

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

Anonymous ACL submission

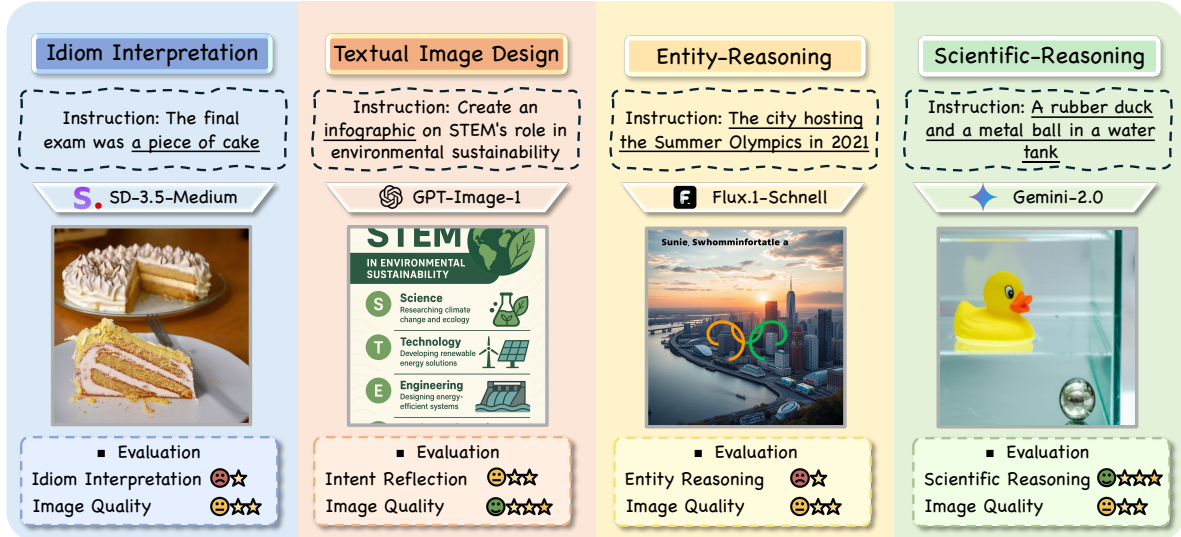


Figure 1: **Overview of T2I-ReasonBench.** We propose T2I-ReasonBench, a benchmark evaluating reasoning-informed generation of text-to-image (T2I) models. It consists of four dimensions: **Idiom Interpretation**, **Textual Image Design**, **Entity-Reasoning** and **Scientific-Reasoning**. We propose 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.

## 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), such as T2I-CompBench (Huang et al., 2025) and PartiPrompts (Yu et al., 2022), 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-ReasonBench 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, in-

cluding 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 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 lan-

151 guage model with a diffusion module to integrate  
152 understanding and generation. For instance, GoT  
153 (Fang et al., 2025) uses an MLLM for semantic-  
154 spatial reasoning before diffusion-based synthesis,  
155 and Bagel (Deng et al., 2025) unifies an LLM and  
156 a diffusion model within a single transformer to  
157 generate reasoning chains prior to image creation.

## 158 2.2 Text-to-image Benchmarks and 159 Evaluation Metrics.

160 **Benchmarks.** Current T2I benchmarks (Yu et al.,  
161 2022; Hu et al., 2023a; Huang et al., 2023; Ghosh  
162 et al., 2023; Hu et al., 2024; Li et al., 2024a; Wu  
163 et al., 2024; Huang et al., 2025; Wei et al., 2025)  
164 primarily evaluate literal prompt-image alignment.  
165 For example, GenEval (Ghosh et al., 2023) utilizes  
166 object detection techniques to test whether gener-  
167 ated images correctly capture object co-occurrence,  
168 position, count, and color described in the prompts.  
169 Recent benchmarks have shifted focus from literal  
170 alignment to reasoning capabilities of T2I models.  
171 For instance, Commonsense-T2I (Fu et al., 2024)  
172 tests everyday logic through adversarial prompt  
173 pairs; PhyBench (Meng et al., 2024) evaluates phys-  
174 ical common sense; WISE (Niu et al., 2025) as-  
175 sesses broader world knowledge; and R2I-Bench  
176 (Chen et al., 2025c) includes both composition and  
177 reasoning categories. Our benchmark encompasses  
178 a comprehensive set of reasoning dimensions and  
179 provides a deeper, more insightful analysis of mod-  
180 els’ reasoning and generation abilities.

181 **Evaluation Metrics.** Conventional text-image  
182 alignment metrics like CLIPscore (Hessel et al.,  
183 2021) and VQAscore (Lin et al., 2024) work as  
184 bag-of-words models, lacking the expertise needed  
185 to evaluate specific composition and reasoning gen-  
186 erations. To address this, many works adopt a more  
187 targeted, disentangled question-answering frame-  
188 work. Leveraging powerful LLMs and MLLMs,  
189 this method first generates specific diagnostic ques-  
190 tions and then uses VQA models to answer them  
191 by inspecting the image. This approach has been  
192 successfully applied across studies evaluating text-  
193 image alignment (Hu et al., 2023b; Cho et al.,  
194 2023a; Yarom et al., 2023; Cho et al., 2023b), com-  
195 position (Wu et al., 2024), reasoning (Chen et al.,  
196 2025c), and factual correctness (Lim et al., 2025).  
197 Given the complexity of our benchmark, this tar-  
198 geted, disentangled approach is more reliable than  
199 metrics with identical instructions.

