SciBench: Addressing Scientific Illusions in Image Synthesis

Anonymous CVPR submission

Paper ID xxxx

Abstract

This paper presents a novel approach to integrating scientific knowledge into generative models, enhancing their realism and consistency in image synthesis. We present SciScore*, an end-to-end reward model that refines the assessment of generated images based on scientific knowledge, which is achieved by augmenting both the scientific comprehension and visual capabilities of pre-trained CLIP model. We also introduce* SciBench*, an expert-annotated adversarial dataset comprising 30k image pairs with 9k prompts, covering wide distinct scientific knowledge categories. Leveraging* SciBench*, we propose a two-stage training framework, comprising a supervised fine-tuning phase and a masked online fine-tuning phase, to incorporate scientific knowledge into existing generative models. Through comprehensive experiments, we demonstrate the effectiveness of our framework in establishing new standards for evaluating the scientific realism of generated content. Specifically,* SciScore *attains performance comparable to human-level, demonstrating a 5% improvement similar to evaluations conducted by experienced human experts. Furthermore, by applying our proposed fine-tuning method to FLUX, we achieve a performance enhancement exceeding 50% based on* SciScore*.*

1. Introduction

The quest to conceptualize the visual world and construct real world simulators has been a longstanding endeavor in the computer vision community [8, 17, 19, 28, 65, 66]. As articulated by [10], “The goal of image synthesis is to create, using the computer, a visual experience that is identical to what a viewer would experience when viewing a real environment.” In alignment with this vision, recent advances in generative modeling have notably improved the performance of image synthesis [47, 51, 54]. While these advancements enable the generation of higher resolution, more aesthetically pleasing images with superior Frechet Inception Distance (FID) scores [1, 5, 47, 63], these models often produce superficial imitations rather than authentic representations of the real visual world [6, 18, 43, 44]. This limitation often arises from an inadequate understanding of the underlying scientific principles of realism, as demonstrated in the lower row of FLUX [1] generated images in Figure 1. Consequently, the images generated tend to mirror imaginative constructs, resulting in a noticeable gap between these creations and the tangible reality we inhabit.

This paper seeks to bridge the gap between imagination and realism in image synthesis by integrating scientific knowledge, a crucial element often overlooked in previous approaches. We introduce SciBench, a comprehensive and expert-annotated dataset comprising over 30k image pairs and 9k prompts that span diverse fields such as physics, chemistry, and biology, and cover 16 unique scientific phenomena. Each data pair is collected in an adversarial setup, consisting of one image that accurately aligns with reality and another that does not, thereby facilitating preference modeling. To ensure quality and accuracy, all data were reviewed by human experts whose assessments were based on their professional expertise and consultation of an extensive knowledge base.

Leveraging SciBench, we further present SciScore, an end-to-end reward model infused with diverse expert-level scientific knowledge, designed to evaluate generated images as a science teacher would. Our results demonstrate that SciScore outperforms complex, prompt-engineering-reliant large multimodal models (LMMs) such as `GPT-4o`. Compared to `GPT-4o`, SciScore excels in capturing fine-grained visual details that LMMs often neglect as illustrated in Figure 1, and functions as a comprehensive end-to-end reward model – eliminating the dependence on language-guided inference processes, which can occasionally fail in LMMs due to hallucinations.

Utilizing SciScore , we introduce a two-stage training methodology to develop an enhanced image synthesis model that conform to the realist with world knowledge. Specifically, we begin with supervised fine-tuning (SFT) on FLUX.1[dev][1] using SciBench. This initial phase is subsequently followed by an additional stage of online fine-tuning, where SciScore functions as the reward model and employs a masking strategy to improve the performance.

