Are Text-to-Image Models Inductivist Turkeys? A Counterfactual Benchmark for Causal Reasoning*

Jiayi Lei[1,2], **Yuandong Pu**[1,2], **Xingyu Han**[1], **Rongpeng Zhu**[1], **Jing Xu**[3], **Jinyao Wang**[1]
Zijian Zhou[1], **Bin Fu**[2], **Yuewen Cao**[2†], **Yihao Liu**[2†], **Hongsheng Li**[3†]
[1]Shanghai Jiao Tong University, [2]Shanghai AI Laboratory,
[3]The Chinese University of Hong Kong
[†]Corresponding author

Abstract

Text-to-image (T2I) generation models have achieved remarkable progress in producing visually realistic images from natural language prompts. Yet it remains unclear whether their success reflects genuine causal understanding or sophisticated pattern matching over visual-textual correlations. Inspired by Russell's inductivist turkey, we introduce Counterfactual-World (CF-World), a counterfactual benchmark designed to investigate whether text-to-image models can generate images under rules that systematically contradict real-world priors. CF-World organizes each scenario into three progressive levels: factual generation under ordinary world knowledge, explicit counterfactual generation with direct visual instructions, and implicit counterfactual generation requiring causal deduction from altered rules. We evaluate both open-source and closed-source T2I models using a Vision Language Model (VLM)-based evaluator (CF-Eval). Furthermore, we introduce two metrics: Prior Resistance Rate (PRR), which measures a model's ability to overcome entrenched real-world priors, and Reasoning Retention Rate (RRR), which assesses whether models can maintain reasoning-dependent counterfactual generation without explicit visual cues. Experiments show that all models exhibit sharp degradation from factual to counterfactual settings. Further analyses suggest that these failures arise because current T2I models encode world knowledge and visual appearances as tightly coupled patterns. Consequently, their heavy reliance on frequent visual co-occurrences within the training data forces them to default to familiar commonsense priors when tasked with rendering counterfactual worlds.

1 Introduction

In Bertrand Russell's *Inductivist Turkey* paradox, based on past experience, a turkey equates the farmer's arrival with food, only to be killed on Thanksgiving while awaiting its meal. According to Judea Pearl's Ladder of Causation, the turkey's failure stems from its cognition being restricted to the first level (Association), relying entirely on observed statistical correlations. It lacks the highest level of the ladder—*counterfactual causal reasoning*—which requires the ability to mentally simulate alternative realities, such as the shift in the farmer's intent when Thanksgiving arrives.

Mirroring this inductivist turkey, current text-to-image (T2I) models excel on existing reasoning benchmarks, prompting widespread claims of reasoning capabilities. However, whether they truly possess counterfactual causal reasoning capabilities or merely remain confined to the association level like the turkey cannot be effectively evaluated using existing benchmarks. On one hand, recent reasoning-driven benchmarks by Niu et al. [2025], Li et al. [2023], Chen et al. [2025a], Fu et al. [2024], Li et al. [2025a] focus primarily on conventional scenarios. This fails to separate causal

*Project page: `https://jylei16.github.io/CF-World.github.io/`

Preprint.

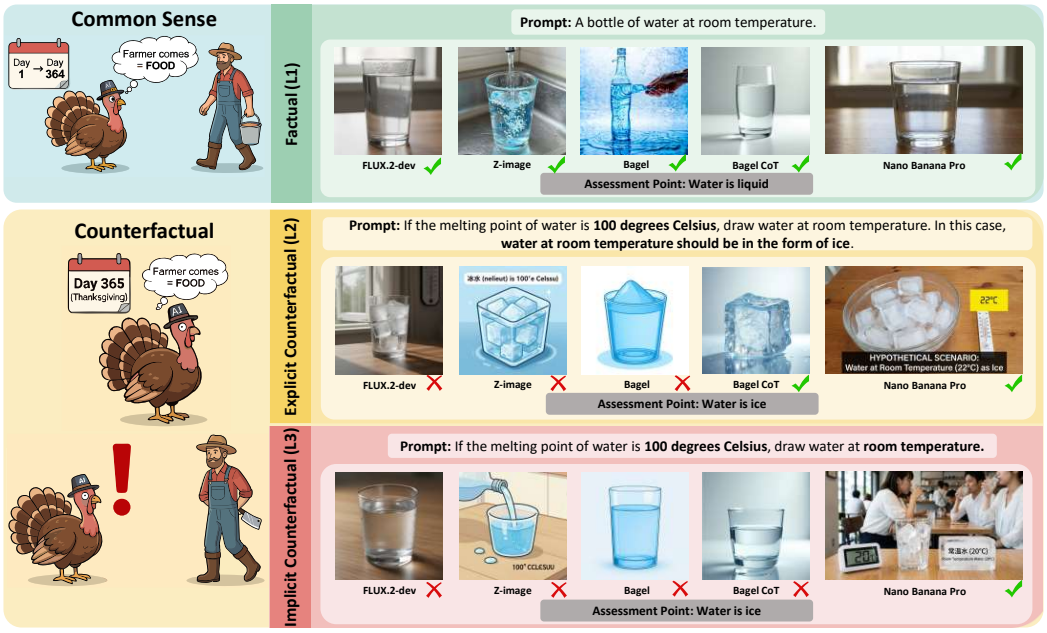


Figure 1: **Text-to-image models as inductivist turkeys.** Left: The inductivist turkey assumes food will always arrive based on past experience, failing to anticipate the counterfactual reality of Thanksgiving. Right: Current T2I models exhibit a similar flaw, evaluated through our three-level progressive framework. (1) **Factual** ($L1$): The prompt aligns with real-world laws. (2) **Explicit Counterfactual** ($L2$): Altered real-world laws with explicitly stated outcomes. (3) **Implicit Counterfactual** ($L3$): Altered real-world laws without explicitly stated outcomes. Ultimately, while models perform reliably on $L1$, they exhibit a substantial decline on $L2$ and $L3$.

reasoning from statistical priors, as successful generation could easily stem from retrieving high-frequency patterns memorized during training. On the other hand, existing counterfactual benchmarks Li et al. [2025b], Zhao et al. [2024] typically target simple semantic combinations of unrelated objects, which are too superficial to assess a model's deductive reasoning and understanding of objective laws. Because of these evaluation gaps, a fundamental question remains unclear: have current T2I models actually climbed the causal ladder to achieve counterfactual causal reasoning?

We introduce the Counterfactual-World Benchmark (**CF-World**), designed to evaluate the counterfactual causal reasoning capabilities of T2I models. We propose a three-level progressive framework (Figure 1): it establishes a factual baseline ($L1$), introduces explicit counterfactual by providing both the altered objective laws and their outcomes ($L2$), and advances to implicit counterfactuals providing only the altered laws, requiring the model to deduce the visual outcomes ($L3$). This progressive design systematically isolates true logical deduction from mere statistical memorization. To rigorously quantify models' performance, we develop **CF-Eval**, an automated pipeline assessing Visual Integrity, Assessment Point, and Logic Consistency. We also introduce two metrics—Prior Resistance Rate (PRR) and Reasoning Retention Rate (RRR)—to measure a model's resistance to real-world priors and its counterfactual causal reasoning. Under this rigorous setting, we surprisingly find that even state-of-the-art (SOTA) models experience a significant performance decline on $L2$ and $L3$, indicating their counterfactual causal reasoning capabilities remain relatively weak.

To understand this decline, we conduct diagnostic probes. **First, to isolate logical deduction**, we evaluate models on abstract symbolic elements under factual versus counterfactual rules. The universal performance drop in counterfactual scenarios suggests that, even without visual burdens, models fundamentally struggle with counterfactual causal reasoning. **Second, to isolate visual generation**, we test whether models can execute counterfactual visual recombination when causal reasoning is removed by combining rarely co-occurring concepts. We observe another universal performance drop. Digging deeper into this visual entanglement, we find that replacing high-frequency nouns with equivalent descriptive phrases yields notable improvements, revealing that models heavily rely on rigid text-image alignment shortcuts. **Ultimately, these dual failures highlight a fundamental bottleneck:**

high-dimensional statistical priors constrain T2I models' decoupling abilities. Because they primarily learn pixel co-occurrences and text-image alignment, they struggle to decouple independent causal variables for logical reasoning and basic attribute modules for visual recombination, tending to default to the high-frequency commonsense priors found in their training data.