## 3 Benchmark Construction 200

201 While modern T2I models are good at explicit  
202 prompt-to-image translation, their capacity for  
203 reasoning-informed generation remains underex-  
204 plored. Existing benchmarks focus predominantly  
205 on literal text-image alignment (e.g., object exist-  
206 ence, spatial arrangements) but fail to evaluate  
207 whether models possess reasoning abilities to un-  
208 cover the deeper meaning behind the text and gen-  
209 erate logically coherent visual content. To this  
210 end, we identify four dimensions that challenge  
211 T2I models to reason about the instructions with  
212 domain knowledge before visualizing them:

213 **Idiom Interpretation.** An idiom is a phrase or  
214 combination of words with a figurative meaning  
215 that differs from its literal meaning. Idioms are  
216 common in everyday language, and their meanings  
217 usually cannot be deduced by analyzing individ-  
218 ual words. For T2I models, prompts containing  
219 idioms demand reasoning to obtain the latent mean-  
220 ing before generating correct visual content. This  
221 process requires leveraging linguistic knowledge  
222 and effectively analyzing context.

223 By sourcing from a book (idi, 2023) and the in-  
224 ternet, we collect 200 idioms that are commonly  
225 used in daily life but challenging for T2I models.  
226 We then use an LLM to generate sentences con-  
227 taining the idioms but without explicitly revealing  
228 their meanings. These idioms span diverse topics  
229 such as social interactions, lifestyle, and emotions.  
230 For example, the sentence “He told a funny joke to  
231 break the ice at the start of the meeting” uses the  
232 idiom “break the ice”, which means to ease tension,  
233 rather than literally destroying the ice.

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

242 In this dimension, we first collected 200 real-world  
243 images featuring rich text from different datasets.  
244 Using an MLLM, we then extracted the underlying  
245 design intentions from these images, resulting in  
246 200 design prompts. Each focuses on the functional  
247 purpose of the image rather than describing visual  
248 details. For example, “Design an infographic on  
249 online risks for children aged 9-16”. Based on  
250 the image sources, the prompts span categories

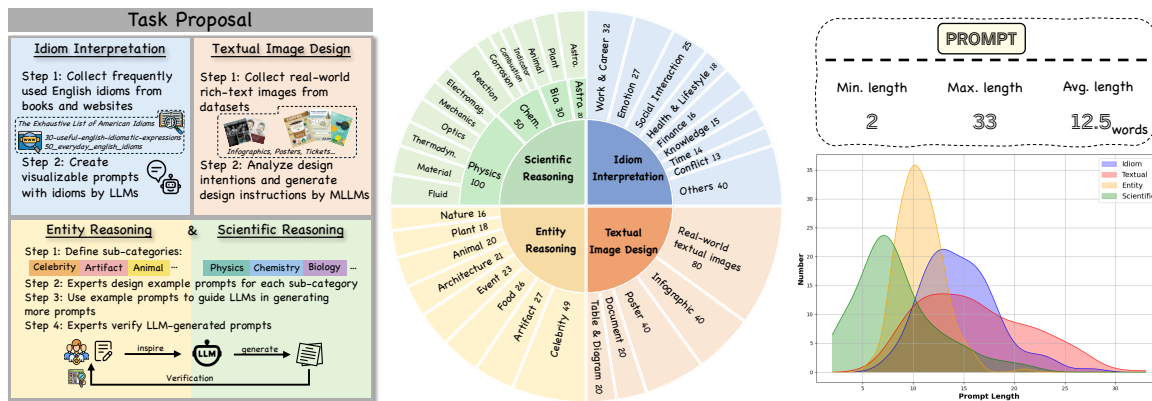


Figure 2: **Left: Prompt Generation Pipeline.** Idiom Interpretation: collect idioms from established sources and use LLMs to create prompts. Textual Image Design: source real-world rich-text images, use MLLMs to analyze design intentions and generate corresponding instructions. Entity & Scientific Reasoning: 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: The prompt suite features an average length of 12.5 words. Bottom: The distribution of prompt lengths across the four dimensions.

like infographics, posters, documents, tables, 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. **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

