompts. After collecting 200 such prompts, we carefully review them to ensure overall consistency and correctness.

Scientific Reasoning. Creating scientifically realistic images remains a persistent challenge for T2I models, which often produce counterintuitive results that violate scientific laws. This highlights the need to evaluate models' awareness of scientific implications in image generation. To design prompts in this dimension, we first identify four key scientific disciplines: physics, chemistry, biology, and astronomy, then create several example prompts manually. We use these examples to inspire the LLM to generate more. Finally, each prompt is manually validated to ensure it requires reasoning with scientific knowledge and the expected visual outcome is not explicitly stated. For instance, the prompt "A trampoline with an iron ball on it" implies that the heavy iron ball would deeply stretch the surface of the trampoline due to its weight.

Figure 2 demonstrates the prompt collection process (left), shows the subcategories in each dimension (middle) and provides the prompt suite statistics (right). We visualize the word distribution in Figure 8. For more information about the prompt suite, please refer to Appendix B.

4 Evaluation

4.1 Evaluation Metric

In recent years, MLLMs have demonstrated remarkable capabilities in understanding complex visual content, becoming the primary tool for evaluating visual contents. However, the prompts in our benchmark are highly complex, often involving multiple objects, intricate relationships, and challenging scenarios. As a result, using generic evaluation instructions that are identical for all prompts proves ineffective. This is because each image, generated from a unique prompt, demands specific and targeted checks that generic instructions cannot provide. To address this, we develop a two-stage evaluation framework with customized evaluation instructions for each prompt. These instructions take into account the prompt category, the reasoning needed, the explicit content the image should exhibit. Figure 3 illustrates the evaluation process.

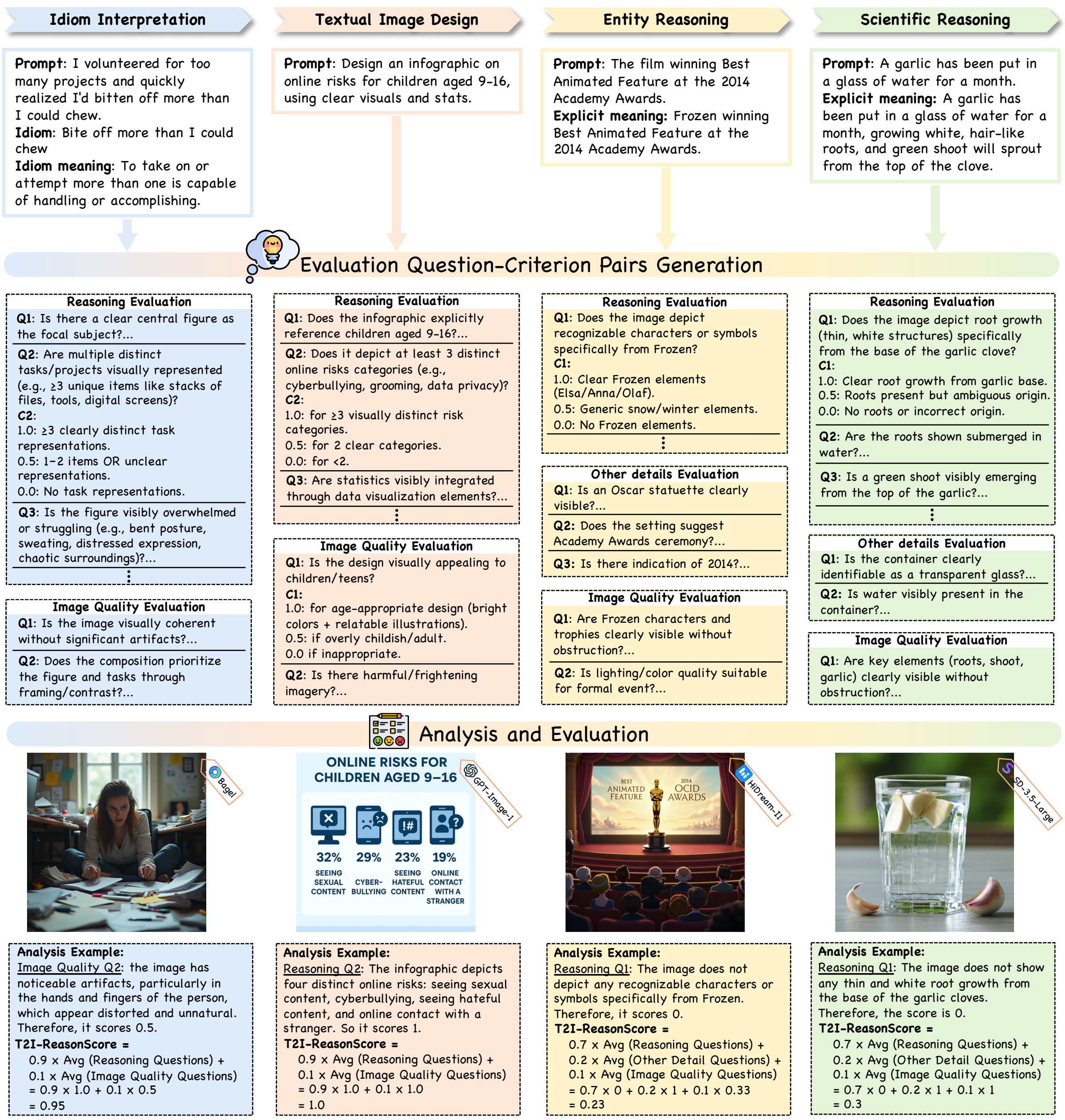


Figure 3: **Evaluation Framework of T2I-ReasonBench.** We adopt a two-stage evaluation framework: prompt-specific evaluation question-criterion pairs generation by an LLM, then image analysis and scoring by an MLLM. This figure shows one evaluation example for each dimension.

Prompt-specific question-criterion pairs generation. In the first stage, we use an LLM to generate question-criterion pairs based on the prompt and dimension-specific assistant information (e.g., idiom meaning for Idiom Interpretation and explicit meaning for Entity or Scientific Reasoning). For each dimension, two sets of questions are generated to separately examine the reasoning accuracy and the image quality. For Entity and Scientific Reasoning, where prompts may involve explicit details that do not need reasoning, an additional set of questions examines these details. Examples are shown in the upper part of Figure 3, with sample questions (Q1, Q2...) and criteria (C1, C2...) in different evaluation sets for each prompt.

Image-level evaluation score. In the second stage, we employ an MLLM to evaluate the generated images with a Chain-of-Thought (Wei et al., 2022) (CoT) mechanism: the model first describes the image, then answers the specific questions posed in Stage 1. For each question, the MLLM provides an analysis prior to assigning a score, ensuring thorough and reliable evaluation. Scores within each set are averaged:

$$S_j = \frac{\sum_{i=1}^{n_j} score_i}{n_j}, \tag{1}$$

where S_j is the score for the jth question set and n_j is the question count for that set. The question sets cover reasoning, detail, and image quality. The final metric, which we call 'T2I-ReasonScore', is calculated as:

$$\text{T2I-ReasonScore} = w_1 S_{reason} + w_2 S_{detail} + w_3 S_{quality} \tag{2}$$

The lower part of Figure 3 illustrates this process to obtain the image-level evaluation score.

In this way, our evaluation metrics reflect the reasoning challenges and provide a comprehensive assessment. For more details of our evaluation framework, including how we generate the questions and criteria, please refer to Appendix C.

4.2 Human Evaluation Correlation Analysis

To validate the effectiveness of 'T2I-ReasonScore', we conduct a human evaluation study to assess its correlation with human judgment. We randomly sample 20 prompts from each of the four dimensions. Using 8 distinct models, we generate 160 images per dimension, yielding a total of 640 images. These images are evaluated by a cohort of postgraduate students based on dimension-specific criteria. Each image is independently scored by 3 annotators, and the results are averaged to establish a gold-standard human score. Then we calculate the correlation between automatic metrics and the human scores using Kendall's τ and Spearman's ρ. To determine the most effective model configuration for our metric, we compute T2I-ReasonScore using 12 distinct LLM-MLLM combinations. For question-criterion generation, we employ three different LLMs: DeepSeek-R1 (Guo et al., 2025), GPT-5.1 (OpenAI, 2025), and Gemini-2.5-pro (Deepmind, 2025). For image evaluation, we use five MLLMs: Qwen2.5-VL-72B (Bai et al., 2025b), GPT-5.1, LLaVA-OneVision-1.5 (An et al., 2025), Gemini-2.5-pro, and Qwen3-VL-235B-A22B (Bai et al., 2025a). Finally, to assess the robustness of our two-stage evaluation pipeline, we compare our metric against two widely-used T2I metrics: CLIPscore (Hessel et al., 2021) and VQAscore (Lin et al., 2024).

The correlation results, presented in Table 1, reveal two key findings: first, the proposed metric consistently outperforms traditional text-image alignment metrics (CLIPscore and VQAscore). Second, Gemini-2.5-pro proves to be a stable and high-performing MLLM evaluator, provided it is paired with a capable LLM in the preceding stage.