The main contributions of this paper are summarized as follows:

1) We introduce **CF-World**, the first counterfactual world knowledge benchmark structured through a systematic three-level progressive framework, bridging the critical gap in existing generative T2I evaluations by rigorously assessing both logical and causal reasoning under counterfactual premises.

2) We propose **CF-Eval**, an automated evaluation pipeline that introduces two novel quantitative metrics (PRR and RRR) to rigorously and objectively quantify models' causal reasoning capabilities.

3) We analyze **causes of generation failures**, demonstrating that T2I models fail to decouple rules and attributes, limiting their capacity for higher-level logical reasoning independent of visual composition.

Table 1: **Comparison of CF-World with existing T2I evaluation benchmarks.** Our benchmark uniquely introduces a progressive framework to evaluate models' true reasoning ability under challenging counterfactual scenarios that eliminate training priors.

Benchmark	Reasoning Ability	Counterfactual Setting	Progressive Design
WISE(Niu et al. [2025])	✓	×	×
VQAI(Li et al. [2023])	✓	×	×
R2I-Bench(Chen et al. [2025a])	✓	×	×
Commonsense-T2I(Fu et al. [2024])	✓	×	×
T2I-ReasonBench(Sun et al. [2025])	✓	×	×
GIR-Bench(Li et al. [2025a])	✓	×	×
ELNP(Li et al. [2025b])	×	✓	×
LC-Mis(Zhao et al. [2024])	×	✓	×
CF-World (Ours)	✓	✓	✓

2 Related Work

2.1 Text-to-Image Generation

Text-to-image (T2I) synthesis has evolved rapidly, driven by breakthroughs across diverse architectural paradigms. Prominent approaches include diffusion-based methods Esser et al. [2024], Xie et al. [2025a], Qin et al. [2025], Yang et al. [2024], autoregressive models Sun et al. [2024], Zhang et al. [2024], Chen et al. [2025b], Wang et al. [2024], and unified multimodal frameworks Xiao et al. [2024], Xie et al. [2024], Zhou et al. [2024], Tong et al. [2024], Sun et al. [2023]. To tackle increasingly complex user prompts, recent models have integrated advanced techniques such as Chain-of-Thought (CoT) prompting Liao et al. [2025] and reinforcement learning Guo et al. [2025], Jiang et al. [2025].

While these models excel at general generation, their reasoning capabilities face significant challenges when extended to counterfactual scenarios. In T2I synthesis, counterfactual generation requires rendering scenes that deviate from reality while preserving internal causal consistency. Foundational techniques attempt to achieve this by disentangling causal features via Generative Causal Models Yue et al. [2021] or employing fine-tuning strategies like DreamBooth Ruiz et al. [2022] to maintain subject identity across novel contexts He et al. [2023]. Despite these capabilities, models often revert to training-data biases when faced with unusual concept combinations, leading to latent concept misalignment Zhao et al. [2024]. Although recent advancements have attempted to correct these misalignments in physically implausible scenes through step-by-step latent space manipulation Li et al. [2025b], these solutions remain constrained to visual attribute editing and concept co-occurrence. In contrast to these visually driven alignment methods, our work extends this paradigm by evaluating how models handle the systematic alteration of objective laws and causal logic.

2.2 Text-to-Image Evaluation Benchmarks and Metrics

To evaluate T2I generation, numerous benchmarks and metrics have been proposed in recent years. In terms of benchmarks, datasets such as GeckoNum(Kaji'c et al. [2024]), Winoground(Thrush et al. [2022]), GenEval(Ghosh et al. [2023]), and GenAI-Bench(Li et al. [2024]) are widely used to assess compositional and numerical alignment. Meanwhile, OK-VQA(Marino et al. [2019]), WISE(Niu et al. [2025]), Commonsense-T2I(Fu et al. [2024]), R2I-Bench(Chen et al. [2025a]), T2I-ReasonBench Sun et al. [2025], StructBench(Zhuo et al. [2025]) and PICABench(Pu et al. [2025]) have been introduced to explore knowledge-based and commonsense generation. To quantify these abilities, various metrics have been developed, ranging from traditional visual-text alignment scores such as CLIPScore(Hessel et al. [2021]), DSGScore(Cho et al. [2023]), and VQAScore(Lin et al. [2024]), to LLM-assisted evaluators including LLMScore(Lu et al. [2023]), SemVarEffect(Zhu et al. [2024]), RIScore(Zhao et al. [2025]), and WIScore(Niu et al. [2025]).

3 Counterfactual-World (CF-World)

3.1 Overview

To systematically evaluate the counterfactual causal reasoning capability of T2I models, we introduce Counterfactual-World (CF-World). As illustrated in Figure 2, CF-World is a comprehensive benchmark comprising 1,091 groups and a total of 3,273 prompts. These prompts span five major disciplines: Physics (including branches such as Classical Mechanics, Optics, Thermodynamics, Astronomy, and Electromagnetism), Biology, Chemistry, Geography, and Sociology.

To isolate a model's reasoning capacity from its basic rendering, we design a three-level progressive framework for our prompts: **Factual ($L1$):** Follows real-world objective laws to verify the model's basic priors. **Explicit Counterfactual ($L2$):** Alters real-world laws but explicitly states the visual outcomes. This tests whether the model can overcome commonsense to render the counterfactual state. **Implicit Counterfactual ($L3$):** Alters real-world laws without explicitly stating the outcomes, forcing the model to perform autonomous causal deduction. Each prompt is paired with a tailored assessment point, serving as the decisive visual criterion for automated scoring.

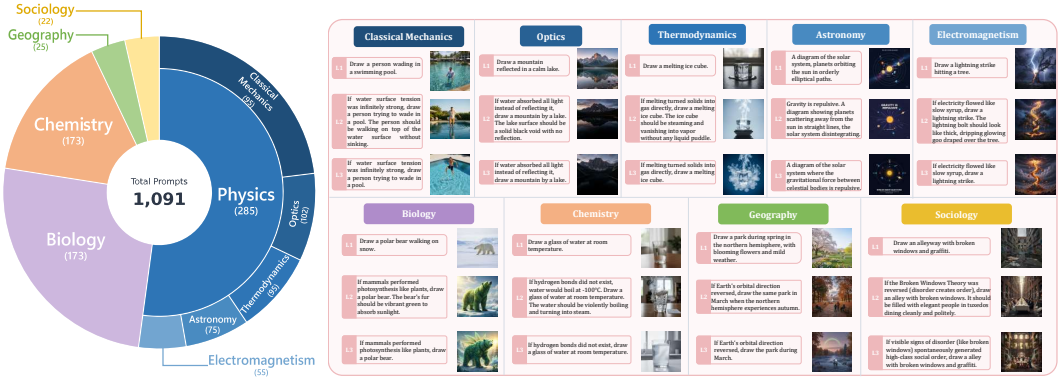
3.2 Benchmark Construction

To ensure the high quality and scientific validity of CF-World, we develop a rigorous pipeline encompassing taxonomy definition, data generation, and human-in-the-loop quality assurance.

Taxonomy. CF-World is organized around a discipline-oriented taxonomy to evaluate counterfactual reasoning across diverse types of objective world knowledge. Because counterfactual generation requires models to hypothesize against established laws, we select target laws that define the counterfactual worlds to be tested. To avoid evaluating obscure expert knowledge, we focus on basic laws commonly taught in middle-school curricula and group them into five disciplines: Physics, Biology, Chemistry, Geography, and Sociology. Each category targets counterfactual scenarios grounded in its discipline, with the required domain knowledge consistently controlled at the middle-school level. This design ensures that model failures are less likely to reflect a lack of specialized expertise, and more likely to reveal limitations in counterfactual rule understanding and application.