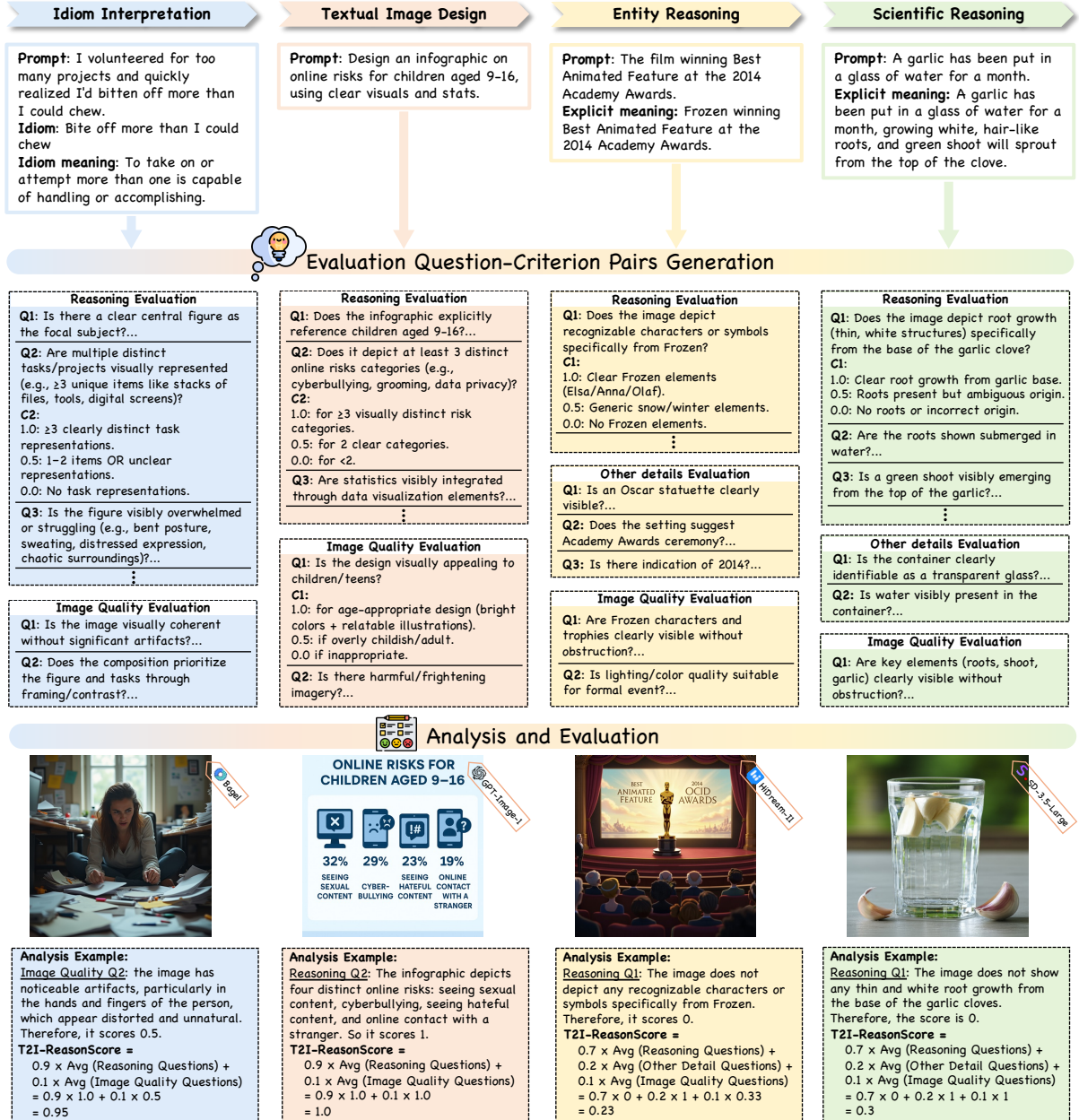


Figure 3: **Evaluation Framework of T2I-ReasonBench.** We adopt a two-stage evaluation framework: prompt-specific evaluation question-criterion pairs generation by an LLM, then image analysis and scoring by an MLLM. This figure shows one evaluation example for each dimension.

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:

$$S_j = \frac{\sum_{i=1}^{n_j} score_i}{n_j}, \quad (1)$$

where  $S_j$  is the score for the  $j^{th}$  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:

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

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  $\tau$  and Spearman’s  $\rho$ .

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	
		$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$
<i>T2I-ReasonScore</i>											
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	0.5285	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	0.5491	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	0.6728	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
<i>Other metrics</i>											
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	0.3098	0.4939	0.4291	0.5834

The average correlation is the highest when pairing Gemini-2.5-pro with questions from GPT-5.1. However, for Entity Reasoning, Gemini-2.5-pro with questions generated by itself yields higher correlation. Therefore, we select Gemini-2.5-pro as our 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 align with human perception.

## 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 settings 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

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
<b>Diffusion Models</b>					
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-II-full (hidream, 2024)	59.1	<b>80.3</b>	58.0	59.4	64.2
Qwen-Image (Wu et al., 2025)	<b>62.7</b>	77.1	<b>61.8</b>	<b>68.2</b>	<b>67.5</b>
<b>Unified Multimodal Models</b>					
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)	<b>50.6</b>	<b>50.5</b>	<b>56.5</b>	<b>64.9</b>	<b>55.6</b>
<b>Proprietary Models</b>					
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)	<b>89.8</b>	<b>94.8</b>	<b>87.3</b>	<b>86.0</b>	<b>89.5</b>