Table 1: **Correlation between automatic metrics and human evaluation.** We use different LLM-MLLM combinations to calculate our metric 'T2I-ReasonScore'. For each dimension, **Bold** signifies the highest correlation, underline denotes the second highest. We highlight the combination adopted for each dimension in yellow .

LLM	MLLM	Idiom		Textual		Entity		Scientific		Average	
		$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$
T2I-ReasonScore											
Deepseek	Qwen2.5-VL	0.5095	0.6792	0.6051	0.7686	0.4767	0.6242	0.4984	0.6438	0.5224	0.6790
Deepseek	GPT-5.1	0.4943	0.6659	0.6052	0.7493	0.5432	0.6831	0.4704	0.6168	0.5283	0.6788
Deepseek	LLaVA-1V	0.4664	0.6202	0.5561	0.6936	0.5116	0.6377	0.4342	0.5802	0.4921	0.6329
GPT-5.1	Qwen2.5-VL	0.5391	0.7004	0.6372	0.8002	0.5937	0.7431	0.5062	0.6592	0.5691	0.7257
GPT-5.1	GPT-5.1	0.5576	0.7168	0.6446	0.8076	0.6077	0.7642	0.4589	0.6098	0.5672	0.7246
GPT-5.1	LLaVA-1V	0.5165	0.6594	0.5343	0.6619	0.5762	0.7074	0.3952	0.5130	0.5055	0.6354
GPT-5.1	Gemini-2.5-pro	0.5540	0.7115	0.6175	0.7848	0.6102	0.7605	**0.5673**	**0.7231**	**0.5872**	**0.7450**
GPT-5.1	Qwen3-VL	**0.5596**	**0.7257**	0.6324	0.7913	0.5611	0.7114	0.5285	0.6778	0.5704	0.7265
Gemini-2.5-pro	Qwen2.5-VL	0.4822	0.6320	**0.6829**	**0.8224**	0.6344	0.7882	0.4245	0.5641	0.5560	0.7017
Gemini-2.5-pro	GPT-5.1	0.5218	0.6672	0.6458	0.7896	**0.6514**	**0.7968**	0.4343	0.5781	0.5633	0.7079
Gemini-2.5-pro	Gemini-2.5-pro	0.4975	0.6566	0.6096	0.7641	0.6390	0.7830	0.4471	0.5918	0.5483	0.6989
Gemini-2.5-pro	Qwen3-VL	0.4693	0.6132	0.6527	0.7947	0.6071	0.7399	0.5109	0.6545	0.5600	0.7006
Other metrics											
CLIPscore (Hessel et al., 2021)		0.3186	0.4348	0.5372	0.7187	0.2732	0.3837	0.1905	0.2657	0.3299	0.4507
VQAscore (Lin et al., 2024)		0.4091	0.5672	0.4890	0.6590	0.4483	0.6133	0.3698	0.4939	0.4291	0.5834

The average correlation is the highest overall when pairing Gemini-2.5-pro with questions from GPT-5.1. However, in Entity Reasoning, Gemini-2.5-pro with questions generated by itself yields higher correlation. Therefore, we select Gemini-2.5-pro as our unified MLLM evaluator and employ a hybrid strategy for question generation: Gemini-2.5-pro for Entity Reasoning and GPT-5.1 for the other three dimensions. Figure 7 in Appendix A shows more qualitative examples generated by different T2I models with the evaluated T2I-ReasonScore. It can be seen that the scores are highly aligned with human preference. For more details about human evaluation, please refer to Appendix C.

5 Evaluation Results

Evaluated models. We evaluate 16 state-of-the-art T2I models, including 8 diffusion models, 5 unified multimodal models, and 3 proprietary models. The diffusion models are HiDream-I1-full (hidream, 2024), FLUX.1-dev (Labs, 2024), FLUX.1-schnell (Labs, 2024), Playground-v2.5 (Li et al., 2024b), Stable-Diffusion-3-Medium (Esser et al., 2024b), Stable-Diffusion-3.5-Medium (Esser et al., 2024b), Stable-Diffusion-3.5-Large (Esser et al., 2024b), Qwen-Image (Wu et al., 2025). The unified models are: Bagel (Deng et al., 2025), Emu3 (Wang et al., 2024), Janus-Pro-7B (Chen et al., 2025d), show-o-demo-512 (Xie et al., 2024), GoT (Fang et al., 2025). The proprietary models are Gemini-2.0 (Team et al., 2023), GPT-Image-1 (OpenAI, 2023), Nano-Banana (Google, 2025).

Implementation Details. We use the default set-

Table 2: **Evaluation results of T2I-ReasonBench.** Scores are normalized between 0-100. A higher score indicates better performance. Blue highlights the top score in diffusion models. Yellow highlights the top score in unified multimodal models. **Bold** signifies the highest score of all models.

Model	Idiom	Textual	Entity	Scientific	Overall
	Diffusion Models				
SD-3-Medium (Esser et al., 2024b)	41.7	66.9	43.1	56.6	52.1
SD-3.5-Medium (Esser et al., 2024b)	39.6	66.0	47.6	55.7	52.2
SD-3.5-Large (Esser et al., 2024b)	44.2	66.5	50.9	59.3	55.2
FLUX.1-dev (Labs, 2024)	49.3	67.9	50.5	51.8	54.9
FLUX.1-schnell (Labs, 2024)	47.4	74.2	50.1	58.7	57.6
Playground-v2.5 (Li et al., 2024b)	50.8	45.5	51.1	54.4	50.5
HiDream-I1-full (hidream, 2024)	59.1	80.3	58.0	59.4	64.2
Qwen-Image (Wu et al., 2025)	62.7	77.1	61.8	68.2	67.5
	Unified Multimodal Models				
Emu3 (Wang et al., 2024)	39.2	40.2	42.7	44.6	41.7
show-o-demo-512 (Xie et al., 2024)	38.9	43.2	41.0	48.3	42.8
Janus-Pro-7B (Chen et al., 2025d)	32.7	42.9	41.3	52.4	42.3
GoT (Fang et al., 2025)	33.7	40.7	37.6	41.3	38.3
Bagel w/ Thinking (Deng et al., 2025)	50.6	50.5	56.5	64.9	55.6
	Proprietary Models				
Gemini-2.0 (Team et al., 2023)	65.5	77.9	78.0	76.3	74.4
GPT-Image-1 (OpenAI, 2023)	84.1	92.4	85.7	82.9	86.3
Nano-Banana (Google, 2025)	**89.8**	**94.8**	**87.3**	**86.0**	**89.5**

tings for all T2I models in image generation. We set the weights $[w_1, w_2, w_3]$ in T2I-ReasonScore to $[0.9, 0.0, 0.1]$ for Idiom Interpretation and Textual Image Design, and $[0.7, 0.2, 0.1]$ for Entity and Scientific Reasoning to prioritize reasoning while maintaining a balanced final score.

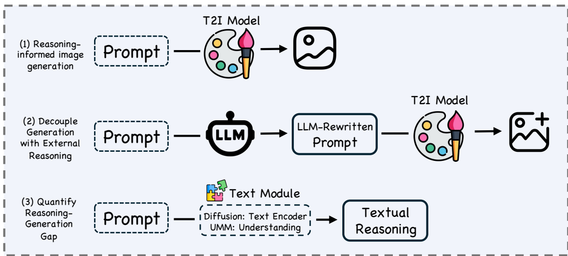


Figure 4: **Overview of the Experiment Settings.** The first experiment (§5.1) establishes a baseline, where images are generated directly from the original benchmark prompts. The second (§5.2) introduces an external LLM to reason and rewrite the prompts before they are sent to the T2I models for image generation. The third experiment (§5.3) isolates and evaluates the models' internal reasoning by using their text encoders or understanding mode to generate textual reasoning, the quality of which is measured as a text score.

5.1 Quantitative Evaluation

The evaluation results on T2I-ReasonBench, presented in Table 2, reveal significant limitations and pronounced performance gaps in current T2I models. Proprietary models demonstrate a clear lead, with scores consistently exceeding 70 points and Nano-Banana approaching 90. In contrast, open-source models lag far behind: specialized diffusion

Table 3: **Evaluation results with LLM-rewritten prompts.** ↑ Red and ↓ Blue subscripts denote change in scores when using LLM-rewritten prompts.