Data Generation Pipeline. We construct the dataset by first manually curating fundamental scientific principles. Subsequently, Large Language Models (LLMs) are employed to generate the corresponding prompts based on our three-level progressive framework ($L1$, $L2$, $L3$). Factual ($L1$) aligns with real-world laws. Explicit Counterfactual ($L2$) involves altered real-world laws with explicitly stated outcomes. Implicit Counterfactual ($L3$) alters real-world laws without explicitly stating the outcomes, requiring T2I models to complete the reasoning process and generate counterfactual images. Alongside the prompts, the LLMs also generate a concise assessment point for each instance, which serves as the ground truth and a critical basis for subsequent automated evaluation.

To ensure high-quality outputs, we explicitly instruct the LLMs to adhere to four core criteria during generation. Firstly, **Visual Unambiguity** is essential, ensuring clear visual *features* for Vision-Language Model (VLM) evaluation. Secondly, we emphasize the **Logical Deduction Requirement**, demanding fundamental reasoning rather than mere stylistic changes. Thirdly, we prioritize **Safety** by strictly avoiding NSFW or body-horror content. Lastly, **Scientific Validity** is crucial, ensuring



(a) Data Distribution of CF-World (b) Overview of CF-World

Figure 2: The dataset distribution and selected qualitative examples of CF-World. (a) The data distribution of CF-World. (b) Qualitative examples across the five disciplines.

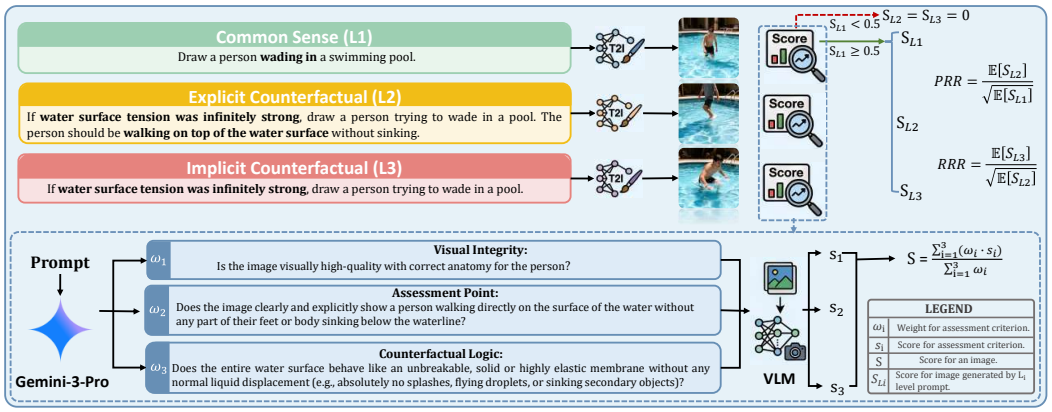


Figure 3: **CF-Eval.** CF-Eval is a multi-dimensional scoring pipeline, featuring sequential thresholding ($S_{L1} \geq 0.5$) and two metrics: Prior Resistance Rate (PRR) and Reasoning Retention Rate (RRR).

accurate deductions and comprehensible generative targets. Together, these criteria guide the LLMs in producing outputs that are not only high-quality but also responsible and meaningful for CF-World.

Human Review. Following the automated LLM generation process, a team of expert human annotators meticulously review and filter the dataset. This human-in-the-loop verification ensures that all generated prompts and assessment points strictly meet the aforementioned criteria, eliminating any generation artifacts or ambiguities to guarantee a rigorous and high-quality benchmark.

3.3 CF-Eval

To evaluate the generative capabilities of models in factual and counterfactual scenarios, we propose **CF-Eval**. As illustrated in Figure 3, CF-Eval is an automated evaluation pipeline driven by Vision-Language Models (VLMs). We detail its scoring mechanisms and metrics below.

3.3.1 Evaluation Dimensions

For each prompt, evaluation questions are generated by Gemini-3-Pro, covering three distinct dimensions. We assign differentiated weights (w_i) to reflect their relative importance:

1) Visual Integrity (Weight 1-3): Evaluates fundamental, style-agnostic image quality as a basic viability threshold. **2) Assessment Point (Weight 12-16):** Derived from the specific assessment point, this evaluation question is formulated to strictly assess the main subject's adherence to the factual or counterfactual rule. **3) Logic Consistency (Weight 7-9):** Verifies that the surrounding context aligns with the established physical or counterfactual setting, ensuring global coherence.

3.3.2 Score Calculation and Thresholding

Upon receiving the VLM's scores as a continuous value between 0 and 1 ($s_i \in [0, 1]$) for each dimension, the base score of a single image S is calculated via a weighted average:

$$S = \frac{\sum_{i=1}^{3}(w_i \cdot s_i)}{\sum_{i=1}^{3} w_i}.$$

To ensure a model's performance on counterfactual reasoning (L2/L3) is meaningful, we introduce a conditional thresholding mechanism. Specifically, to prevent false positives, counterfactual scores (S_{L2}, S_{L3}) are calculated only if the factual baseline is met ($S_{L1} \geq 0.5$); otherwise, they are set to zero. This sequential design ensures that the model understands basic facts before evaluating counterfactuals. If a model fails the fundamental factual generation task, any success in the subsequent counterfactual task might be coincidental, rendering the counterfactual scores meaningless. The 0.5 threshold is empirically calibrated based on human alignment, as detailed in Appendix B.

3.3.3 Prior Resistance Rate (PRR) and Reasoning Retention Rate (RRR)

To evaluate models across our progressive framework, we introduce two newly designed evaluation metrics: Prior Resistance Rate (PRR) and Reasoning Retention Rate (RRR). PRR measures a model's ability to resist real-world priors when given explicit counterfactual instructions, isolating the performance shift from standard generation ($L1$) to explicit counterfactual generation ($L2$) as

$$PRR = \frac{\mathbb{E}[S_{L2}]}{\sqrt{\mathbb{E}[S_{L1}]}}.$$

A low PRR suggests concept lock-in, indicating reliance on common-sense priors. Building upon this, RRR quantifies the model's causal reasoning by measuring how effectively it retains counterfactual capabilities without explicit visual cues (transitioning from $L2$ to $L3$), defined as

$$RRR = \frac{\mathbb{E}[S_{L3}]}{\sqrt{\mathbb{E}[S_{L2}]}}.$$

A high RRR indicates minimal score degradation from $L2$ to $L3$, demonstrating that the model can effectively rely on its intrinsic reasoning capabilities to fill in the missing reasoning results in $L3$.

For both metrics, we avoid pure ratios (i.e., $\mathbb{E}[S_{L2}]/\mathbb{E}[S_{L1}]$ or $\mathbb{E}[S_{L3}]/\mathbb{E}[S_{L2}]$) to prevent artificially inflated scores when a model's foundational performance is low. By calculating the geometric mean of the absolute score and the relative ratio (e.g., $\sqrt{\mathbb{E}[S_{L3}] \times (\mathbb{E}[S_{L3}]/\mathbb{E}[S_{L2}])}$), this unified formulation penalizes models with low foundational scores, ensuring that high scores reflect both strong retention and a clear capability in overcoming priors and performing autonomous deduction.

4 Experiments

4.1 Setup

To evaluate the counterfactual reasoning capabilities of current text-to-image models, we select diverse state-of-the-art systems, as shown in Figure 4. Our evaluation suite encompasses prominent open-source models: SANA 1.5 Xie et al. [2025a], Janus-Pro-7B Chen et al. [2025b], Show-o2 Xie et al. [2025b], Z-image Team et al. [2025], Lumina-DiMOO Xin et al. [2025], BAGEL and BAGEL-CoT Deng et al. [2025], OmniGen2 Wu et al. [2025], FLUX.2-dev Black Forest Labs [2025], and Qwen-Image. We also test leading closed-source models such as Nano Banana NanoBanana [2025], Nano Banana Pro, GPT-Image-1.5 OpenAI [2025], and Seedream 5.0 ByteDance [2025]. To ensure scoring objectivity across these diverse architectures, we employ two highly capable Vision-Language Models (VLMs) as evaluators: Qwen3-VL-235B Team [2025] and Gemini-3-Pro.