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
<b>Diffusion Models</b>					
SD-3-Medium	76.7 <sub>↑34.9</sub>	74.1 <sub>↑7.2</sub>	75.0 <sub>↑32.0</sub>	68.5 <sub>↑11.9</sub>	73.6 <sub>↑21.5</sub>
SD-3.5-Medium	76.4 <sub>↑36.8</sub>	73.1 <sub>↑7.1</sub>	76.4 <sub>↑28.8</sub>	69.5 <sub>↑13.9</sub>	73.8 <sub>↑21.7</sub>
SD-3.5-Large	76.7 <sub>↑32.5</sub>	74.2 <sub>↑7.7</sub>	82.1 <sub>↑31.1</sub>	71.4 <sub>↑12.1</sub>	76.1 <sub>↑20.9</sub>
FLUX.1-dev	77.8 <sub>↑28.5</sub>	78.0 <sub>↑10.1</sub>	75.6 <sub>↑25.1</sub>	71.6 <sub>↑19.9</sub>	75.8 <sub>↑20.9</sub>
FLUX.1-schnell	79.9 <sub>↑32.5</sub>	78.2 <sub>↑4.0</sub>	76.0 <sub>↑26.5</sub>	74.4 <sub>↑15.7</sub>	77.3 <sub>↑19.7</sub>
Playground-v2.5	65.3 <sub>↑14.5</sub>	48.9 <sub>↑3.5</sub>	77.4 <sub>↑26.4</sub>	57.3 <sub>↑2.8</sub>	62.2 <sub>↑11.8</sub>
HiDream-II-full	77.0 <sub>↑17.9</sub>	81.8 <sub>↑1.4</sub>	79.6 <sub>↑21.6</sub>	71.6 <sub>↑12.1</sub>	77.5 <sub>↑13.2</sub>
Qwen-Image	<b>84.7</b> <sub>↑22.0</sub>	83.0 <sub>↑5.9</sub>	<b>84.2</b> <sub>↑22.4</sub>	<b>83.7</b> <sub>↑15.5</sub>	<b>83.9</b> <sub>↑16.4</sub>
<b>Unified Multimodal Models</b>					
Emu3	67.7 <sub>↑28.5</sub>	48.8 <sub>↑8.5</sub>	67.6 <sub>↑24.9</sub>	57.3 <sub>↑12.8</sub>	60.3 <sub>↑18.7</sub>
show-o-demo-512	74.8 <sub>↑35.8</sub>	48.6 <sub>↑5.4</sub>	69.4 <sub>↑28.4</sub>	65.2 <sub>↑16.9</sub>	64.5 <sub>↑21.6</sub>
Janus-Pro-7B	72.6 <sub>↑39.9</sub>	57.1 <sub>↑14.2</sub>	71.0 <sub>↑29.7</sub>	68.6 <sub>↑16.2</sub>	67.3 <sub>↑25.0</sub>
GoT	62.3 <sub>↑28.7</sub>	45.9 <sub>↑5.2</sub>	55.2 <sub>↑17.6</sub>	50.2 <sub>↑8.9</sub>	53.4 <sub>↑15.1</sub>
Bagel w/o Thinking	<b>77.9</b> <sub>↑27.4</sub>	<b>65.4</b> <sub>↑14.9</sub>	<b>73.9</b> <sub>↑17.4</sub>	<b>74.0</b> <sub>↑9.1</sub>	<b>72.8</b> <sub>↑17.2</sub>
<b>Proprietary Models</b>					
Gemini-2.0	80.3 <sub>↑14.8</sub>	80.8 <sub>↑2.9</sub>	86.6 <sub>↑8.6</sub>	82.5 <sub>↑6.3</sub>	82.5 <sub>↑8.1</sub>
GPT-Image-1	87.1 <sub>↑3.1</sub>	89.2 <sub>↑3.2</sub>	<b>90.8</b> <sub>↑5.0</sub>	<b>87.3</b> <sub>↑4.4</sub>	88.6 <sub>↑2.3</sub>
Nano-Banana	<b>87.7</b> <sub>↑2.2</sub>	<b>92.2</b> <sub>↑2.6</sub>	90.2 <sub>↑2.9</sub>	86.5 <sub>↑0.6</sub>	<b>89.2</b> <sub>↑0.3</sub>

maintaining a balanced final score.

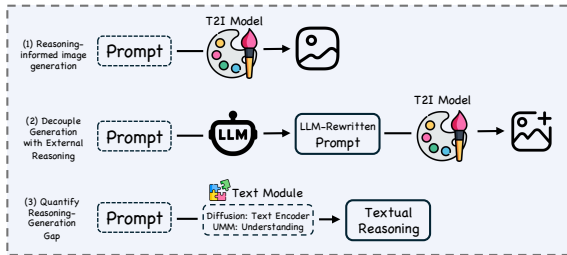


Figure 4: **Three Experiment Pipeline in Section 5.**

## 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 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 involves 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**

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
<i>Diffusion Models</i>				
HiDream-11-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
<i>Unified Multimodal Models</i>				
Emu3-Gen/Emu3-Und	<b>31.3/36.9</b>	<b>34.9/80.3</b>	<b>34.9/56.9</b>	<b>37.7/63.2</b>
Bage-Gen/Bage1-Und	41.7/52.0	43.2/84.4	47.8/67.6	56.0/76.4
<i>Proprietary Models</i>				
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

**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 demonstrate significant challenges in both areas. As shown in Figure 6 (Row #4), its approach to reasoning is superficial; rather than analyzing the scientific implications, its textual output merely extends the original prompt with more detail. This reasoning deficit is compounded by poor generation quality, as the model fails to compose all specified objects and struggles to follow the prompts. More advanced UMM like Bage1 exhibit 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 correctly 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.

**Other failure cases.** The last two rows in Figure 6 exhibit two additional failure cases by proprietary models. The textual reasoning in Row #5 generated by Nano-Banana fails to uncover the physical implications, so does the image. In Row #6, although the textual reasoning by GPT-4o successfully recognize the entity, the image fails to follow it but just shows an object that superficially align with the prompt, demonstrating a failure in reasoning integration.

**Implication for future model design.** The fact that an integrated model like Nano-Banana (Google, 2025) outperforms the two-stage pipeline suggests that models with inherent, built-in reasoning capacity are superior to a decoupled approach and represent a promising future trend for development.