Model	Idiom	Textual	Entity	Scientific	Overall
	Diffusion Models				
SD-3-Medium	76.7$_{\uparrow 34.9}$	74.1$_{\uparrow 7.2}$	75.0$_{\uparrow 32.0}$	68.5$_{\uparrow 11.9}$	73.6$_{\uparrow 21.5}$
SD-3.5-Medium	76.4$_{\uparrow 36.8}$	73.1$_{\uparrow 7.1}$	76.4$_{\uparrow 28.8}$	69.5$_{\uparrow 13.9}$	73.8$_{\uparrow 21.7}$
SD-3.5-Large	76.7$_{\uparrow 32.5}$	74.2$_{\uparrow 7.7}$	82.1$_{\uparrow 31.1}$	71.4$_{\uparrow 12.1}$	76.1$_{\uparrow 20.9}$
FLUX.1-dev	77.8$_{\uparrow 28.5}$	78.0$_{\uparrow 10.1}$	75.6$_{\uparrow 25.1}$	71.6$_{\uparrow 19.9}$	75.8$_{\uparrow 20.9}$
FLUX.1-schnell	79.9$_{\uparrow 32.5}$	78.2$_{\uparrow 4.0}$	76.6$_{\uparrow 26.5}$	74.4$_{\uparrow 15.7}$	77.3$_{\uparrow 19.7}$
Playground-v2.5	65.3$_{\uparrow 14.5}$	48.9$_{\uparrow 3.5}$	77.4$_{\uparrow 26.4}$	57.3$_{\uparrow 2.8}$	62.2$_{\uparrow 11.8}$
HiDream-I1-full	77.0$_{\uparrow 17.9}$	81.8$_{\uparrow 1.4}$	79.6$_{\uparrow 21.6}$	71.6$_{\uparrow 12.1}$	77.5$_{\uparrow 13.2}$
Qwen-Image	84.7$_{\uparrow 22.0}$	83.0$_{\uparrow 5.9}$	84.2$_{\uparrow 22.4}$	83.7$_{\uparrow 15.5}$	83.9$_{\uparrow 16.4}$
	Unified Multimodal Models				
Emu3	67.7$_{\uparrow 28.5}$	48.8$_{\uparrow 8.5}$	67.6$_{\uparrow 24.9}$	57.3$_{\uparrow 12.8}$	60.3$_{\uparrow 18.7}$
show-o-demo-512	74.8$_{\uparrow 35.8}$	48.6$_{\uparrow 5.4}$	69.4$_{\uparrow 28.4}$	65.2$_{\uparrow 16.9}$	64.5$_{\uparrow 21.6}$
Janus-Pro-7B	72.6$_{\uparrow 39.9}$	57.1$_{\uparrow 14.2}$	71.0$_{\uparrow 29.7}$	68.6$_{\uparrow 16.2}$	67.3$_{\uparrow 25.0}$
GoT	62.3$_{\uparrow 28.7}$	45.9$_{\uparrow 5.2}$	55.2$_{\uparrow 17.6}$	50.2$_{\uparrow 8.9}$	53.4$_{\uparrow 15.1}$
Bagel w/o Thinking	77.9$_{\uparrow 27.4}$	65.4$_{\uparrow 14.9}$	73.9$_{\uparrow 17.4}$	74.0$_{\uparrow 9.1}$	72.8$_{\uparrow 17.2}$
	Proprietary Models				
Gemini-2.0	80.3$_{\uparrow 14.8}$	80.8$_{\uparrow 2.9}$	86.6$_{\uparrow 8.6}$	82.5$_{\uparrow 6.3}$	82.5$_{\uparrow 8.1}$
GPT-Image-1	87.1$_{\uparrow 3.1}$	89.2$_{\downarrow 3.2}$	**90.8**$_{\uparrow 5.0}$	**87.3**$_{\uparrow 4.4}$	88.6$_{\uparrow 2.3}$
Nano-Banana	**87.7**$_{\downarrow 2.2}$	**92.2**$_{\downarrow 2.6}$	90.2$_{\uparrow 2.9}$	86.5$_{\uparrow 0.6}$	**89.2**$_{\downarrow 0.3}$

models typically score around 50; unified multimodal models are even lower, with most around 40 points, though Bagel is a notable exception, performing on par with the diffusion models.

5.2 Decoupling Generation via External Reasoning

We conduct an additional experiment using a pipeline that decouples image generation from reasoning, as shown in (2) of Figure 4. In this setup, GPT-4o (Hurst et al., 2024) first reasons about the original prompt and converts it into a visually explicit description, which is then fed to the T2I model. Table 3 presents the quantitative results of using the LLM-rewritten prompts.

Reasoning is the bottleneck for model performance. By providing models with pre-reasoned, visually explicit prompts, this pipeline setting allows us to isolate and clarify their actual T2I generation capability. The results show that this approach substantially improves the accuracy of reasoning for open-source models. The more a model improves, the more it initially lacks reasoning ability, as opposed to generation ability. As a result, the performance gap between open-source and proprietary models shrinks with decoupled reasoning, suggesting that the previous disparity in Table 2 stems primarily from differences in reasoning ability.

5.3 Quantifying the Internal Reasoning-Generation Gap

We design an experiment to measure the reasoning capability of each model's text module and quantify the internal reasoning-generation gap. It in-

Table 4: **Comparison of Reasoning Ability in Image Generation vs. Text-Only Interpretation.** Scores are presented as *Image Score/Text Score*. The *Text Score* is derived by evaluating an explicit description generated by the model's isolated text module. The *Image Score* evaluates the image generated by the full model. Both scores are composed of $w_1 S_{reason} + w_2 S_{detail}$, which excludes the image quality component, for a maximum possible value of 90. **Bold** marks the lowest scores.

Model/Text-Module	Idiom	Textual	Entity	Scientific
Diffusion Models				
HiDream-I1-full/Llama-3.1-8b	50.3/74.1	71.2/81.5	49.1/66.8	50.9/69.6
Qwen-Image/Qwen2.5-VL-7b	53.1/70.2	69.2/85.1	52.5/63.0	59.3/70.5
Unified Multimodal Models				
Emu3-Gen/Emu3-Und	**31.3/36.9**	**34.9/80.3**	**34.9/56.9**	**37.7/63.2**
Bage-Gen/Bagel-Und	41.7/52.0	43.2/84.4	47.8/67.6	56.0/76.4
Proprietary Models				
GPT-Image-1/GPT-4o	74.3/78.2	82.7/87.6	76.0/73.6	73.2/77.4
Nano-Banana-Gen/Nano-Banana-Und	80.1/76.7	85.5/86.8	77.7/75.1	76.2/80.9

volves isolating the text understanding component and having it generate a textual description with concrete visual details for each prompt, as illustrated in (3) of Figure 4. This textual output is then evaluated using our metric (without evaluating the image quality) to produce a "text score", as shown in Table 4. This score reflects the pure reasoning ability of the text module, representing a theoretical upper limit on the model's performance if its reasoning were perfectly utilized by generation. The difference between this text score and the model's final image score provides an approximation of its reasoning-generation gap.

Proprietary models demonstrate superior reasoning ability and minimal reasoning-generation gap. Proprietary models achieve the highest text reasoning scores, representing their superior ability to understand reasoning-intensive prompts. Gemini-2.0 (Team et al., 2023), for example, verbalizes its reasoning before generation. For the prompt "The city hosting the Summer Olympics in 2021", it explicitly identifies Tokyo and then plans to include its landmarks and Olympic imagery. While other top models like GPT-image-1 and Nano-Banana do not provide such textual response, the minimal gap between their text scores and image scores demonstrate a highly efficient transfer of reasoning to synthesis. Our experiment further validates this conclusion. As shown in Figure 5, there is a high degree of alignment between the final images in the third column and the correct textual reasoning that we prompt the models to generate in a separate query. It demonstrates that the models' internal reasoning is stable and that the generation process fully integrates this reasoning, reflecting it in the final image without information loss.

Leading diffusion models still limited by semantic mapping. The cases of Qwen-Image (Wu et al., 2025) and HiDream (hidream, 2024) reveal a critical reasoning-generation gap in top-tier diffusion models. Both incorporate powerful LLMs as text encoders (Qwen2.5-VL-7b and Llama 3.1-8b), whose reasoning abilities are comparable to proprietary models. However, this advantage is lost during image synthesis. As shown in Figure 6 (Row#1,#2), while the models are able to generate correct imagery from LLM pre-reasoned prompts, they fail with the original implicit prompts. In the latter case, even though the isolated text encoders can correctly interpret them, the generated images only reflect a superficial keyword mapping, leading to incorrect outputs. This demonstrates that the models fail to leverage the deep reasoning capabilities of their LLM encoders. Instead, they function merely as semantic mappers rather than integrated reasoners, which ultimately limits the models' performance on reasoning-aware image generation.

Unified multimodal models face both reasoning and generation challenges. Early UMM like Emu3 struggles with both reasoning and generation. As shown in Figure 6 (Row #4). Rather than analyzing the scientific implications, its textual output merely extends the original prompt with more descriptive details. This reasoning deficit is compounded by poor generation quality, as the model fails to compose all objects specified in the prompt. More advanced UMM like Bagel exhibits stronger reasoning capabilities. Its bottleneck-free architecture unifies LLM and diffusion model within a single transformer, enhancing interaction between understanding and generation modules. It can analyze a prompt's implicit meaning and verbalize its reasoning before generation (Figure 7, Row #3). However, even when its textual reasoning is correct, the final generated image may fail to align with that plan, indicating a failure to transfer its understanding into the visual synthesis.