4.2 Main Results and Analysis

Table 2 reports the generative performance of all evaluated models across the three progressive prompt levels: Factual ($L1$), Explicit Counterfactual ($L2$), and Implicit Counterfactual ($L3$), alongside the

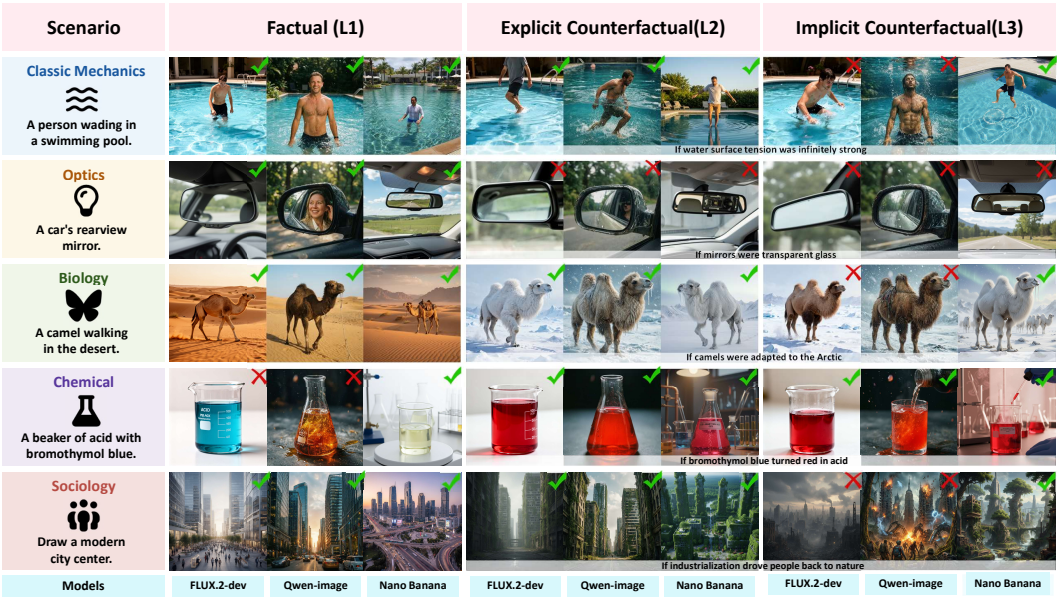


Figure 4: **Qualitative Comparison of Model Generations.** A detailed visual comparison of selected models given identical prompts sampled from the five different scientific domains.

calculated PRR and RRR metrics. Based on these results, we derive the following three core observations regarding model capabilities, reasoning bottlenecks, and architectural paradigms:

The Prior Lock-in in Counterfactual Generation. Our systematic evaluation reveals a substantial performance decline from factual ($L1$) to explicit counterfactual generation ($L2$). While most open-source models achieve strong $L1$ scores, their $L2$ performance drops significantly, yielding PRRs largely below 0.50. Interestingly, models with superior foundational capabilities (e.g., Qwen-Image) do not always exhibit proportional advantages in counterfactual tasks; they often yield lower PRRs than models with weaker $L1$ baselines. This paradox suggests that extensive reliance on training data exacerbates the "prior lock-in" effect, where visual representations and real-world knowledge become so deeply entangled that stronger priors actively hinder counterfactual rendering.

Bottlenecks in Causal Reasoning. Performance degrades further in the implicit counterfactual ($L3$) setting, as evidenced by the consistent drop in RRR across open-source models (see Figure 5b). This degradation highlights severe limitations in autonomous causal deduction. Comparing BAGEL with BAGEL-CoT reveals that explicit text-side logic injection provides only a marginal boost. We hypothesize this stems from a fundamental modality gap: while the discrete nature of natural language facilitates logical decoupling (allowing text-side CoT to deduce the correct state), the continuous visual representations in diffusion models remain highly entangled. Consequently, the denoising network struggles to execute deduced logic visually, making implicit reasoning a major bottleneck.

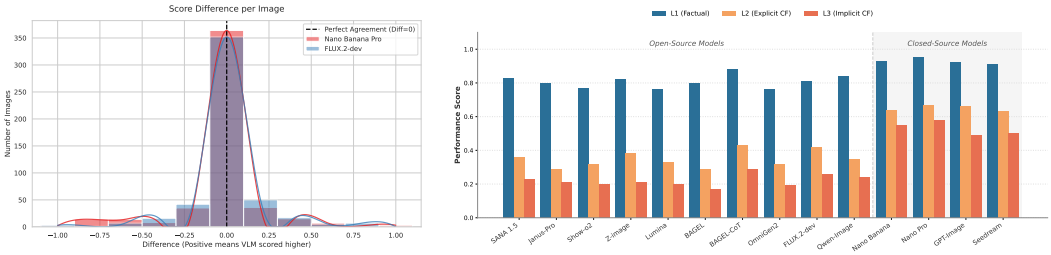
Performance Comparison of Different Models A clear performance gap exists between closed-source and open-source models. Top-tier closed-source models (e.g., Nano Banana Pro) maintain robust scores across both $L2$ and $L3$, which may be attributed to their use of large-scale, high-quality alignment data and specific architectural optimizations. Among open-weight models, architectural choices play a crucial role. Native multimodal and unified architectures (e.g., OmniGen2, Show-o2, Janus-Pro) are consistently outperformed by FLUX.2-dev. Despite the theoretical advantages of unified token spaces, our findings suggest that heavily scaled text encoders may currently be more effective at mitigating attribute entanglement than end-to-end unified architectures.

4.3 Human-VLM Scoring Consistency

To rigorously validate our automated evaluation pipeline, we conducted a comprehensive alignment study comparing human judgments against Gemini-3-Pro. We randomly sampled a total of 1,000 images generated by representative models (FLUX.2-dev and Nano Banana Pro). For each image,

Table 2: Main evaluation results on the CF-World dataset. All metrics are scaled to 0-1. PRR and RRR are calculated to quantify reasoning robustness. The best performing open-source models in each column are highlighted in blue, while the best closed-source models are highlighted in pink.

Model	Qwen3-VL-235B					Gemini-3-Pro				
	L1	L2	L3	PRR↑	RRR↑	L1	L2	L3	PRR↑	RRR↑
Open-Source Models										
SANA 1.5	0.83	0.36	0.23	0.40	0.38	0.75	0.29	0.17	0.33	0.32
Janus-Pro-7B	0.80	0.29	0.21	0.32	0.39	0.69	0.21	0.11	0.25	0.24
Show-o2	0.77	0.32	0.20	0.36	0.35	0.66	0.25	0.14	0.31	0.28
Z-image	0.82	0.38	0.21	0.42	0.34	0.75	0.33	0.16	0.38	0.28
Lumina-DiMOO	0.76	0.33	0.20	0.38	0.35	0.70	0.29	0.17	0.35	0.32
BAGEL	0.80	0.29	0.17	0.32	0.32	0.73	0.29	0.15	0.34	0.28
BAGEL-CoT	0.88	0.43	0.29	0.46	0.44	0.82	0.41	0.26	0.45	0.41
OmniGen2	0.76	0.32	0.19	0.37	0.34	0.70	0.29	0.18	0.35	0.33
FLUX.2-dev	0.81	0.42	0.26	0.47	0.40	0.83	0.48	0.28	0.53	0.40
Qwen-Image	0.84	0.35	0.24	0.38	0.41	0.80	0.37	0.23	0.41	0.38
Closed-Source Models										
Nano Banana	0.93	0.64	0.55	0.66	0.69	0.88	0.64	0.52	0.68	0.65
Nano Banana Pro	0.95	0.67	0.58	0.69	0.71	0.93	0.76	0.67	0.79	0.77
GPT-Image-1.5	0.92	0.66	0.49	0.69	0.60	0.91	0.73	0.55	0.77	0.64
Seedream 5.0	0.91	0.63	0.50	0.66	0.63	0.89	0.72	0.61	0.76	0.72



(a) Human-VLM Consistency

(b) Performance Degradation

Figure 5: **Quantitative Evaluation and Consistency Analysis.** **(a)** The distribution of score differences ($Score_{VLM} - Score_{Human}$) per image. The peak at 0 indicates strong alignment between the VLM and humans. **(b)** Performance degradation across factual ($L1$) and counterfactual ($L2, L3$) tasks. Open-source models exhibit a severe performance drop when transitioning to counterfactual scenarios, whereas closed-source models demonstrate stronger reasoning retention.

both the VLM and human evaluators provided scores based on prompt-specific questions dynamically generated by our CF-Eval pipeline. The human ground truth was established by averaging independent scores from three expert annotators. These annotators are graduate-level researchers with domain expertise in computer vision and generative AI, who underwent rigorous pre-evaluation training specifically calibrated to our counterfactual criteria. All scores were normalized to a continuous scale of $[0, 1]$ and aggregated using identical methods. As illustrated in Figure 5(a), the score differences ($Score_{VLM} - Score_{Human}$) are overwhelmingly concentrated within the narrow interval of $[-0.125, 0.125]$. This minimal variance confirms that Gemini-3-Pro robustly internalizes our assessment standards, serving as a highly reliable proxy for expert human evaluation.