557  
558  
559  
  
560  
561  
562  
563  
564  
565  
  
566  
567  
568  
569  
570  
  
571  
572  
573  
574  
575  
576  
577  
  
578  
579  
580  
581  
582  
  
583  
584  
585  
586  
587  
588  
  
589  
590  
591  
592  
593  
  
594  
595  
596  
597  
598  
  
599  
600  
601  
602  
603  
604  
  
605  
606  
607  
608  
  
609  
610  
611

## 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-v1 technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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-v1 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, Shiron Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

668	Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. <i>arXiv preprint arXiv:2104.08718</i> .	723
669		724
670		
671		
672	hidream. 2024. hidream. <a href="https://github.com/HiDream-ai/HiDream-I1">https://github.com/HiDream-ai/HiDream-I1</a> .	
673		
674	Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. <i>Advances in neural information processing systems</i> , 33:6840–6851.	
675		
676		
677		
678	Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. <i>arXiv preprint arXiv:2403.05135</i> .	
679		
680		
681		
682	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 <i>Proceedings of the IEEE/CVF International Conference on Computer Vision</i> , pages 20406–20417.	
683		
684		
685		
686		
687		
688		
689	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 <i>Proceedings of the IEEE/CVF International Conference on Computer Vision</i> , pages 20406–20417.	
690		
691		
692		
693		
694		
695		
696	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. <i>IEEE Transactions on Pattern Analysis &amp; Machine Intelligence</i> , 47(05):3563–3579.	
697		
698		
699		
700		
701		
702	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. <i>Advances in Neural Information Processing Systems</i> , 36:78723–78747.	
703		
704		
705		
706		
707	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. <i>arXiv preprint arXiv:2410.21276</i> .	
708		
709		
710		
711		
712	Black Forest Labs. 2024. Flux. <a href="https://github.com/black-forest-labs/flux">https://github.com/black-forest-labs/flux</a> .	
713		
714	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. <i>arXiv preprint arXiv:2406.13743</i> .	
715		
716		
717		
718		
719		
720	Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. 2024b. Playground v2. 5: Three insights towards enhancing aesthetic	
721		
722		
	quality in text-to-image generation. <i>arXiv preprint arXiv:2402.17245</i> .	725
		726
		727
		728
		729
	Youngsun Lim, Hojun Choi, and Hyunjung Shim. 2025. Evaluating image hallucination in text-to-image generation with question-answering. In <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , 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 <i>European Conference on Computer Vision</i> , pages 366–384. Springer.	730
		731
		732
		733
		734
	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 pre-training. <i>arXiv preprint arXiv:2408.02657</i> .	735
		736
		737
		738
		739
	Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In <i>Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision</i> , pages 1697–1706.	740
		741
		742
		743
		744
	Fanqing Meng, Wenqi Shao, Lixin Luo, Yahong Wang, Yiran Chen, Quanfeng Lu, Yue Yang, Tianshuo Yang, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. Phylabench: A physical commonsense benchmark for evaluating text-to-image models. <i>arXiv preprint arXiv:2406.11802</i> .	745
		746
		747
		748
		749
		750
	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. <i>arXiv preprint arXiv:2112.10741</i> .	751
		752
		753
		754
		755
		756
	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. <i>arXiv preprint arXiv:2503.07265</i> .	757
		758
		759
		760
		761
	OpenAI. 2023. Gpt-image-1. <a href="https://openai.com/index/image-generation-api/">https://openai.com/index/image-generation-api/</a> .	762
		763
	OpenAI. 2025. Gpt-5.1: A smarter, more conversational chatgpt. <a href="https://openai.com/index/gpt-5-1/">https://openai.com/index/gpt-5-1/</a> .	764
		765
	Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. <i>arXiv preprint arXiv:2102.12092</i> .	766
		767
		768
		769
	Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i> , pages 10684–10695.	770
		771
		772
		773
		774
		775

776	Chitwan Saharia, William Chan, Saurabh Saxena,	Yue Yang, Ajay Patel, Matt Deitke, Tanmay Gupta, Luca	833
777	Lala Li, Jay Whang, Emily L Denton, Kam-	Weihs, Andrew Head, Mark Yatskar, Chris Callison-	834
778	yar Ghasemipour, Raphael Gontijo Lopes, Burcu	Burch, Ranjay Krishna, Aniruddha Kembhavi, and 1	835
779	Karagol Ayan, Tim Salimans, and 1 others. 2022.	others. 2025. Scaling text-rich image understanding	836
780	Photorealistic text-to-image diffusion models with	via code-guided synthetic multimodal data genera-	837
781	deep language understanding. <i>Advances in Neural</i>	tion. <i>arXiv preprint arXiv:2502.14846</i> .	838
782	<i>Information Processing Systems</i> , 35:36479–36494.		
783	Christoph Schuhmann, Romain Beaumont, Richard	Michal Yarom, Yonatan Bitton, Soravit Changpinyo,	839
784	Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti,	Roe Aharoni, Jonathan Herzig, Oran Lang, Eran	840
785	Theo Coombes, Aarush Katta, Clayton Mullis,	Ofek, and Idan Szpektor. 2023. What you see is what	841
786	Mitchell Wortsman, and 1 others. 2022. Laion-5b:	you read? improving text-image alignment evalua-	842
787	An open large-scale dataset for training next genera-	tion. <i>Advances in Neural Information Processing</i>	843
788	tion image-text models. <i>Advances in neural informa-</i>	<i>Systems</i> , 36:1601–1619.	844
789	<i>tion processing systems</i> , 35:25278–25294.		
790	Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang,	Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Lu-	845
791	Bingyue Peng, Ping Luo, and Zehuan Yuan.	ong, Gunjan Baid, Zirui Wang, Vijay Vasudevan,	846
792	2024. Autoregressive model beats diffusion: Llama	Alexander Ku, Yinfei Yang, Burcu Karagol Ayan,	847
793	for scalable image generation. <i>arXiv preprint</i>	and 1 others. 2022. Scaling autoregressive models for	848
794	<i>arXiv:2406.06525</i> .	content-rich text-to-image generation. <i>arXiv preprint</i>	849
795	Chameleon Team. 2024. Chameleon: Mixed-modal	<i>arXiv:2206.10789</i> , 2(3):5.	850
796	early-fusion foundation models. <i>arXiv preprint</i>	Shijie Zhou, Ruiyi Zhang, Yufan Zhou, and Changyou	851
797	<i>arXiv:2405.09818</i> .	Chen. 2024. A high-quality text-rich image instruc-	852
798	Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-	tion tuning dataset via hybrid instruction generation.	853
799	Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan	<i>arXiv preprint arXiv:2412.16364</i> .	854
800	Schalkwyk, Andrew M Dai, Anja Hauth, Katie Mil-		
801	lican, and 1 others. 2023. Gemini: a family of		
802	highly capable multimodal models. <i>arXiv preprint</i>		
803	<i>arXiv:2312.11805</i> .		
804	Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo,		
805	Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang,		
806	Yueze Wang, Zhen Li, Qiying Yu, and 1 others. 2024.		
807	Emu3: Next-token prediction is all you need. <i>arXiv</i>		
808	<i>preprint arXiv:2409.18869</i> .		
809	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten		
810	Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,		
811	and 1 others. 2022. Chain-of-thought prompting elic-		
812	its reasoning in large language models. In <i>NeurIPS</i> .		
813	Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei,		
814	Zhen Guo, and Lei Zhang. 2025. Tiif-bench: How		
815	does your t2i model follow your instructions? <i>arXiv</i>		
816	<i>preprint arXiv:2506.02161</i> .		
817	Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin,		
818	Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai		
819	Bai, Xiao Xu, Yilei Chen, and 1 others. 2025.		
820	Qwen-image technical report. <i>arXiv preprint</i>		
821	<i>arXiv:2508.02324</i> .		
822	Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Rus-		
823	sakovsky, and Sanjeev Arora. 2024. Conceptmix:		
824	A compositional image generation benchmark with		
825	controllable difficulty. <i>Advances in Neural Informa-</i>		
826	<i>tion Processing Systems</i> , 37:86004–86047.		
827	Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao		
828	Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao		
829	Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng		
830	Shou. 2024. Show-o: One single transformer to unify		
831	multimodal understanding and generation. <i>arXiv</i>		
832	<i>preprint arXiv:2408.12528</i> .		