Insights for future model design. The last two rows in Figure 6 highlight two distinct failure cases in proprietary models: one of reasoning and one of integration. The first is a reasoning failure (Row #5), where Nano-Banana's textual analysis fails to uncover the prompt's physical implications, leading directly to an incorrect image. The second is a reasoning-integration failure (Row #6). Here, GPT-4o's textual reasoning correctly recognizes the entity, yet the model neglects the implicit meaning when generating the image. These examples highlight a key challenge: improving reasoning-aware text-to-image generation requires not only stronger foundational reasoning in current T2I models, but also more effective integration of that reasoning into the final image synthesis.

6 Conclusion

We introduce T2I-ReasonBench, a benchmark for evaluating reasoning-informed text-to-image generation. Unlike prior benchmarks that mainly emphasize literal prompt following, T2I-ReasonBench focuses on whether models can infer implicit meaning, apply world knowledge, and translate reasoning outcomes into coherent visual outputs. To support this goal, we curate 800 prompts across four challenging dimensions: Idiom Interpretation, Textual Image Design, Entity Reasoning, and Scientific Reasoning, and propose a prompt-specific two-stage evaluation framework with T2I-ReasonScore for fine-grained and interpretable assessment.

Through a comprehensive study of 16 state-of-the-art models, we show that current T2I systems still face substantial limitations in reasoning-aware generation. Our results reveal two major bottlenecks: first, many models struggle to correctly reason about prompts containing implicit meaning; second, even when a model's text understanding module is capable of correct reasoning, this understanding is often not faithfully reflected in the final generated image.

Overall, T2I-ReasonBench highlights that high image quality and literal text alignment are not sufficient to measure true text-to-image intelligence. Future progress in T2I generation will require both stronger foundational reasoning and tighter integration between reasoning and visual synthesis. We hope this benchmark can serve as a useful testbed for the community, encouraging the development of next-generation T2I systems that not only generate realistic images, but also genuinely understand what they are asked to depict.

7 Societal Impacts

This work may support the development of more reliable and reasoning-aware text-to-image models, which could benefit applications such as education, design, and scientific communication. At the same time, stronger reasoning ability may also enable the generation of more convincing misleading or deceptive images, including fake advertisements, documents, or propaganda. The benchmark itself may also introduce risks. By emphasizing reasoning performance, it could incentivize building more capable generative systems without corresponding safeguards. Overall, we view T2I-ReasonBench as an evaluation tool, and we believe progress on reasoning-aware generation should be accompanied by appropriate safeguards.

Limitations

This work has several limitations.

First, although T2I-ReasonBench covers four important reasoning dimensions, it does not exhaust the full spectrum of reasoning required in T2I generation. Other forms of reasoning could be explored in future extensions.

Second, while we design a prompt-specific two-stage evaluation framework and show that T2I-ReasonScore correlates well with human judgments, the evaluation still relies on LLMs and MLLMs as automatic judges. Such evaluators may themselves introduce bias, make recognition errors, or favor certain visual styles and model outputs. Although human correlation analysis supports the reliability of our metric, automatic evaluation cannot fully replace large-scale human assessment.

Third, the benchmark is currently limited in scale and language coverage. Our prompt set contains 800 English prompts, which enables controlled and diverse evaluation, but may not fully represent the richness of real-world user instructions, multilingual usage, or culturally dependent expressions such as region-specific idioms and design conventions. Extending the benchmark to broader languages and cultural contexts would improve its generality.

Fourth, our analysis of the reasoning-generation gap is only conducted on a subset of models. As a result, the conclusions about internal reasoning and its transfer to image synthesis cannot yet be generalized to all T2I systems.

Finally, benchmark construction and prompt verification involve LLM assistance and expert review, which may still leave room for subjective choices in prompt design, and expected visual interpretation. While we carefully validate prompts to ensure quality and consistency, some cases may still admit multiple plausible visual realizations. We hope future work can further expand the benchmark, diversify the reasoning dimensions, and develop even more robust evaluation protocols.

Acknowledgement

This work is partially supported by the National Nature Science Foundation of China (No. 62402406).

References

2023. The exhaustive list of american idioms. https://learn-esl.org/index.html.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, and 1 others. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Haoyu Chen, Xiaojie Xu, Wenbo Li, Jingjing Ren, Tian Ye, Songhua Liu, Ying-Cong Chen, Lei Zhu, and Xinchao Wang. 2025a. Posta: A go-to framework for customized artistic poster generation. *arXiv preprint arXiv:2503.14908*.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, and 1 others. 2025b. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*.

Kaijie Chen, Zihao Lin, Zhiyang Xu, Ying Shen, Yuguang Yao, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. 2025c. R2i-bench: Benchmarking reasoning-driven text-to-image generation. *arXiv preprint arXiv:2505.23493*.

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

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

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023b. Visual programming for step-by-step text-to-image generation and evaluation. *Advances in Neural Information Processing Systems*, 36:6048–6069.

Google Deepmind. 2025. Introducing openai o3 and o4-mini. https://deepmind.google/models/gemini/pro/.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, and 1 others. 2025. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. 2021. Cogview: Mastering text-to-image generation via transformers. *arXiv preprint arXiv:2105.13290*.

Chengqi Duan, Rongyao Fang, Yuqing Wang, Kun Wang, Linjiang Huang, Xingyu Zeng, Hongsheng Li, and Xihui Liu. 2025. Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. *arXiv preprint arXiv:2505.17022*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 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, and 1 others. 2024b. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, and 1 others. 2025. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*.

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*.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *arXiv preprint arXiv:2310.11513*.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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*.

hidream. 2024. hidream. https://github.com/HiDream-ai/HiDream-I1.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

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*.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023a. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023b. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2025. T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 47(05):3563–3579.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Black Forest Labs. 2024. Flux. https://github.com/black-forest-labs/flux.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and 1 others. 2024a. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. 2024b. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*.

Youngsun Lim, Hojun Choi, and Hyunjung Shim. 2025. Evaluating image hallucination in text-to-image generation with question-answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26290–26298.

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.

Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. 2024. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Fanqing Meng, Wenqi Shao, Lixin Luo, Yahong Wang, Yiran Chen, Quanfeng Lu, Yue Yang, Tianshuo Yang, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. Phybench: A physical commonsense benchmark for evaluating text-to-image models. *arXiv preprint arXiv:2406.11802*.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

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*.

OpenAI. 2023. Gpt-image-1. https://openai.com/index/image-generation-api/.

OpenAI. 2025. Gpt-5.1: A smarter, more conversational chatgpt. https://openai.com/index/gpt-5-1/.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, and 1 others. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

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

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, and 1 others. 2024. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. 2025. Tiif-bench: How does your t2i model follow your instructions? *arXiv preprint arXiv:2506.02161*.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, and 1 others. 2025. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. 2024. Conceptmix: A compositional image generation benchmark with controllable difficulty. *Advances in Neural Information Processing Systems*, 37:86004–86047.

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*.

Yue Yang, Ajay Patel, Matt Deitke, Tanmay Gupta, Luca Weihs, Andrew Head, Mark Yatskar, Chris Callison-Burch, Ranjay Krishna, Aniruddha Kembhavi, and 1 others. 2025. Scaling text-rich image understanding via code-guided synthetic multimodal data generation. *arXiv preprint arXiv:2502.14846*.

Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. 2023. What you see is what you read? improving text-image alignment evaluation. *Advances in Neural Information Processing Systems*, 36:1601–1619.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, and 1 others. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5.

Shijie Zhou, Ruiyi Zhang, Yufan Zhou, and Changyou Chen. 2024. A high-quality text-rich image instruction tuning dataset via hybrid instruction generation. *arXiv preprint arXiv:2412.16364*.

Appendix Contents

A Qualitative Examples

Figure 5, 6 compares the qualitative outputs of different T2I models. Figure 7 shows more image outputs from different models with the evaluated T2I-ReasonScore.

B More Details on Prompt Collection Process

Idiom Interpretation. In idiom collection, we leverage a book titled "The Exhaustive List of American Idioms" (idi, 2023), which systematically documents over 11k idioms. These idioms were collected from diverse sources, including TV shows, movies, and everyday conversations. Each idiom in the book is accompanied by its actual meaning in context. In addition to this resource, we also refer to idioms available on the Internet. From this extensive pool, we manually select 200 idioms that are commonly used in daily life and challenging for T2I models due to their figurative meanings. We input the selected idioms and their actual meanings into an LLM and prompt it to generate new sentences. These sentences are designed to describe visible scenes involving the idioms, providing contextual clues for reasoning while avoiding directly revealing the idiom's meaning.

Textual Image Design. For textual image design, we collect 6 types of text-rich images from 4 distinct sources.

(1) LLAVAR-2 Dataset (Zhou et al., 2024): This dataset contains 42k text-rich images sourced from LAION (Schuhmann et al., 2022), representing various categories such as quotes, memes, book covers, posters, and product packaging. However, images in this dataset are of various quality and formats, so we filter out 80 high-quality images that have a resolution greater than 384x384 and exhibit clear design intentions.

