Science-T2I: Addressing Scientific Illusions in Image Synthesis

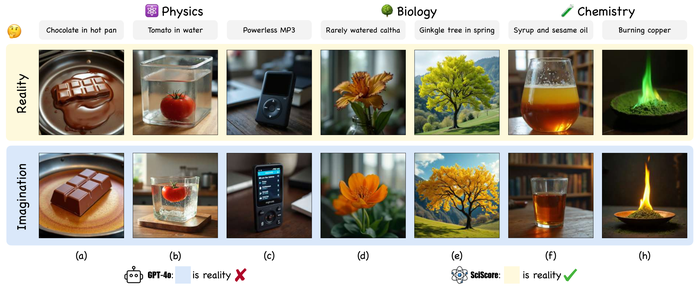
Jialuo Li [1] Wenhao Chai [2] Xingyu Fu [3] Haiyang Xu [4] Saining Xie[1]

[1]New York University [2]University of Washington [3]University of Pennsylvania [4]University of California San Diego

Motivation

Can Text-to-Image models generate scientifically accurate images?

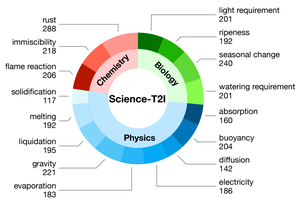
No! Modern generative models often produce scientifically implausible results—red unripe apples, or copper burning with a yellow flame. Moreover, multimodal models like GPT-4o often struggle to reliably differentiate between realistic depictions and physically implausible ones.

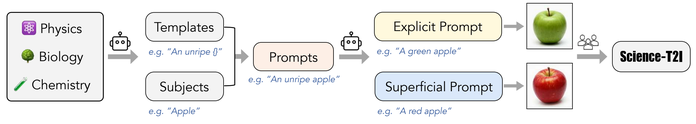


Dataset: SCIENCE-T2I

We introduce Science-T2I, a novel scientific dataset consisting of 20,000 adversarial image pairs and 9,000 prompts. The prompts are categorized into three main types:

- **Implicit Prompt:** Contains terms or phrases that suggest specific visual phenomena, requiring scientific reasoning to interpret.
- **Explicit Prompt:** A clear, descriptive reformulation of the implicit prompt that accurately conveys the intended image.
- **Superficial Prompt:** Offers an explicit description but lacks scientific depth, focusing on surface-level or simplistic interpretations.





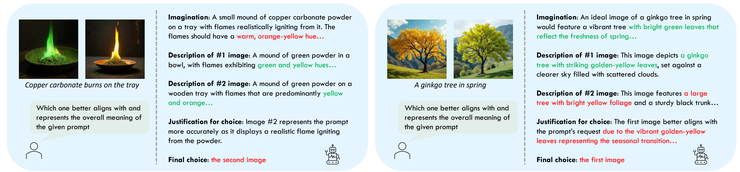
Benchmark	Type	Category	Training Set	Evaluation	
				Generation	LMM
Commonsense-T2I	Commonsense	5	✘	✔	✘
T2I-FactualBench	Commonsense	8	✘	✔	✘
PhyBench	Science	31	✘	✔	✘
Science-T2I (Ours)	Science	16	✔	✔	✔

Method: SCISCORE

Using Science-T2I, we introduce SciScore, a CLIP-based reward model that evaluates the scientific authenticity of generated images. SciScore assigns a continuous score reflecting the scientific plausibility of an image, based on prompt with implicit scientific concepts.

We also create two benchmarks from the Science-T2I Dataset, Science-T2I-S and Science-T2I-C, to evaluate how well LMMs and VLMs can distinguish real from fake scientific images.

Model	Science-T2I S				Science-T2I C			
	Physics	Chemistry	Biology	Avg.	Physics	Chemistry	Biology	Avg.
Human Eval	87.67	75.85	95.29	87.01	84.71	85.40	89.14	86.02
Random Guess	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
CLIP-H	55.08	52.38	55.88	54.69	56.56	44.44	76.67	59.47
BLIPScore	50.35	43.08	59.86	55.00	49.78	60.00	58.33	51.54
SigLIP ViT-SO-14	59.63	53.17	55.94	57.23	61.48	51.11	70.00	61.67
Qwen2-VL-7B	60.03	67.01	68.82	63.79	66.80	50.00	90.83	69.82
LLaVA-OV-7B	68.22	57.82	64.71	65.05	74.59	50.00	76.67	70.26
InternVL2.5-8B	67.80	62.24	84.41	70.79	73.77	65.56	85.83	75.33
GPT-4o mini	61.97	73.81	86.76	70.83	69.29	70.00	90.00	74.78
GPT-4o mini+ CoT	67.04	76.87	90.00	74.97	72.44	70.00	92.50	77.16
SciScore (ours)	94.92	80.95	100.0	93.14	86.89	91.11	100.0	91.19