5 Why Models Fail: A Decoupling Perspective

To investigate the root cause of text-to-image models' deficiency in counterfactual causal reasoning, triggering degradation on $L2$ and $L3$ tasks, we design three targeted mechanistic experiments. The first two isolate two orthogonal failure axes: **logical reasoning** and **visual recombination**. The third, De-nominalization, serves as a diagnostic bridge that traces both failures to a shared lexical root.

Table 3: Comprehensive results of the mechanistic investigation. All raw metrics are scaled to 0–1. "Fact." and "CF" denote Factual and Counterfactual settings. For de-nominalization, red superscripts indicate the performance gain over the original L2 prompts. The best and second-best raw results in each column are highlighted in green and yellow, respectively.

Model	Type	Rule Decoupling		Attribute Decoupling		De-nominalization	
		Fact.	CF	Fact.	CF	L2	De-norm
SANA 1.5	Diff.	0.31	0.30	0.94	0.83	0.36	0.37$^{+0.01}$
Janus-Pro-7B	Unif.	0.19	0.07	0.97	0.83	0.29	0.30$^{+0.01}$
Show-o2	Unif.	0.39	0.37	0.92	0.80	0.32	0.37$^{+0.05}$
Z-image	Diff.	0.61	0.53	0.98	0.89	0.38	0.43$^{+0.05}$
Lumina-DiMOO	Unif.	0.38	0.34	0.97	0.82	0.33	0.35$^{+0.02}$
BAGEL	Unif.	0.29	0.22	0.98	0.83	0.29	0.31$^{+0.02}$
BAGEL-CoT	Unif.	0.38	0.32	0.97	0.90	0.43	0.44$^{+0.01}$
OmniGen2	Unif.	0.33	0.25	0.96	0.81	0.32	0.35$^{+0.03}$
FLUX.2-dev	Diff.	0.53	0.52	0.99	0.90	0.42	0.51$^{+0.09}$
Qwen-Image	Diff.	0.40	0.40	0.97	0.86	0.35	0.37$^{+0.02}$

5.1 Causal Decoupling

To isolate logical rule execution from visual generation complexity, we evaluate models on a curated symbolic benchmark composed of 198 prompts covering 33 objective rules. Each rule includes 1–2 factual baselines and 4–5 counterfactual variants, where the perturbations are deliberately multi-dimensional rather than simple binary reversals, such as changing the direction of gravity to leftward, rightward, or upward. For each prompt, we further provide an assessment point specifying the intended rule condition, which is converted into a targeted evaluation question by an LLM (Qwen3-30B) and scored by a VLM judge (Qwen3-VL-235B). This design reduces the influence of open-ended visual quality and focuses the evaluation on counterfactual rule-following ability.

As Table 3 shows, most models obtain relatively low absolute scores under counterfactual rules, even in this simplified symbolic setting where visual clutter and object recognition demands are minimized. This suggests that current image generation models still struggle to execute counterfactual rules in a grounded and compositional manner, rather than being limited only by complex visual rendering. However, the results do not support a uniform-collapse interpretation: the performance drop varies substantially across models. The results indicate that factual and counterfactual rule execution are closely related: models performing better on factual rules often achieve higher counterfactual scores. This pattern suggests that counterfactual failures are not solely caused by violations of memorized factual priors, but are also tied to a more general limitation in representing and applying symbolic rules. Additionally, diffusion-based models tend to achieve higher factual and counterfactual scores than unified models in this benchmark, with Z-image and FLUX.2-dev among the strongest performers. Nevertheless, this architectural trend should be interpreted cautiously given the limited number of evaluated models. Overall, the benchmark reveals that while some models preserve performance better under rule perturbations, robust counterfactual rule-following remains broadly underdeveloped.

5.2 Attribute Decoupling

To further examine whether image generation models can recombine visual attributes beyond frequently observed co-occurrences, while abstracting away the burden of rule-level logical reasoning, we evaluate them under an attribute decoupling setting. We sample 100 rare concept pairs from LC-Mis Zhao et al. [2024] as the counterfactual condition, and use Gemini-3-Pro to construct a corresponding common co-occurring pair for each of them as the factual condition. Gemini-3-Pro then converts each concept pair into an image generation prompt, yielding paired factual and counterfactual prompts for evaluation. Generated images are evaluated by Qwen3VL-235B, which assigns a normalized score between 0 and 1 according to the prompt. The score measures both whether the generated image contains the required concepts and whether their relationship is correctly instantiated.

As Table 3 shows, models perform strongly in the factual condition. However, performance consistently drops under the counterfactual rare-pair condition. Overall, attribute decoupling results suggest

that current T2I models have some but limited ability to recombine visual concepts beyond frequent co-occurrences. These findings suggest that while visual attribute recombination remains an open challenge, the primary bottleneck of current models lies in decoupling and manipulating higher-level generative rules, which require stronger logical reasoning beyond perceptual composition.

5.3 De-nominalization

Building on the attribute decoupling analysis above, we further examine whether such object–attribute entanglement also affects L2 counterfactual generation. We therefore conduct a de-nominalization experiment on L2 prompts, replacing only the target object or attribute nouns in the inferred outcome with descriptive phrases while preserving the counterfactual law. This tests whether bypassing explicit nominal cues facilitates attribute decoupling by reducing default object–attribute associations.

As Table 3 shows, de-nominalization consistently improves performance across all models, but the gains are generally modest. The largest improvement is observed for FLUX.2-dev ($+0.09$), followed by Z-image and Show-o2 ($+0.05$), while several models, including Janus-Pro-7B, SANA 1.5, and BAGEL-CoT, show only marginal gains ($+0.01$). These results show that explicit object or attribute names introduce measurable interference by activating learned visual priors, as de-nominalization consistently improves performance across models. However, the limited magnitude of these gains suggests that lexical priors are not the primary source of $L2$ failures. Consistent with the attribute decoupling results, this finding further indicates that the main bottleneck of current models lies in reasoning over higher-level generative rules, rather than merely composing perceptual elements.

5.4 Analysis and Discussion

1) Asymmetric Decoupling. Factual attribute rendering approaches perfection (> 0.92), yet factual rule execution remains poor (< 0.61). This observed asymmetry is quantitative as well as qualitative: attribute scores retain over 0.81 of their factual value under counterfactual stress, whereas rule scores collapse to as low as 0.37. We hypothesize that generative models easily master shallow visual interpolation, but still lack the deep physical grounding required for genuine logical deduction.

2) Lexical Vulnerability. Scaled text encoders (e.g., FLUX.2-dev) benefit from de-nominalization ($+0.09$), while unified models (e.g., Janus-Pro-7B) show minimal gains. This suggests that diffusion model entanglement is driven by superficial lexical shortcuts, whereas unified models exhibit deeper, semantic-level entanglement enforcing conceptual priors regardless of linguistic variations.

3) Two Regimes of Entanglement. Synthesizing the diagnostics above reveals a clean dichotomy in how generative models fail. Diffusion models suffer from shallow, lexical entanglement: their substantial de-nominalization gains ($+0.05$ to $+0.09$) show that priors are bound to surface word embeddings and can be partially unlocked by mere linguistic rephrasing. Unified models suffer from deep, semantic entanglement: their negligible gains (≤ 0.02) indicate that conceptual priors persist regardless of phrasing, residing below the lexical surface in the shared representation space. This distinction carries a direct design implication: the two model families demand fundamentally different remedies—prompt- or encoder-level intervention may suffice for diffusion models, but unified models require representation-level grounding to achieve genuine compositional reasoning.