(2) InfographicVQA Dataset (Mathew et al., 2022): This dataset comprises 5k high-quality infographics. We select 40 with normal height-width ratio that exemplify well-crafted layouts to convey structured information.

(3) POSTA Dataset (Chen et al., 2025a): This dataset includes over 300 posters with professional background, layout, and text formats designed by experts. We select 40 posters that demonstrate a balance between text and visual design elements.

(4) CoSyn-400k Dataset (Yang et al., 2025): This dataset consists of 400k synthetic text-rich images, generated by LLM-drive codes. These images cover diverse formats, such as charts, diagrams, tables, documents (e.g., menus or business cards), math examples, and musical scores. From this dataset, we select 40 samples that exemplify structured text-visual integration, including 10 tables, 10 diagrams, and 20 documents.

The design intentions of the collected images are summarized using Qwen2.5-VL (Bai et al., 2025b), yielding 200 prompts in this dimension.

Entity Reasoning and Scientific Reasoning. Our prompt generation pipeline for each of these two dimensions consists of three stages: initial design, LLM-based expansion, and expert verification. The process begins with human experts defining key subcategories (e.g., different scientific disciplines, or well-known entities with clear visual features). For each subcategory, the experts then authored a set of high-quality "seed" prompts to serve as a foundation. Next, these seed prompts were used as exemplars to prompt multiple LLMs to generate a larger, more diverse corpus of prompts. The LLMs also generated an explicit description of the expected visual phenomena to assist with evaluation. Finally, all generated prompts and their explicit descriptions underwent a rigorous verification by human experts. This review includes consulting reliable sources to fact-check all domain-specific information and ensure its accuracy.

We visualize the word distribution of each dimension in our prompt suite in Figure 8.

C Evaluation Framework

C.1 Templates

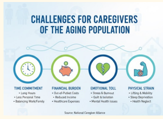
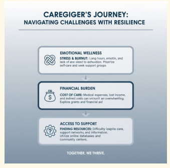
Templates for evaluation stage I: LLM question-criterion generation. Table 9, 10, 11, and 12 present the templates used to generate the prompt-specific question-criterion pairs for Idiom Interpretation, Textual Image Design, Entity-Reasoning and Scientific-Reasoning, respectively. Each template is tailored to focus on the unique aspects of its corresponding dimension. For example, when generating the question-criterion pairs for Idiom Interpretation, the LLM receives the prompt together with the idiom included, and its actual meaning to ensure it has the full background information to generate accurate questions and criteria. For Entity Reasoning, the explicit prompt with the specific entity name is provided. For Scientific Reasoning, the explicit prompt specifying the visible outcome derived from the scientific law is provided. As for Textual Image Design, rather than prescribing visual content, the core of evaluation is to assess whether generated images fulfill their intended communicative function. Therefore, only the original prompt is provided, this allows the T2I models to have full design freedom. The generated evaluation questions and criteria for the prompt "Create a minimalist promotional poster for a workshop on simplicity in design" are shown in Table 13, we can see they focus on evaluating functional elements including:

- Clear presentation of workshop topic
- Explicit identification as an event
- Inclusion of essential details (date/time/location)
- Consistency with minimalist design principles

Templates for evaluation stage II: MLLM image evaluation. Table 14 presents the template used to evaluate the generated images for all four dimensions. Only the evaluation question-criterion pairs need to be replaced for each prompt.

C.2 Human Evaluation

The human evaluation is conducted on eight models: Stable-Diffusion-3-Medium, FLUX.1-schnell, HiDream-I1-full, Qwen-Image, Bagel, Janus-Pro-7B, GPT-Image-1, and Nano-Banana (4 diffusion models, 2 unified multimodal models, 2 proprietary model). We randomly select 20 prompts from each dimension and use these 8 models to generate 160 images per dimension. This results in 640 images in total for human evaluation. The evaluation is conducted by a group of volunteer postgraduate students. Three participants independently score each image, and we average their scores for each prompt-image pair. The evaluation criteria provided to the human annotators are shown in Table 15, 16, 17, and 18. These criteria are tailored to each dimension in order to address different aspects. The correlations are calculated between the averaged human scores and the automatic metric score using Kendall's τ and Spearman's ρ. The results are shown in Table 1.

	Prompt from Textual Image Design	Textual Reasoning by Nano-Banana	Image generated by Nano-Banana with original prompt	LLM-Rewritten Prompt	Image generated by Nano-Banana with LLM-rewritten prompt
#1	Create an infographic highlighting the challenges faced by caregivers for the aging population.	...The infographic should feature distinct sections or icons representing different challenges, such as financial strain (e.g., piggy bank), emotional burnout, and physical demands. Include relevant statistics or short descriptive texts next to each visual element...		A vertical infographic with a bold, engaging title, icons for caregiver challenges (such as stress, finances, and resources), calming blue-gray color tones, and neatly organized text sections to educate on the topic.	

	Prompt from Entity Reasoning	Textual Reasoning by Nano-Banana	Image generated by Nano-Banana with original prompt	LLM-Rewritten Prompt	Image generated by Nano-Banana with LLM-rewritten prompt
#2	The Southeast Asian fruit with pungent odor and custard-like texture	A close-up photograph of a durian fruit, cut open to reveal its creamy, custard-like, pale yellow flesh. The fruit has a spiky, green-brown exterior. The texture of the flesh should be clearly visible, smooth and rich.		A freshly split open durian with its spiky, greenish-brown husk and bright yellow custard-like flesh displayed prominently in a bustling Southeast Asian marketplace, surrounded by other tropical fruits like bananas and mangosteens.	

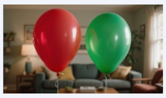
	Prompt from Scientific Reasoning	Textual Reasoning by Nano-Banana	Image generated by Nano-Banana with original prompt	LLM-Rewritten Prompt	Image generated by Nano-Banana with LLM-rewritten prompt
#3	A heavy rock placed on a beach ball	A large, jagged, grey rock is resting on top of a colorful, inflated beach ball. The rock is heavy enough to visibly indent the surface of the beach ball, causing it to flatten. The beach ball is situated on a sandy beach, with a clear blue ocean and sky in the background.		A heavy rock compresses a brightly colored beach ball into a flattened shape beneath it, showing distortion in the beach ball's curvature and slight stretching of its colors.	

	Prompt from Idiom Interpretation	Textual Reasoning by GPT-4o	Image generated by GPT-Image-1 with original prompt	LLM-Rewritten Prompt	Image generated by GPT-Image-1 with LLM-rewritten prompt
#4	She hadn't played the piano in years, but it was like riding a bicycle; she picked it up instantly.	The scene depicts a woman sitting at a piano. She has medium-length dark hair and is wearing a casual, comfortable sweater with jeans. Her expression is a mix of concentration and joy as she looks down at the piano keys with her fingers gracefully poised above them.		A woman is seated at a piano. The room includes bookshelves, framed photos, and a teapot on a small table behind her. She's dressed casually with rolled-up sleeves. She looks surprised and pleased, her fingers moving naturally over the piano keys...	

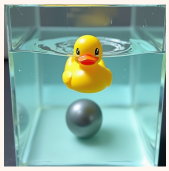
	Prompt from Entity Reasoning	Textual Reasoning by GPT-4o	Image generated by GPT-Image-1 with original prompt	LLM-Rewritten Prompt	Image generated by GPT-Image-1 with LLM-rewritten prompt
#5	The ancient stone circle in Wiltshire, England	A wide-angle view of the ancient stone circle in Wiltshire, England, known as Stonehenge. The scene shows a grassy plain with a circle of massive, upright stones arranged in a circular formation. The sky is partly cloudy with patches of blue...		A prehistoric stone circle known as Stonehenge, featuring large, standing stones arranged in a circular formation on a green, grassy plain at Salisbury Plain, Wiltshire, England. The time is dawn, with soft light casting shadows across the stones.	

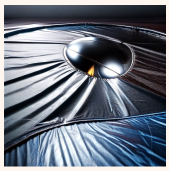
	Prompt from Scientific Reasoning	Textual Reasoning by GPT-4o	Image generated by GPT-Image-1 with original prompt	LLM-Rewritten Prompt	Image generated by GPT-Image-1 with LLM-rewritten prompt
#6	A seesaw with a 1 cubic decimeter block of cotton on one end and 1 cubic decimeter iron cube on the other	A scene focusing on a large yellow bus abruptly stopping. The bus is full and some passengers are standing. The standing passengers are visibly unsteady, leaning forward with surprised expressions, grasping onto overhead rails and support poles.		A bus interior is shown in detail. Standing passengers lean or stumble forward due to sudden deceleration. Some grab the railings or straps for balance. A few seated passengers are startled, leaning forward slightly.	

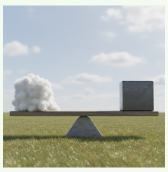
Figure 5: Qualitative examples with original prompt and generated image, textual reasoning by text module, LLM-rewritten prompt and generated image.