T2I Model	Science-T2I S				Science-T2I C			
	SP	EP	IP	ND	SP	EP	IP	ND
Stable Diffusion v1.5	19.35	26.88	22.37	40.11	22.45	28.19	23.40	16.55
Stable Diffusion XL	21.80	31.90	25.47	36.34	26.21	34.22	30.89	58.43
Stable Diffusion 3	18.99	32.53	22.31	24.52	24.01	34.65	27.88	36.37
FLUX.1[schnell]	18.45	32.87	24.43	41.47	25.12	36.05	29.66	41.54
FLUX.1[dev]	17.69	32.85	23.56	38.72	23.78	34.70	27.26	31.87

Method: Two-Stage Training Framework

To overcome the limitation of current generative models in producing scientifically accurate images, we propose a two-stage fine-tuning framework. Our approach, using FLUX as the base model, combines supervised fine-tuning with subsequent online fine-tuning.

For supervised fine-tuning, we use Science-T2I with the following training objective:

$$\mathcal{L}_{SFT} = \mathbb{E}_{t,p_t(z|\epsilon),p(\epsilon)} \|v_\theta(z,t) - u_t(z|\epsilon)\|_2^2 \tag{1}$$

For online fine-tuning, specifically, we model the denoising process in the diffusion model as a multi-step Markov Decision Process (MDP):

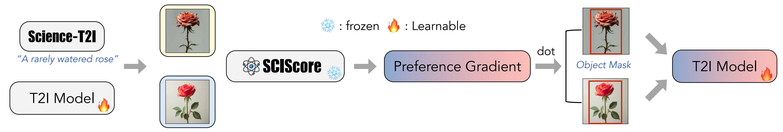
$$s_t \triangleq (c, t, x_{1-t}),\ \pi_\theta(a_t \mid s_t) \triangleq p_\theta(x_{1-\Delta t-t} \mid c, t, x_{1-t}),\ \rho_0(s_0) \triangleq (p(c), \delta_0, \mathcal{N}(0, I)) \tag{2}$$

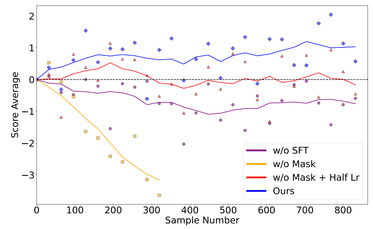
$$a_t \triangleq x_{1-\Delta t-t},\ P(s_{t+\Delta t} \mid s_t, a_t) \triangleq (\delta_c, \delta_{t+\Delta t}, \delta_{x_{1-t-\Delta t}}),\ r(s_t, a_t) \triangleq \begin{cases} r(x_0, c) & \text{if } t = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

We then adopt the DPO objective with a masking strategy, where the mask is denoted by $\mathcal{M}$:

$$\mathcal{L}_{OFT} = -\mathbb{E}\left[\log \rho\left(\beta \log \frac{\mathcal{M}^w \odot \pi_\theta(a_k^w \mid s_k^w)}{\mathcal{M}^w \odot \pi_{\text{ref}}(a_k^w \mid s_k^w)} - \beta \log \frac{\mathcal{M}^l \odot \pi_\theta(a_k^l \mid s_k^l)}{\mathcal{M}^l \odot \pi_{\text{ref}}(a_k^l \mid s_k^l)}\right)\right]$$

Here, w and l denote the preferred and less preferred trajectories, respectively.





Method	Science-T2I S		Science-T2I C	
	SciScore	RI	SciScore	RI
FLUX.1[dev]	23.56	/	27.26	/
+EP	32.85	/	34.70	/
+SFT	27.43	41.66	29.49	29.97
+SFT+OFT	28.52	53.39	30.11	38.31