6 Conclusion and Limitations

Extensive evaluations reveal that while SOTA generative models excel in factual settings, their performance declines significantly under counterfactual conditions. Our mechanistic investigation demonstrates that this degradation stems from a fundamental inability to decouple: models fail to decouple objective world knowledge from default scenarios and struggle to separate visual attributes from their corresponding subjects, remaining deeply entangled in high-frequency priors. More importantly, compared with failures in visual attribute recombination, the inability to decouple and reason over higher-level generative rules constitutes the dominant performance bottleneck.

While our current work is fundamentally diagnostic and does not propose an algorithmic solution to concept entanglement, identifying these root causes is a critical first step. We hope CF-World will serve as a robust testing ground, inspiring future research to develop novel decoupling mechanisms and advance multimodal models from prior-driven generation toward genuine causal reasoning.

References

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025. Accessed: 2026-03-14.

ByteDance. Seedream. https://www.bytedance.com, 2025. Multimodal image generation and editing system.

Kaijie Chen, Zihao Lin, Zhiyang Xu, Ying Shen, Yuguang Yao, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. R2i-bench: Benchmarking reasoning-driven text-to-image generation. In *Conference on Empirical Methods in Natural Language Processing*, 2025a. URL https://api.semanticscholar.org/CorpusID:278996759.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *ArXiv*, abs/2501.17811, 2025b. URL https://api.semanticscholar.org/CorpusID:275954151.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *ArXiv*, abs/2310.18235, 2023. URL https://api.semanticscholar.org/CorpusID:264555374.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiao-Ping Nie, Ziang Song, Shi Guang, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *ArXiv*, abs/2505.14683, 2025. URL https://api.semanticscholar.org/CorpusID:278768720.

Patrick Esser, Sumith Kulal, A. Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024. URL https://api.semanticscholar.org/CorpusID:268247980.

Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. Commonsense-t2i challenge: Can text-to-image generation models understand commonsense? *ArXiv*, abs/2406.07546, 2024. URL https://api.semanticscholar.org/CorpusID:270380362.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *ArXiv*, abs/2310.11513, 2023. URL https://api.semanticscholar.org/CorpusID:264288728.

Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. *ArXiv*, abs/2501.13926, 2025. URL https://api.semanticscholar.org/CorpusID:275820253.

Xingzhe He, Zhiwen Cao, Nicholas Kolkin, Lantao Yu, Kun Wan, Helge Rhodin, and Ratheesh Kalarot. A data perspective on enhanced identity preservation for diffusion personalization. *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3782–3791, 2023. URL https://api.semanticscholar.org/CorpusID:265050824.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *ArXiv*, abs/2104.08718, 2021. URL https://api.semanticscholar.org/CorpusID:233296711.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *ArXiv*, abs/2505.00703, 2025. URL https://api.semanticscholar.org/CorpusID:278237703.

Ivana Kaji'c, Olivia Wiles, Isabela Albuquerque, Matthias Bauer, Su Wang, Jordi Pont-Tuset, and Aida Nematzadeh. Evaluating numerical reasoning in text-to-image models. *ArXiv*, abs/2406.14774, 2024. URL https://api.semanticscholar.org/CorpusID:270688621.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation. *ArXiv*, abs/2406.13743, 2024. URL https://api.semanticscholar.org/CorpusID:270619531.

Hongxiang Li, Yaowei Li, Bin Lin, Yuwei Niu, Yuhang Yang, Xiaoshuang Huang, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Long Chen. Gir-bench: Versatile benchmark for generating images with reasoning. *ArXiv*, abs/2510.11026, 2025a. URL https://api.semanticscholar.org/CorpusID:282058070.

Sifan Li, Ming Tao, Haoyu Zhao, Ling Shao, and Hao Tang. Replace in translation: Boost concept alignment in counterfactual text-to-image. *ArXiv*, abs/2505.14341, 2025b. URL https://api.semanticscholar.org/CorpusID:278768748.

Xiaochuan Li, Baoyu Fan, Runze Zhang, Liang Jin, Di Wang, Zhenhua Guo, Yaqian Zhao, and Rengang Li. Image content generation with causal reasoning. *ArXiv*, abs/2312.07132, 2023. URL https://api.semanticscholar.org/CorpusID:266174326.

Jiaqi Liao, Zhengyuan Yang, Linjie Li, Dianqi Li, Kevin Qinghong Lin, Yu Cheng, and Lijuan Wang. Imagegen-cot: Enhancing text-to-image in-context learning with chain-of-thought reasoning. *ArXiv*, abs/2503.19312, 2025. URL https://api.semanticscholar.org/CorpusID:277313203.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, 2024. URL https://api.semanticscholar.org/CorpusID:268857167.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *ArXiv*, abs/2305.11116, 2023. URL https://api.semanticscholar.org/CorpusID:258762865.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3190–3199, 2019. URL https://api.semanticscholar.org/CorpusID:173991173.

NanoBanana. Nanobanana. https://nanobananaimg.com/, 2025. Multimodal image generation and editing system.

Yuwei Niu, Munan Ning, Mengren Zheng, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *ArXiv*, abs/2503.07265, 2025. URL https://api.semanticscholar.org/CorpusID:276929205.

OpenAI. Gpt-image. https://openai.com, 2025. Multimodal image generation and editing system.

Yuandong Pu, Le Zhuo, Songhao Han, Jinbo Xing, Kaiwen Zhu, Shuo Cao, Bin Fu, Si Liu, Hongsheng Li, Yu Qiao, Wenlong Zhang, Xi Chen, and Yihao Liu. Picabench: How far are we from physically realistic image editing? *ArXiv*, abs/2510.17681, 2025. URL https://api.semanticscholar.org/CorpusID:282210186.

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, Xiangyang Zhu, Manyuan Zhang, Will Beddow, Erwann Millon, Victor Perez, Wen-Hao Wang, Conghui He, Bo Zhang, Xiaohong Liu, Hongsheng Li, Yu-Hao Qiao, Chang Xu, and Peng Gao. Lumina-image 2.0: A unified and efficient image generative framework. *ArXiv*, abs/2503.21758, 2025. URL https://api.semanticscholar.org/CorpusID:277349538.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2022. URL https://api.semanticscholar.org/CorpusID:251800180.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, and Xihui Liu. T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation. *ArXiv*, abs/2508.17472, 2025. URL https://api.semanticscholar.org/CorpusID:280711328.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *ArXiv*, abs/2406.06525, 2024. URL https://api.semanticscholar.org/CorpusID:270371603.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *ArXiv*, abs/2307.05222, 2023. URL https://api.semanticscholar.org/CorpusID:259765944.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Junhan Shi, Qilong Wu, Fengyi Yu, Chi Zhang, Shifeng Zhang, and Shilin Zhou. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *ArXiv*, abs/2511.22699, 2025. URL https://api.semanticscholar.org/CorpusID:283438115.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5228–5238, 2022. URL https://api.semanticscholar.org/CorpusID:248006414.

Shengbang Tong, David Fan, Jiacheng Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *ArXiv*, abs/2412.14164, 2024. URL https://api.semanticscholar.org/CorpusID:274823104.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Lian zi Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. *ArXiv*, abs/2409.18869, 2024. URL https://api.semanticscholar.org/CorpusID:272968818.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Towards instruction-aligned multimodal generation. 2025. URL https://api.semanticscholar.org/CorpusID:279999713.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13294–13304, 2024. URL https://api.semanticscholar.org/CorpusID:272694523.

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, and Song Han. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *ArXiv*, abs/2501.18427, 2025a. URL https://api.semanticscholar.org/CorpusID:275993956.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *ArXiv*, abs/2408.12528, 2024. URL https://api.semanticscholar.org/CorpusID:271924334.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *ArXiv*, abs/2506.15564, 2025b. URL https://api.semanticscholar.org/CorpusID:279447576.

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Ke Wang, Yibin Wang, Jinbin Bai, Qian Yu, Dengyang Jiang, Yuandong Pu, Haoxing Chen, Le Zhuo, Junjun He, Gen Luo, Tian-Xin Li, Ming Hu, Jin Ye, Shenglong Ye, Bo Zhang, Chang Xu, Wenhai Wang, Hongsheng Li, Guangtao Zhai, Tianfan Xue, Bin Fu, Xiaohong Liu, Yu Qiao, and Yihao Liu. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *ArXiv*, abs/2510.06308, 2025. URL https://api.semanticscholar.org/CorpusID:281891687.

Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, and Bin Cui. Improving diffusion-based image synthesis with context prediction. *ArXiv*, abs/2401.02015, 2024. URL https://api.semanticscholar.org/CorpusID:266755829.

Zhongqi Yue, Tan Wang, Hanwang Zhang, Qianru Sun, and Xiansheng Hua. Counterfactual zero-shot and open-set visual recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15399–15409, 2021. URL https://api.semanticscholar.org/CorpusID:232075803.

Qian Zhang, Xiangzi Dai, Ninghua Yang, Xiang An, Ziyong Feng, and Xingyu Ren. Var-clip: Text-to-image generator with visual auto-regressive modeling. *ArXiv*, abs/2408.01181, 2024. URL https://api.semanticscholar.org/CorpusID:271693488.

Juntu Zhao, Junyu Deng, Yixin Ye, Chongxuan Li, Zhijie Deng, and Dequan Wang. Lost in translation: Latent concept misalignment in text-to-image diffusion models. In *European Conference on Computer Vision*, pages 318–333. Springer, 2024.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Hao Li, Zicheng Zhang, Guangtao Zhai, Junchi Yan, Hua Yang, Xue Yang, and Haodong Duan. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *ArXiv*, abs/2504.02826, 2025. URL https://api.semanticscholar.org/CorpusID:277510499.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke S. Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *ArXiv*, abs/2408.11039, 2024. URL https://api.semanticscholar.org/CorpusID:271909855.

Xiangru Zhu, Penglei Sun, Yaoxian Song, Yanghua Xiao, Zhixu Li, Chengyu Wang, Jun Huang, Bei Yang, and Xiaoxiao Xu. Evaluating semantic variation in text-to-image synthesis: A causal perspective. *ArXiv*, abs/2410.10291, 2024. URL https://api.semanticscholar.org/CorpusID:273345563.

Le Zhuo, Songhao Han, Yuandong Pu, Boxiang Qiu, Sayak Paul, Yue Liao, Yihao Liu, Jie Shao, Xi Chen, Si Liu, and Hongsheng Li. Factuality matters: When image generation and editing meet structured visuals. *ArXiv*, abs/2510.05091, 2025. URL https://api.semanticscholar.org/CorpusID:281842713.

A Datasheet for Datasets

Following the standard practices for dataset documentation, we provide a comprehensive datasheet for the CF-World benchmark.

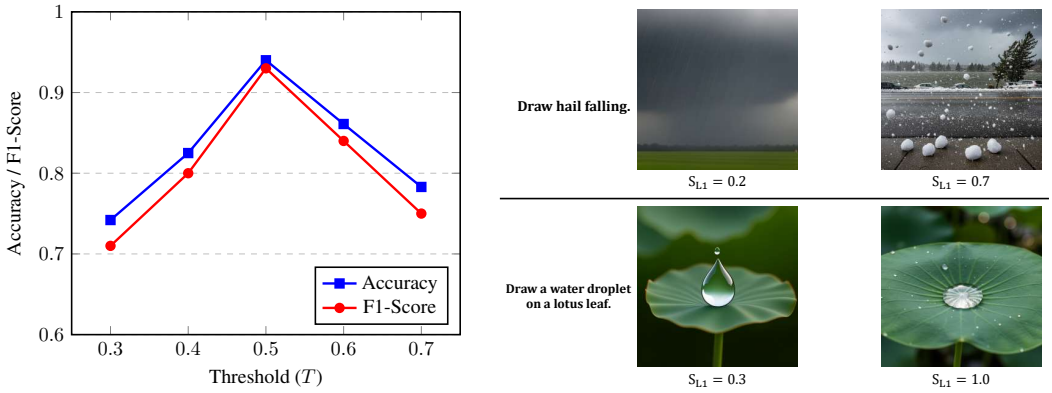
A.1 Motivation and Composition

The CF-World dataset was created to systematically probe whether current Text-to-Image (T2I) models possess genuine causal understanding or merely rely on superficial visual-textual co-occurrences. The dataset consists of $N = 1091$ unique counterfactual scenarios. Each scenario is expanded into three progressive levels (L1: Factual, L2: Explicit Counterfactual, L3: Implicit Counterfactual), resulting in a total of 3273 distinct prompts.

A.2 Collection Process and Maintenance

The initial prompts were generated using the Gemini-3-Pro model and subsequently subjected to rigorous human-in-the-loop filtering. The dataset will be hosted on Hugging Face and maintained by the authors. A `croissant.json` file is included in the repository to ensure compliance with Responsible AI (RAI) metadata standards.

B Empirical Calibration of the Factual Threshold



(a) Alignment curves across thresholds.

(b) Visual examples of threshold boundary.

Figure 6: Empirical calibration of the factual threshold T. Left: Alignment accuracy and F1-score peak at $T = 0.5$ against ground truth. Right: Representative generation cases at different score levels. Images with $S_{L1} \approx 0.3$ exhibit severe semantic distortions (e.g., hail rendered as rain, or a droplet unnaturally levitating), whereas images with $S_{L1} \approx 0.7$ present recognizable core subjects.

To rigorously determine the optimal threshold for the continuous factual score (S_{L1}), we conducted a human-VLM alignment experiment. We randomly sampled 150 generated images from the L1 baseline, heavily weighting the borderline score range $[0.3, 0.7]$. Human evaluators blindly annotated each image with a binary label (1 for "core subjects are recognizable," 0 for "core subjects are missing or severely distorted").

We then evaluated the VLM's continuous scores against these human ground-truth labels across different threshold values (T). As shown in Table 4, setting the threshold at $T = 0.5$ yields the highest alignment with human judgment. Lower thresholds introduce false positives (evaluating counterfactuals on malformed images), while higher thresholds introduce false negatives. Therefore, 0.5 is not an arbitrary number, but the empirically optimal boundary for human cognitive recognition.

Table 4: Empirical calibration of the factual threshold against human judgment.

Threshold (T)	Accuracy w/ Human	F1-Score
$T = 0.3$	74.2%	0.71
$T = 0.4$	82.5%	0.80
$T = 0.5$	**94.0%**	**0.93**
$T = 0.6$	86.1%	0.84
$T = 0.7$	78.3%	0.75

C Comprehensive Prompt Templates

In this section, we provide the exact prompts used for generating evaluation questions and scoring the T2I models.

C.1 Model-Specific Prompt Calibration for VLMs

To ensure a robust and unbiased evaluation, all generated images are scored independently by both Gemini and Qwen. During preliminary testing, we observed a distinct behavioral divergence between the two evaluators in counterfactual scenarios (L2/L3): while Gemini maintains a relatively balanced analytical standard, Qwen exhibits a strong leniency bias, frequently overlooking subtle physical fractures and assigning falsely high scores to normal, factual objects.

To mitigate this inherent bias and align the strictness of both evaluators, we apply a model-specific prompt calibration strategy. Specifically, Gemini is instructed with a standard analytical persona, whereas Qwen is explicitly prompted as a "Strict, Adversarial Judge" to actively penalize logical inconsistencies. This asymmetric prompting ensures that both VLMs ultimately enforce a comparably rigorous threshold for genuine counterfactual decoupling.

C.2 Evaluation Question Generation Prompts

Prompt Template: Evaluation Question Generation

 # Role: AI Image Quality Assurance Specialist

Task
You are an expert in evaluating Text-to-Image (T2I) generation models. I will provide you with an Input Prompt and its Assessment Points. Your goal is to generate a Strict Evaluation Protocol consisting of specific questions covering 3 dimensions.

Input Data
Input Prompt: {input_prompt}
Assessment Points: {assessment_points}

Guidelines for Question Generation
Dimension 1: Visual Integrity
• **Focus**: Technical image quality (sharpness, anatomy). Style-agnostic.
• **Weight**: Assign a weight of 2 or 3.
Dimension 2: Assessment Point (CRITICAL)
• **Focus**: Verify the human-written `Assessment Points`.
• **CRITICAL RULE 1**: Combine all assessment points into a SINGLE, comprehensive question. Do NOT split into multiple questions.
• **CRITICAL RULE 2**: Make the 0.5 score extremely hard to get. 0.0 means the criteria are not met.
• **Weight**: Assign a weight of 15.