#1	Prompt from Scientific Reasoning	Textual Reasoning by Qwen2.5-VL-7b	Image generated by Qwen-Image with original prompt	LLM-Rewritten Prompt	Image generated by Qwen-Image with LLM-rewritten prompt
	A red helium balloon and a green water balloon in a living room.	A red helium balloon floating in the air with some strands of string visible, alongside a green water balloon on the floor, surrounded by a cozy living room setting with plush furniture and homey decor		A living room with a red helium balloon floating near the ceiling, and a green water balloon resting on the floor. The two balloons showcase distinct positions due to their interaction with gravity and air buoyancy.	

#2	Prompt from Entity Reasoning	Textual Reasoning by Llama-3.1-8b	Image generated by Hidream with original prompt	LLM-Rewritten Prompt	Image generated by Hidream with LLM-rewritten prompt
	The city hosting the Summer Olympics in 2008	A cityscape of Beijing with the Bird's Nest Stadium in the foreground, surrounded by Olympic venues and athletes, set against a clear blue sky with subtle Beijing city signs or elements blurred to focus on the main activity.		A living room with a red helium balloon floating near the ceiling, and a green water balloon resting on the floor. The two balloons showcase distinct positions due to their interaction with gravity and air buoyancy.	

#3	Prompt from Entity Reasoning	Textual Reasoning by Bagel-Und	Image generated by Bagel with original prompt	LLM-Rewritten Prompt	Image generated by Bagel with LLM-rewritten prompt
	A rubber duck and a metal ball in a water tank	A clear depiction of a water tank filled with transparent water, showcasing a rubber duck floating effortlessly on the surface and a metal ball submerged at the bottom, illustrating the contrasting buoyancy and density.		In a transparent water tank, a yellow rubber duck floats near the surface, while a gray metal ball sits at the bottom after sinking. The tank sides are glass, showing clear distinctions between the floating and sunken objects.	

#4	Prompt from Scientific Reasoning	Textual Reasoning by Emu3-Und	Image generated by Emu3 with original prompt	LLM-Rewritten Prompt	Image generated by Emu3 with LLM-rewritten prompt
	A trampoline with an iron ball on it	...The trampoline is a circular object with a soft, cushioned surface, typically made of a flexible material such as rubber or foam. The iron ball is a heavy, metallic object. The trampoline and the iron ball are both located on a flat surface...		A trampoline with an iron ball on it appears with the trampoline's surface visibly sagging into a deep concave dip due to the ball's mass. The iron ball, metallic and reflective, sits in the center while the trampoline springs stretch outward.	

#5	Prompt from Scientific Reasoning	Textual Reasoning by Nano-Banana	Image generated by Nano-Banana with original prompt	LLM-Rewritten Prompt	Image generated by Nano-Banana with LLM-rewritten prompt
	A seesaw with a 1 cubic decimeter block of cotton on one end and 1 cubic decimeter iron cube on the other	A seesaw is perfectly balanced in the center on a fulcrum. On the left side of the seesaw, there is a 1 cubic decimeter block of cotton, white and fluffy. On the right side of the seesaw, there is a 1 cubic decimeter iron cube, dark grey, metallic, and dense.		A wooden seesaw tilts steeply due to a 1 cubic decimeter iron cube weighing more than a 1 cubic decimeter block of cotton on opposite ends.	

#6	Prompt from Entity Reasoning	Textual Reasoning by GPT-4o	Image generated by GPT-Image-1 with original prompt	LLM-Rewritten Prompt	Image generated by GPT-Image-1 with LLM-rewritten prompt
	The Southeast Asian fruit with pungent odor and custard-like texture	A close-up image of a Mexican fruit known as pitaya or dragon fruit. The fruit has a leathery skin with a magenta color. The fruit is sliced open, revealing the blood-red pulp inside. The pulp is dotted with numerous small, shiny black seeds.		A vibrant Mexican market featuring a bright pink-red dragon fruit (pitaya) with leathery skin and eye-catching blood-red pulp, surrounded by small, shiny black seeds. The market scene is lively with traditional woven cloths and tropical ambiance.	

Figure 6: Qualitative examples with original prompt and generated image, textual reasoning by text module, LLM-rewritten prompt and generated image.

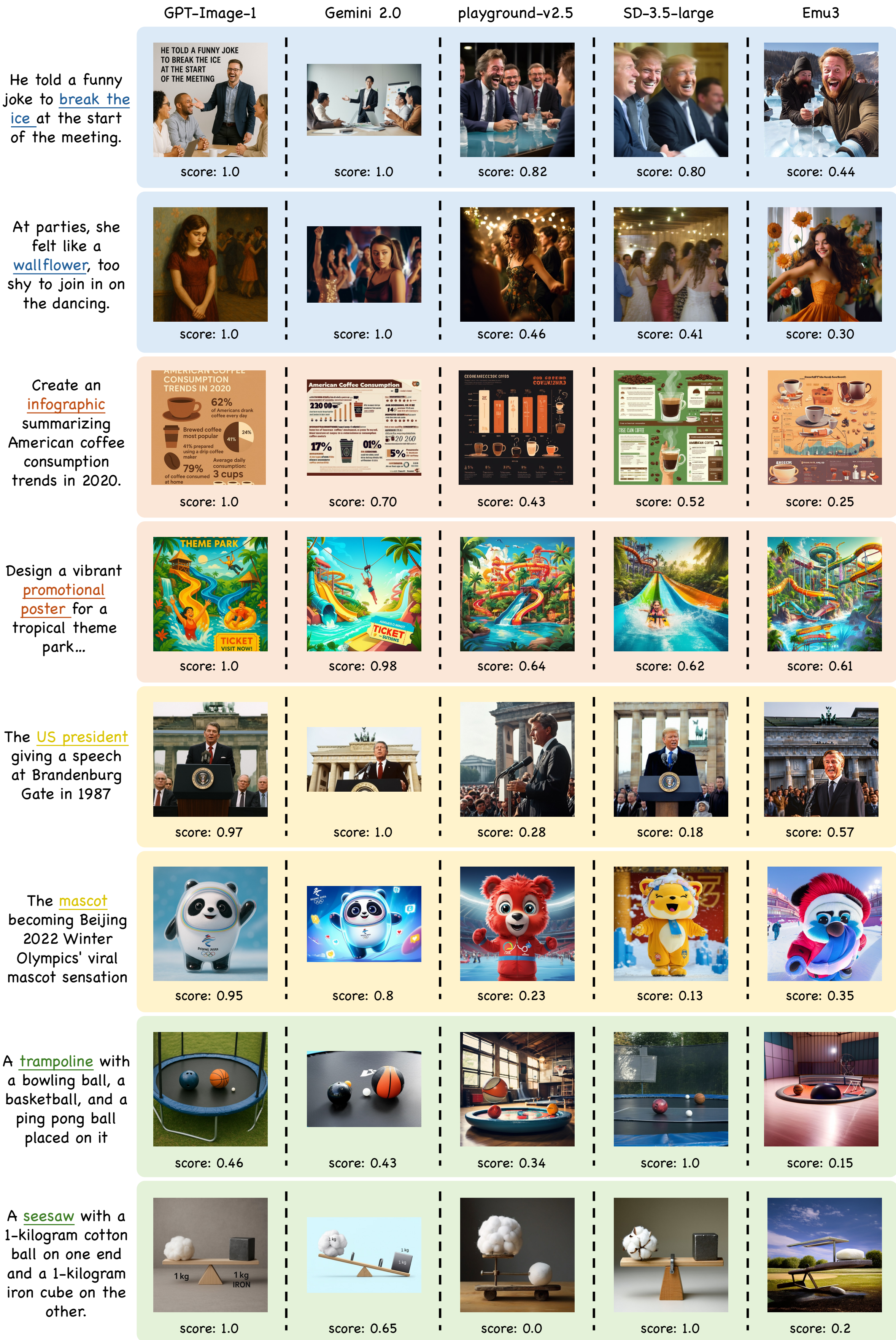


Figure 7: Qualitative examples with evaluated T2I-ReasonScore.

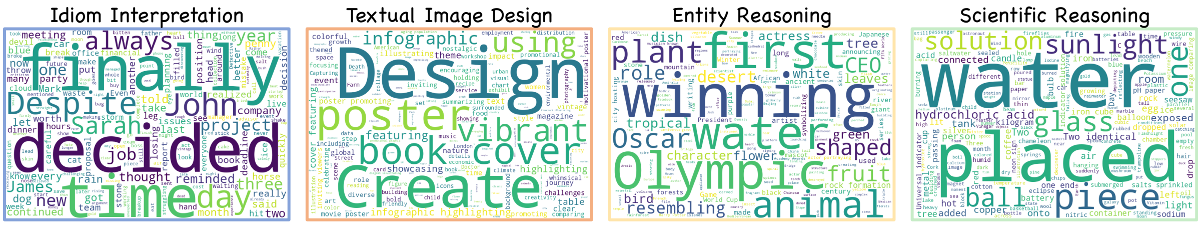


Figure 8: Word cloud to visualize the word distribution of each dimension in our prompt suite.