855	<b>Appendix Contents</b>	
856	<b>A Qualitative Examples</b>	<b>12</b>
857	<b>B More Details on Prompt Collection Process</b>	<b>12</b>
858		
859	<b>C Evaluation Framework</b>	<b>12</b>
860	C.1 Templates . . . . .	12
861	C.2 Human Evaluation . . . . .	13
862	<b>D Limitation</b>	<b>13</b>
863	<b>A Qualitative Examples</b>	
864	Figure 5, 6 compares the qualitative outputs of	
865	different T2I models. Figure 7 shows more image	
866	outputs from different models with the evaluated	
867	T2I-ReasonScore.	
868	<b>B More Details on Prompt Collection</b>	
869	<b>Process</b>	
870	<b>Idiom Interpretation.</b> In idiom collection, we	
871	leverage a book titled “The Exhaustive List of	
872	American Idioms” (idi, 2023), which systemati-	
873	cally documents over 11k idioms. These idioms	
874	were collected from diverse sources, including TV	
875	shows, movies, and everyday conversations. Each	
876	idiom in the book is accompanied by its actual	
877	meaning in context. In addition to this resource,	
878	we also refer to idioms available on the Internet.	
879	From this extensive pool, we manually select 200	
880	idioms that are commonly used in daily life and	
881	challenging for T2I models due to their figurative	
882	meanings. We input the selected idioms and their	
883	actual meanings into an LLM and prompt it to gener-	
884	ate new sentences. These sentences are designed	
885	to describe visible scenes involving the idioms, pro-	
886	viding contextual clues for reasoning while avoid-	
887	ing directly revealing the idiom’s meaning.	
888	<b>Textual Image Design.</b> For textual image design,	
889	we collect 6 types of text-rich images from 4 dis-	
890	tinct sources.	
891	(1) LLAVar-2 Dataset (Zhou et al., 2024): This	
892	dataset contains 42k text-rich images sourced from	
893	LAION (Schuhmann et al., 2022), representing	
894	various categories such as quotes, memes, book	
895	covers, posters, and product packaging. However,	
896	images in this dataset are of various quality and	
897	formats, so we filter out 80 high-quality images	
898	that have a resolution greater than 384x384 and	
899	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.	900
(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.	901
(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.	902
The design intentions of the collected images are summarized using Qwen2.5-VL (Bai et al., 2025b), yielding 200 prompts in this dimension.	903
<b>Entity Reasoning and Scientific Reasoning.</b> 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.	904
We visualize the word distribution of each dimension in our prompt suite in Figure 8.	905
<b>C Evaluation Framework</b>	906
<b>C.1 Templates</b>	907
<b>Templates for evaluation stage I: LLM question-criterion generation.</b> 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	908
	909
	910
	911
	912
	913
	914
	915
	916
	917
	918
	919
	920
	921
	922
	923
	924
	925
	926
	927
	928
	929
	930
	931
	932
	933
	934
	935
	936
	937
	938
	939
	940
	941
	942
	943
	944
	945
	946
	947
	948
	949