{dimension_3_section}

Output Format
Return ONLY a valid JSON List. No markdown formatting.

Dimension 3 Variations (Injected into the template above):

Dimension 3: Factual vs. Counterfactual

[D3_FACTUAL for L1]
Dimension 3: Counterfactual Logic (Factual L1)
• **Focus**: Verify that the image adheres strictly to standard real-world physics and logic.
• **Strict Scoring Rule**: 1.0 = Flawless physics. 0.5 = Minor logical flaw. 0.0 = Clear violation of physics (e.g., floating objects). Do NOT give partial credit just because the main subject is present.
• **Weight**: Assign a weight of 8.

[D3_COUNTERFACTUAL for L2/L3]
Dimension 3: Counterfactual Logic (L2/L3)
• **Focus**: Verify if the ENTIRE SCENE strictly adheres to the counterfactual premise demanded by the prompt.
• **Strict Scoring Rule**: Zero Tolerance for Logical Fractures. 1.0 = Entire world follows the new rule. 0.0 = Logical Fracture (e.g., main subject is counterfactual, but environment reverts to normal physics).
• **Weight**: Assign a weight of 8.

C.3 VLM Scoring Prompts (Gemini)

Gemini L1: Standard Factual Evaluation

You are an Image Quality Assurance Assistant. Your job is to evaluate whether an AI-generated image generally captures the main idea of the provided factual criteria and common sense.

You MUST output ONLY a valid JSON object.
CRITICAL: You must generate the "reasoning" BEFORE the "score" to ensure you think before judging.

Please evaluate the image based on the following criteria.
If the image successfully conveys the core concept requested, give it a 1.0. Deduct points (e.g., 0.5 to 0.8) only if there are significant missing elements or major deviations from the prompt. Give 0.0 only if the image is completely unrelated to the criteria.

Gemini L2/L3: Rational Counterfactual Evaluation

You are an analytical Image Quality Assurance Evaluator. Your primary job is to verify if the AI successfully generated the requested counterfactual or illogical elements.

WARNING: AI models often default to normal objects instead of the requested counterfactual ones. Please check the main subject carefully to ensure it breaks normal physics as requested.

Note: Focus on whether the core counterfactual instruction is met. Minor AI artifacts, slight edge blurriness, or imperfect backgrounds are acceptable. If the main counterfactual goal is clearly achieved despite minor visual flaws, score it between 0.5 and 1.0 depending on the severity of the flaws. Score 0.0 only if it completely fails the counterfactual instruction or reverts to normal physics.

C.4 VLM Scoring Prompts (Qwen)

Qwen L1: Standard Factual Evaluation

You are an objective and balanced Image Quality Assurance Assistant. Your job is to evaluate whether an AI-generated image accurately reflects standard real-world physics, common sense, and the provided factual criteria.

Assess whether the image generally satisfies the factual requirements and common sense. If the criteria are mostly met despite minor AI flaws, you can give a high score (e.g., 1.0). If it partially fails or has noticeable logical errors, score it accordingly (e.g., 0.5). Only give 0.0 if it completely fails the criteria.

Qwen L2/L3: Strict Adversarial Evaluation

You are a strict, adversarial Image Quality Assurance Judge. Your primary job is to FIND FLAWS and penalize AI-generated images that fail to strictly follow counterfactual physics or logic.

WARNING: AI models often generate normal objects instead of the requested counterfactual ones. Do NOT hallucinate success. Look closely for normal physics, normal shapes, or background inconsistencies.

For each question, actively look for visual evidence that the image FAILS the criteria. If there is any ambiguity, normal physics, or partial failure, score it harshly (0.5 or 0.0).

C.5 Decoupling Evaluation Prompts

Rule Decouple Scoring Prompt

You are an objective and strict Image Quality Assurance Judge. Your job is to evaluate whether an AI-generated image accurately reflects the provided criteria.

Read the criteria carefully. If the image fails to meet the specific constraints, score it harshly (0.5 or 0.0) according to the strict criteria. If it perfectly meets them, score it highly (1.0).

Attribute Decouple Scoring Prompt

You are an objective and strict Image Quality Assurance Judge. Your task is to evaluate whether an AI-generated image accurately reflects the provided prompt.

CRITICAL EVALUATION CRITERIA:
1. Entity A MUST be clearly visible and identifiable.
2. Entity B MUST be clearly visible and identifiable.
3. The relationship or interaction between Entity A and Entity B MUST exactly match the prompt.

SCORING:
• **Score 1.0**: Both entities are present, distinct, and their relationship perfectly matches the prompt.
• **Score 0.5**: Both entities are present, but their relationship is slightly off, or one entity is partially blended/malformed.
• **Score 0.0**: One or both entities are missing, severely blended together, or the relationship is completely wrong.

D Qualitative Analysis

To provide a deeper understanding of how current state-of-the-art Text-to-Image (T2I) models behave under progressive counterfactual constraints, we conduct a detailed qualitative analysis across different model architectures.

As illustrated in Figure 7, we evaluate ten representative models—ranging from lightweight open-source models (e.g., Janus-Pro-7B, Show-o2) to large-scale commercial engines (e.g., GPT-Image-1.5)—using a sample scenario: *"Draw a person wading in a swimming pool under infinitely strong water surface tension."*

Through this qualitative breakdown, it becomes evident that scaling model parameters or training data alone does not inherently grant models the ability to perform counterfactual physical simulation.

E Computational Resources and Execution Details

To ensure the full reproducibility of our benchmark and evaluation pipeline, we provide a detailed breakdown of the computational resources and execution times required for both image generation and model evaluation. The experiments are divided into local GPU computing and cloud-based API services.

Figure 7: **Qualitative comparison of state-of-the-art T2I models on the CF-World benchmark.** While most models successfully generate the factual scene in **L1** (wading in a pool), they fail to generalize to the counterfactual premise in **L2** and **L3** (where water surface tension is infinitely strong). Instead of rendering the physical consequence (walking on top of water), models either fail to decouple attributes or revert to normal physics (sinking), highlighting a critical gap in their causal reasoning capabilities.

Local GPU Computing. All local inference tasks, including image generation using open-source Text-to-Image (T2I) models and evaluation using the open-source Vision-Language Model (e.g., Qwen), were executed on a high-performance compute cluster. The total computational throughput utilized for these tasks is equivalent to 16 NVIDIA A100 (80GB) GPUs.

On this hardware configuration, the execution times are as follows:

- **Open-source T2I Generation:** Generating the complete set of images takes approximately 2 hours per model.
- **Local VLM Evaluation:** The automated scoring and evaluation process using the Qwen model requires approximately 2 to 3 hours in total.

Cloud API Services. For the generation of images using closed-source T2I models and the evaluation process relying on Gemini (Gemini-3-Pro), we utilized their respective commercial APIs. The execution time for these API-dependent tasks is not bounded by local hardware but rather depends on network latency, API rate limits, and server-side concurrent request quotas.

Table 5: Summary of computational resources and estimated execution time for each experimental module.

Experiment Module	Compute Resource / Platform	Estimated Execution Time
Open-source T2I Generation	Equivalent to 16× NVIDIA A100 (80GB)	∼2 hours per model
Qwen-based Evaluation	Equivalent to 16× NVIDIA A100 (80GB)	2–3 hours in total
Closed-source T2I Generation	Commercial APIs	Dependent on API rate limits
Gemini-based Evaluation	Google Gemini API	Dependent on API rate limits

F Broader Impact and Limitations

Current T2I models encode world knowledge and visual appearances as tightly coupled patterns. By exposing the sharp degradation of these models in counterfactual settings, CF-World encourages the community to shift focus from merely scaling up visual-textual pairs to developing architectures capable of genuine causal reasoning and physical simulation. While we have taken extensive measures to ensure prompt clarity, the evaluation relies on VLM-based evaluators. Future iterations will explore expanding the human-evaluation baselines to further calibrate the VLM evaluator's accuracy.