Template for evaluation question-criterion generation: I. Idiom Interpretation

<USER>:
I have a text-to-image generation model that can generate images based on given prompts. However, the model is not perfect and may fail to accurately capture the meaning of the prompt or depict it correctly. Your task is to evaluate the generated image based on a specific prompt that contains an idiom.
Given the prompt: {'id': {11}, 'prompt': {After three weeks of burning the midnight oil, she finally submitted her dissertation.}, 'idiom': {Burning the midnight oil}, 'idiom_meaning': {To work or study late into the night, often requiring artificial light.}}, you need to:

- 1. identify what should be depicted in the image or the meaning the image should convey.
- 2. analyze the prompt and create a list of questions based on the key elements that the image should be checked against.
- 3. consider factors that could impact the aesthetics or visual quality of the image and list relevant questions.

Please also design a scoring criterion for each question, where a score of 1 means "yes (to the question)", 0 means "no", and 0.5 means "partially yes".
Provide your answer in json format:

```
{
    'id': [prompt id],
    'prompt': [the prompt],
    'image_content': [what the image should convey],
    'reason_evaluation': (
        here should be a dictionary with 3-5 pairs of question and criterion:
        'q1': [question 1], 'c1': [criterion 1], 'q2': [question 2], 'c2': [criterion 2]...
    ),
    'quality_evaluation': (same format as 'reason_evaluation' with 1-3 pairs of question and criterion)
}
```

Figure 9: **Template used by the LLM to generate evaluation question-criterion pairs for the dimension of Idiom Interpretation.** The text highlighted in cyan should be replaced with details from the specific prompt.

Template for evaluation question-criterion generation: II. Textual Image Design

<USER>:
I have a text-to-image generation model that can generate images based on given prompts. However, the model is not perfect and may fail to accurately reflect the prompt or depict the details correctly. Given a prompt which is a design intention for a text-rich image like infographic or poster, your task is to evaluate whether the generated image correctly fulfill the design intention.
Here is the prompt: {{'id': {75}, 'prompt': {Create a minimalist promotional poster for a workshop on simplicity in design.}}, you need to:

- 1. identify what should be depicted in the image and its functional purposes.
- 2. analyze the design intention and create a list of questions based on the key elements that the image should be checked against, including presence of required text elements.
- 3. consider factors that could impact the aesthetics or visual quality of the image and list relevant questions.

Please also design a scoring criterion for each question, where a score of 1 means "yes (to the question)", 0 means "no", and 0.5 means "partially yes".
Provide your answer in json format:

```
{
    'id': [prompt id],
    'prompt': [the prompt],
    'image_content': [what the image should convey],
    'reason_evaluation': (
        here should be a dictionary with 3-5 pairs of question and criterion:
        'q1': [question 1], 'c1': [criterion 1], 'q2': [question 2], 'c2': [criterion 2]...
    ),
    'quality_evaluation': (same format as 'reason_evaluation' with 1-3 pairs of question and criterion)
}
```

Figure 10: **Template used by the LLM to generate evaluation question-criterion pairs for the dimension of Textual Image Design.** The text highlighted in cyan should be replaced with details from the specific prompt.

Template for evaluation question-criterion generation: III. Entity Reasoning

`<USER>`: I have a text-to-image generation model that can generate images based on given prompts. However, the prompts given to the model may contain implicit meanings or entities that are not directly stated. Your task is to evaluate whether the generated image accurately represents the intended meaning of the prompt.
Given the prompt: {{'id': {31}, 'prompt': {The CEO of Tesla unveiling the first electric car in 2008}, 'explicit_meaning': {Elon Musk unveiling the first electric car at Tesla in 2008}}, you need to:

- 1. identify what should be depicted in the image in order to fully and accurately reflect the explicit meaning of the prompt.
- 2. identify the entity that the model needs to infer from the prompt, and create a list of questions that check whether the image has correctly identified and depicted this entity.
- 3. Consider other elements or details in the prompt (apart from the implicit entity), create a list of questions that check if the image accurately reflects these additional key elements.
- 4. consider factors that could impact the aesthetics or visual quality of the image and list relevant questions.

Please also design a scoring criterion for each question, where a score of 1 means "yes (to the question)", 0 means "no", and 0.5 means "partially yes".
Provide your answer in json format:

```
{
    'id': [prompt id],
    'prompt': [the prompt],
    'explicit_meaning': [the explicit meaning],
    'image_content': [what the image should convey],
    'entity_evaluation': (
        here should be a dictionary with 1-3 pairs of question and criterion:
        'q1': [question 1], 'c1': [criterion 1], 'q2': [question 2], 'c2': [criterion 2]...
    ),
    'other_details_evaluation': (same format as 'entity_evaluation' with 1-3 pairs of question and criterion),
    'quality_evaluation': (same format as 'entity_evaluation' with 1-3 pairs of question and criterion)
}
```

Figure 11: **Template used by the LLM to generate evaluation question-criterion pairs for the dimension of Entity Reasoning.** The text highlighted in cyan should be replaced with details from the specific prompt.

Template for evaluation question-criterion generation: IV. Scientific Reasoning

<USER>: I have a text-to-image generation model that can generate images based on given prompts. However, the prompts given to the model imply scientific laws (e.g., physics, chemistry, biology, or astronomy) that can affect how the scene looks without explicit explanation. Your task is to evaluate whether the generated image accurately reflects the scientific law and correctly portrays the resulting scene.
Given the prompt: {{'id': {7}, 'prompt': {Two springs hanging in a row: one empty, one with a ball}, 'explicit_meaning': {Two springs hanging vertically in a row: one is empty and short, one is stretched by holding a ball}}, you need to:

- 1. describe what should be depicted in the image in order to fully and accurately reflect the explicit meaning of the prompt.
- 2. identify any scientific law(s) that the model needs to infer from the prompt, and create a list of questions that check whether the image correctly demonstrates and complies with these scientific laws.
- 3. consider other elements or details in the prompt that are not directly affected by the scientific law(s), create a list of questions that check if the image accurately represents these additional key elements.
- 4. consider factors that could impact the aesthetics or visual quality of the image and list relevant questions.

Please also design a scoring criterion for each question, where a score of 1 means "yes (to the question)", 0 means "no", and 0.5 means "partially yes".
Provide your answer in json format:

```
{
    'id': [prompt id],
    'prompt': [the prompt],
    'explicit_meaning': [the explicit meaning],
    'image_content': [what the image should depict],
    'scientific_evaluation': (
        here should be a dictionary with 2-4 pairs of question and criterion:
        'q1': [question 1], 'c1': [criterion 1], 'q2': [question 2], 'c2': [criterion 2]...
    ),
    'other_details_evaluation': (same format as 'entity_evaluation' with 1-3 pairs of question and criterion),
    'quality_evaluation': (same format as 'entity_evaluation' with 1-3 pairs of question and criterion)
}
```

Figure 12: **Template used by the LLM to generate evaluation question-criterion pairs for the dimension of Scientific Reasoning.** The text highlighted in cyan should be replaced with details from the specific prompt.

Example of evaluation question-criterion: II. Textual Image Design

<ASSISTANT>:
{ “id”: 75,
“prompt”: “Create a minimalist promotional poster for a workshop on simplicity in design.”,
“image_content”: "The image should be a clean, minimalist promotional poster advertising a workshop focused on simplicity in design. It should prominently communicate that it is a workshop (educational/learning event), clearly mention the theme/topic (simplicity in design), and ideally include essential event information such as title, subtitle/tagline, date, time, and possibly location or registration details. Visually, it should use a minimalist design style: ample white space, limited color palette, simple typography, and minimal decorative elements that reinforce the idea of simplicity......”,
“reason_evaluation”: {
“q1”: “Does the image clearly represent a promotional poster for a workshop (i.e., it looks like an event advertisement rather than a random artwork)?”,
“c1”: “1: The layout and content clearly resemble a promotional poster for an event (e.g., has a title, event-focused text, and structured composition). 0.5: Some elements suggest it might be a poster for an event, but it is ambiguous or missing key cues. 0: It does not look like a promotional poster at all.”,
“q2”: “Is it clear from the text that this is a workshop (e.g., using words like ’workshop’, ’seminar’, ’session’, or similar)?”,
“c2”: “1: The text explicitly indicates it is a workshop or equivalent learning event. 0.5: The text implies some kind of event or session but does not clearly state ’workshop’ or similar. 0: No indication that it is a workshop or event.”,
“q3”: “Does the poster clearly communicate the theme/topic of ’simplicity in design’ in the text?”,
“c3”: “1: The text specifically mentions ’simplicity in design’ or an obviously equivalent phrase (e.g., ’design simplicity’, ’simple design principles’). 0.5: The text suggests something related to design or simplicity but not clearly both together. 0: The text does not convey the theme or is unrelated.”,
“q4”: “Are key event details present, such as workshop title and at least one of: date, time, location, or call-to-action (e.g., registration info or website)?”,
“c4”: “1: The poster includes a clear workshop title and at least one important event detail (date/time/location/registration/website). 0.5: Only a vague title or partial detail is present, or details are incomplete or hard to read. 0: No usable event details are present beyond perhaps a vague title.”,
“q5”: “Is the visual style appropriately minimalist, aligning with the concept of simplicity in design (e.g., limited color palette, ample white space, minimal clutter)?”,
“c5”: “1: The design is clearly minimalist, with restrained use of color, simple shapes, and plenty of negative space. 0.5: The design is somewhat simple but includes some clutter, unnecessary decoration, or busy elements that weaken the minimalist feel. 0: The design is visually busy or complex and does not appear minimalist.” },
“quality_evaluation”: {
“q1”: “Is the typography clean and legible (fonts simple, text easy to read, no obvious text distortions or AI text artifacts)?”,
“c1”: “1: Text is clear, readable, and uses simple, appropriate fonts with minimal or no distortion. 0.5: Most text is readable but some parts are distorted, crowded, or slightly unclear. 0: Text is largely unreadable, heavily distorted, or filled with nonsensical characters.”, ... } }