950 and Scientific-Reasoning, respectively. Each tem- 997  
951 plate is tailored to focus on the unique aspects of 998  
952 its corresponding dimension. For example, when 999  
953 generating the question-criterion pairs for Idiom In- 1000  
954 terpretation, the LLM receives the prompt together 1001  
955 with the idiom included, and its actual meaning to 1002  
956 ensure it has the full background information to 1003  
957 generate accurate questions and criteria. For Entity 1004  
958 Reasoning, the explicit prompt with the specific 1005  
959 entity name is provided. For Scientific Reason- 1006  
960 ing, the explicit prompt specifying the visible out- 1007  
961 come derived from the scientific law is provided.  
962 As for Textual Image Design, rather than prescrib-  
963 ing visual content, the core of evaluation is to as-  
964 sess whether generated images fulfill their intended  
965 communicative function. Therefore, only the origi-  
966 nal prompt is provided, this allows the T2I models  
967 to have full design freedom. The generated evalua-  
968 tion questions and criteria for the prompt “Create a  
969 minimalist promotional poster for a workshop on  
970 simplicity in design” are shown in Table 13, we can  
971 see they focus on evaluating functional elements  
972 including:

- 973 • Clear presentation of workshop topic
- 974 • Explicit identification as an event
- 975 • Inclusion of essential details (date/time/locat- 1004  
976 tion) 1005
- 977 • Consistency with minimalist design principles 1006  
1007

978 **Templates for evaluation stage II: MLLM image**  
979 **evaluation.** Table 14 presents the template used to  
980 evaluate the generated images for all four dimen-  
981 sions. Only the evaluation question-criterion pairs  
982 need to be replaced for each prompt.

## 983 C.2 Human Evaluation

984 The human evaluation is conducted on eight mod-  
985 els: Stable-Diffusion-3-Medium, FLUX.1-schnell,  
986 HiDream-I1-full, Qwen-Image, Bagel, Janus-Pro-  
987 7B, GPT-Image-1, and Nano-Banana (4 diffusion  
988 models, 2 unified multimodal models, 2 propri-  
989 etary model). We randomly select 20 prompts from  
990 each dimension and use these 8 models to generate  
991 160 images per dimension. This results in 640 im-  
992 ages in total for human evaluation. The evaluation  
993 is conducted with a group of college postgradu-  
994 ate participants. Three participants independently  
995 score each image, and we average their scores for  
996 each prompt-image pair. The evaluation criteria

provided to the human annotators are shown in Ta-  
ble 15, 16, 17, and 18. These criteria are tailored  
to each dimension in order to address different as-  
pects. The correlations are calculated between the  
averaged human scores and the automatic metric  
score using Kendall’s  $\tau$  and Spearman’s  $\rho$ . The  
results are shown in Table 1.

## D Limitation

One limitation of this work is not including more  
reasoning dimensions. The other limitation is not  
examining textual scores for more models.

#1	<p><b>Prompt</b> from Textual Image Design</p>	<p><b>Textual Reasoning</b> by Nano-Banana</p>	<p><b>Image generated by Nano-Banana</b> with original prompt</p>	<p><b>LLM-Rewritten Prompt</b></p>	<p><b>Image generated by Nano-Banana</b> with LLM-rewritten prompt</p>
	<p>Create an infographic highlighting the challenges faced by caregivers for the aging population.</p>	<p>...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...</p>		<p>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.</p>	
#2	<p><b>Prompt</b> from Entity Reasoning</p>	<p><b>Textual Reasoning</b> by Nano-Banana</p>	<p><b>Image generated by Nano-Banana</b> with original prompt</p>	<p><b>LLM-Rewritten Prompt</b></p>	<p><b>Image generated by Nano-Banana</b> with LLM-rewritten prompt</p>
	<p>The Southeast Asian fruit with pungent odor and custard-like texture</p>	<p>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.</p>		<p>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.</p>	
#3	<p><b>Prompt</b> from Scientific Reasoning</p>	<p><b>Textual Reasoning</b> by Nano-Banana</p>	<p><b>Image generated by Nano-Banana</b> with original prompt</p>	<p><b>LLM-Rewritten Prompt</b></p>	<p><b>Image generated by Nano-Banana</b> with LLM-rewritten prompt</p>
	<p>A heavy rock placed on a beach ball</p>	<p>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.</p>		<p>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.</p>	
#4	<p><b>Prompt</b> from Idiom Interpretation</p>	<p><b>Textual Reasoning</b> by GPT-4o</p>	<p><b>Image generated by GPT-Image-1</b> with original prompt</p>	<p><b>LLM-Rewritten Prompt</b></p>	<p><b>Image generated by GPT-Image-1</b> with LLM-rewritten prompt</p>
	<p>She hadn't played the piano in years, but it was like riding a bicycle; she picked it up instantly.</p>	<p>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.</p>		<p>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...</p>	
#5	<p><b>Prompt</b> from Entity Reasoning</p>	<p><b>Textual Reasoning</b> by GPT-4o</p>	<p><b>Image generated by GPT-Image-1</b> with original prompt</p>	<p><b>LLM-Rewritten Prompt</b></p>	<p><b>Image generated by GPT-Image-1</b> with LLM-rewritten prompt</p>
	<p>The ancient stone circle in Wiltshire, England</p>	<p>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...</p>		<p>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.</p>	
#6	<p><b>Prompt</b> from Scientific Reasoning</p>	<p><b>Textual Reasoning</b> by GPT-4o</p>	<p><b>Image generated by GPT-Image-1</b> with original prompt</p>	<p><b>LLM-Rewritten Prompt</b></p>	<p><b>Image generated by GPT-Image-1</b> with LLM-rewritten prompt</p>
	<p>A seesaw with a 1 cubic decimeter block of cotton on one end and 1 cubic decimeter iron cube on the other</p>	<p>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.</p>		<p>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.</p>	