Figure 13: An image evaluation example output for Textual Image Design

Template for image evaluation

<USER>: Describe this image.
<ASSISTANT>: The image depicts a group of four people in a meeting setting, with one person standing and the others seated around a table. The standing individual, dressed in a dark suit and light blue shirt, is playfully pointing at the nose of one of the seated individuals...
<USER>: Based on the image and your previous description, answer the following questions: q1, q2, ... For each question, assign a score of 1, 0.5 or 0 according to the corresponding scoring criteria: c1, c2, ...
Here are the questions and criteria: {accuracy_evaluation_qc}.
Carefully consider the image and each question before responding, then provide your answer in json format: {{"reason": [your detailed reasoning], "score": [s1,s2, ...]}}
<ASSISTANT>: { "reason": ["q1: The image clearly shows a meeting environment with a conference table, office-like setting, and individuals in professional attire. This aligns well with the criteria for a clear meeting setting.",
"q2: One person is visibly acting as the joke-teller by making a playful gesture towards another person's face, which suggests an active role in engaging the group. This meets the criteria for a clear speaker with an active role.",
"q3: The audience members are displaying relaxed and post-icebreaking reactions, such as genuine smiles and laughter, indicating that the tension has been relieved. This fulfills the criteria for clear reactions showing tension relief.",
"q4: The idiom 'break the ice' is depicted entirely metaphorically through the playful interaction and the text overlay, without any literal elements like actual ice. This satisfies the criteria for an entirely metaphorical depiction."], "score": [1, 1, 1, 1] }
<USER>: Based on the image and your previous description, answer the following questions: q1, q2, ... For each question, assign a score of 1, 0.5 or 0 according to the corresponding scoring criteria: c1, c2, ...
Here are the questions and criteria: {quality_evaluation_qc}
Carefully consider the image and each question before responding, then provide your answer in json format: {{"reason": [your detailed reasoning], "score": [s1,s2, ...]}}
<ASSISTANT>: { "reason": ["The image is visually clear with well-defined subjects. There is no noticeable blur or distortion, allowing for easy identification of the individuals and their actions. Therefore, it scores high on clarity.",
"The composition is balanced, with the speaker positioned centrally and the audience arranged cohesively around the table. There is no awkward cropping, and all elements contribute to a harmonious arrangement.",
"The colors and lighting are appropriate, enhancing the mood of the scene. Warm tones and even lighting create a positive and inviting atmosphere, supporting the lighthearted interaction depicted."], "score": [1, 1, 1] }

Figure 14: **Template used by MLLM for image evaluation.** The text highlighted in cyan should be replaced with the specific evaluation question and criterion pairs for the given prompt. An example output for Idiom Interpretation is also provided.

Human Evaluation Criteria: I. Idiom Interpretation

Given the prompt "{prompt}", idiom it contains "{idiom}" and idiom meaning "{idiom meaning}", please rate the alignment between the image and the prompt on a scale of 1 to 5 according to the criteria:

- 5 - The image independently and unambiguously depicts all elements of the prompt (actions, emotions, context, and consequences) without requiring additional text for further explanation. Annotators can instantly recognize the intended message and its nuances.
- 4 - The image clearly reflects the core idea of the prompt, capturing major elements (e.g., key actions, settings) but may lack subtle details (e.g., context, specific emotions). Annotators can easily connect it to the prompt with minimal effort.
- 3 - The image partially represents the prompt, focusing on generic aspects (e.g., basic scenario) but missing critical details (e.g., cause-effect relationships, tone, implied consequences). Annotators can only understand the link after reading the prompt and idiom meaning.
- 2 - The image vaguely or superficially relates to the prompt, with weak or unclear ties to its specifics (e.g., missing context, conflicting tone, wrong elements). Even with the prompt, the connection feels unclear or underdeveloped.
- 1 - The image contradicts or ignores the prompt's core message (e.g., misrepresenting outcomes, tone, or relationships). Annotators can find it irrelevant or misleading, even with the prompt.

Figure 15: **Human Evaluation Criterion for Idiom Interpretation.** The text highlighted in cyan should be replaced with information about the specific prompt.

Human Evaluation Criteria: II. Textual Image Design

Given a prompt describing a design intention for a rich-text image "{prompt}", please rate how well the image reflects the design prompt on a scale of 1 to 5 according to the criteria:

- 5 – Exemplary Alignment: The image perfectly reflects the design prompt, addressing all specified elements (e.g., text type, visuals, data, tone), delivers the core message clearly, and has no flaws (no errors, coherent emphasis, and non-superficial intentions fully realized).
- 4 – Good Alignment with Minor Gaps: The image aligns well with the prompt, fulfills core requirements, and conveys the message effectively but has minor oversights (e.g., missing details, slight color/text inconsistency) that do not undermine the overall intent.
- 3 – Partial Fulfillment: The image captures the general idea and addresses key aspects of the prompt (e.g., correct type, basic message) but overlooks or misrepresents notable details (e.g., incorrect text/data visualization, inconsistent tone) or contains errors affecting clarity.
- 2 – Superficial Compliance: The image only superficially resembles the prompt's intent (e.g., correct theme but missing critical elements like key visuals, misaligned focus, or unaddressed design implications) and may include distracting errors or inconsistencies.
- 1 – Mismatched or Incomplete: The image fails to address the prompt's requirements (e.g., wrong image type, missing core message, major design inaccuracies) with pervasive errors, rendering it ineffective or off-topic.

Figure 16: **Human Evaluation Criterion for Textual Image Design.** The text highlighted in cyan should be replaced with information about the specific prompt.

Human Evaluation Criteria: III. Entity Reasoning

Given the prompt "{prompt}" and the actual entity it indicates "{explicit_prompts}", please rate the alignment between the image and the prompt on a scale of 1 to 5 according to the criteria:

- 5 - Perfectly alignment: the image faithfully captures all key elements of the prompt (subject, setting, time period, distinguishing features) with no inaccuracies.
- 4 - Mostly accurate: the image depicts core elements correctly but has minor errors (e.g., slight anachronisms, missing details, or incomplete context).
- 3 - Partially correct: the image includes some relevant elements but mixes in inaccuracies (e.g., wrong context, missing critical details, or moderate deviations from the prompt).
- 2 - Weak representation: the image only loosely connected to the prompt, with significant inaccuracies (e.g., wrong subject identity, era or location).
- 1 - Completely inaccurate: the image fails to reflect the prompt's core theme, details, or context (e.g., unrelated subject, fantasy elements, or contradictory visuals).

Please carefully examine the image and check if all the details in the prompt are correctly addressed in the image.

Figure 17: **Human Evaluation Criterion for Entity Reasoning.** The text highlighted in cyan should be replaced with information about the specific prompt.

Human Evaluation Criteria: IV. Scientific Reasoning

Given a prompt that relates to scientific laws "{prompt}" and its reference prompt "{explicit_prompt}", please rate the image on a scale of 1 to 5 according to the criteria:

- 5 - Excellent: The image accurately depicts all the elements from the prompt (subject, action, setting, state) and strictly adheres to scientific laws. No errors in details or logic.
- 4 - Good: The image includes all key elements from the prompt but has minor scientific inaccuracies or small missing details.
- 3 - Fair: The image includes most elements but has moderate errors: either missing an critical element or clearly violating scientific principles.
- 2 - Poor: The image omits multiple key elements and has significant scientific inaccuracies.
- 1 - Fail: The image fails to represent the prompt (e.g., incorrect subjects/actions) and completely ignores scientific laws.

Please carefully examine the image and check if the image correctly address the scientific law inherent in the prompt.

Figure 18: **Human Evaluation Criterion for Scientific Reasoning.** The text highlighted in cyan should be replaced with information about the specific prompt.