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

#1	<b>Prompt</b> from Scientific Reasoning	<b>Textual Reasoning</b> by Qwen2.5-VL-7b	<b>Image generated by Qwen-Image</b> with original prompt	<b>LLM-Rewritten Prompt</b>	<b>Image generated by Qwen-Image</b> 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	<b>Prompt</b> from Entity Reasoning	<b>Textual Reasoning</b> by Llama-3.1-8b	<b>Image generated by Hidream</b> with original prompt	<b>LLM-Rewritten Prompt</b>	<b>Image generated by Hidream</b> 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	<b>Prompt</b> from Entity Reasoning	<b>Textual Reasoning</b> by Bagel-Und	<b>Image generated by Bagel</b> with original prompt	<b>LLM-Rewritten Prompt</b>	<b>Image generated by Bagel</b> 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	<b>Prompt</b> from Scientific Reasoning	<b>Textual Reasoning</b> by Emu3-Und	<b>Image generated by Emu3</b> with original prompt	<b>LLM-Rewritten Prompt</b>	<b>Image generated by Emu3</b> 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	<b>Prompt</b> from Scientific Reasoning	<b>Textual Reasoning</b> by Nano-Banana	<b>Image generated by Nano-Banana</b> with original prompt	<b>LLM-Rewritten Prompt</b>	<b>Image generated by Nano-Banana</b> 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	<b>Prompt</b> from Entity Reasoning	<b>Textual Reasoning</b> by GPT-4o	<b>Image generated by GPT-Image-1</b> with original prompt	<b>LLM-Rewritten Prompt</b>	<b>Image generated by GPT-Image-1</b> 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.






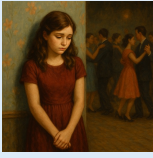
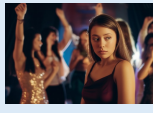

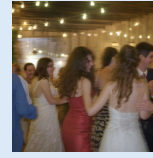
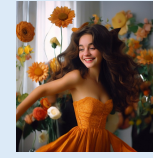
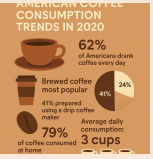





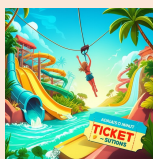

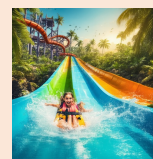
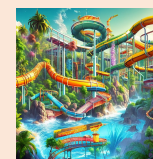



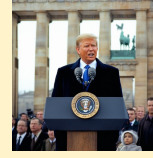


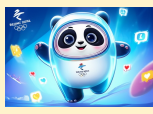


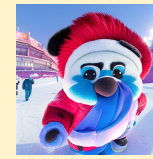



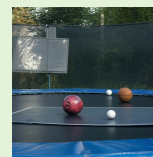



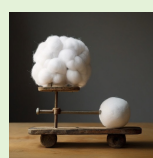
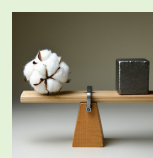

	GPT-Image-1	Gemini 2.0	playground-v2.5	SD-3.5-large	Emu3
He told a funny joke to <u>break the ice</u> at the start of the meeting.	 score: 1.0	 score: 1.0	 score: 0.82	 score: 0.80	 score: 0.44
At parties, she felt like a <u>wallflower</u> , too shy to join in on the dancing.	 score: 1.0	 score: 1.0	 score: 0.46	 score: 0.41	 score: 0.30
Create an <u>infographic</u> summarizing American coffee consumption trends in 2020.	 score: 1.0	 score: 0.70	 score: 0.43	 score: 0.52	 score: 0.25
Design a vibrant <u>promotional poster</u> for a tropical theme park...	 score: 1.0	 score: 0.98	 score: 0.64	 score: 0.62	 score: 0.61
The <u>US president</u> giving a speech at Brandenburg Gate in 1987	 score: 0.97	 score: 1.0	 score: 0.28	 score: 0.18	 score: 0.57
The <u>mascot</u> becoming Beijing 2022 Winter Olympics' viral mascot sensation	 score: 0.95	 score: 0.8	 score: 0.23	 score: 0.13	 score: 0.35
A <u>trampoline</u> with a bowling ball, a basketball, and a ping pong ball placed on it	 score: 0.46	 score: 0.43	 score: 0.34	 score: 1.0	 score: 0.15
A <u>seesaw</u> with a 1-kilogram cotton ball on one end and a 1-kilogram iron cube on the other.	 score: 1.0	 score: 0.65	 score: 0.0	 score: 1.0	 score: 0.2

Figure 7: Qualitative examples with evaluated T2I-ReasonScore.

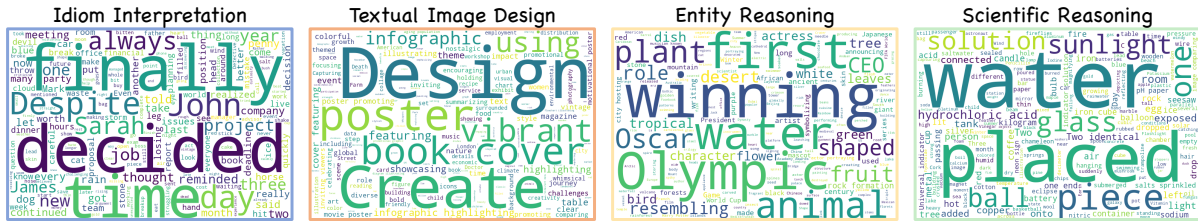


Figure 8: Word cloud to visualize the word distribution of each dimension in our prompt suite.

**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.