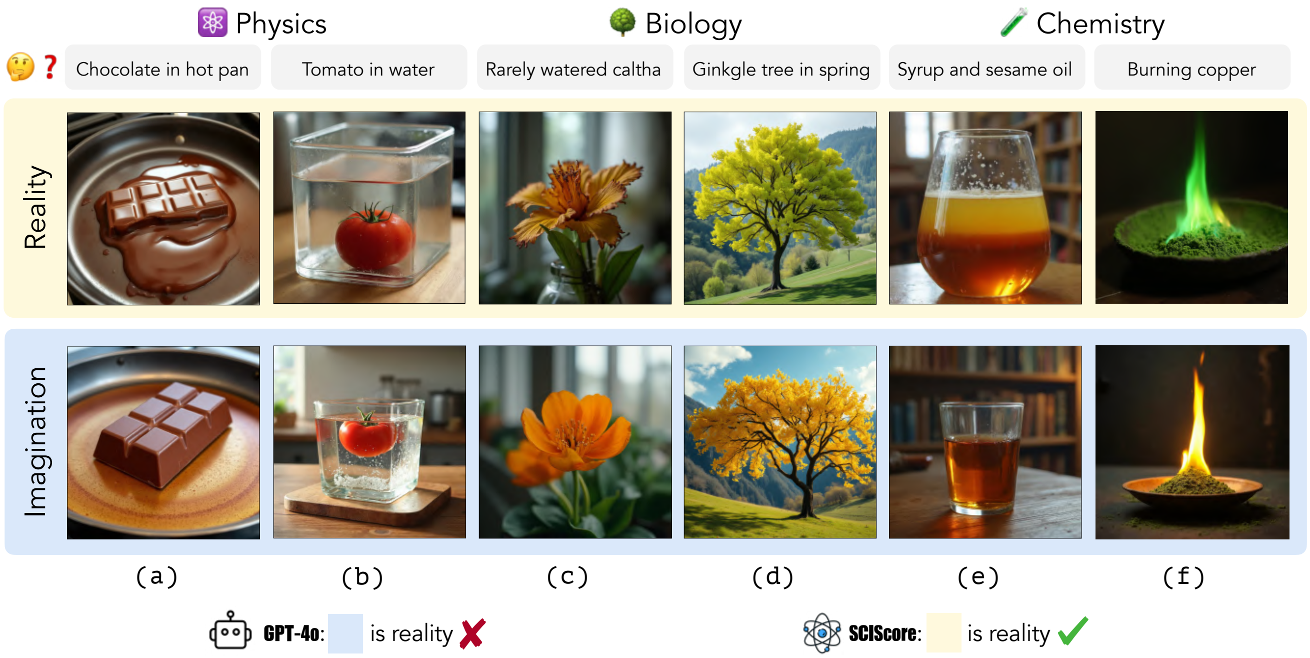


Figure 1. **Comparison between `GPT-4o` and SciScore.** Given a prompt (in grey) requiring scientific knowledge, FLUX [1] model generates imaginary images (lower row) that are far from reality (upper row). LMMs such as GPT-4o [2] fail to distinguish which image aligns better with reality. In contrast, our end-to-end reward model SciScore can successfully do the task. Notice that the prompts here are summarization of the real prompts that we used for illustration purposes.

Our main contributions are summarized as follows:

- We introduce SciBench of over 9,000 prompts and 30,000 image pairs, annotated by experts to reflect reality, enabling the training of a language-guided reward model for text-to-image alignment with scientific knowledge.
- We propose an optimization strategy using the reward model SciScore to enhance diffusion-based generative models, showing improved alignment of generated images with reality on a quantitative benchmark.
- Extensive experiments show that our method outperforms the baseline by over 50%, marking a significant advancement in grounding the model in real-world scenarios.

2. Related Works

2.1. Physics Modeling in Generative Models

Integrating physical laws into generative models has become a vital area of research to enhance the realism and consistency of generated data across various domains, including image synthesis [44], video generation [6, 29, 43], and 3D modeling [21]. PhyBench [44] is a pioneering work that explores the incorporation of physical knowledge into current text-to-image (T2I) models by providing a comprehensive dataset designed to test physical commonsense across various domains. In the realm of text-to-video (T2V) models, benchmarks like VideoPhy [6] and PhyGenBench [43] evaluate whether current generative models can accurately simulate physical commonsense in real-world scenarios involving various material interactions, with PhyGenBench introducing a hierarchical evaluation framework called PhyGenEval. PhysComp [21] introduces a novel approach to single-image 3D reconstruction by decomposing visual geometry into mechanical properties, external forces, and rest-shape geometry, ensuring physical compatibility through static equilibrium constraints. Our work differs by designing tasks as reasoning challenges, requiring models to understand and apply physical laws to generate accurate outputs. This approach pushes the boundaries of physical knowledge integration in generative models by emphasizing implicit reasoning over explicit description.

2.2. RL in Diffusion Models

Reinforcement learning (RL) has been effectively applied in diffusion models to enhance sample quality. For instance, VersaT2I [20] and DreamSync [56] simply use reject sampling. ReNO [14] focus on adapting a diffusion model during inference by purely optimizing the initial latent noise using a differentiable objective. Some other works [59, 64] leverages DPO [50] as optimization strategies. Our work differs by introducing a novel reward function that leverages physical commonsense to guide the diffusion process, ensuring the generated samples are physically plausible.

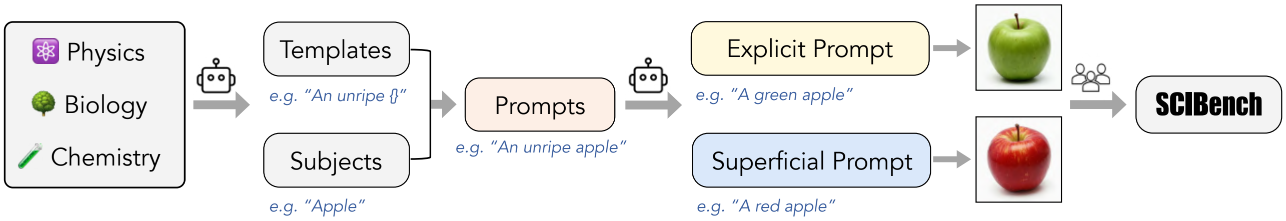


Figure 2. **Data curation pipeline**. For each task, `GPT-4o` [2] first generates structured templates that capture the scientific principles while allowing for variability in objects or substances. These templates are used to create implicit prompts, which `GPT-4o` [2] then expands into explicit and superficial prompts, ultimately guiding the synthesis of corresponding explicit and superficial images.

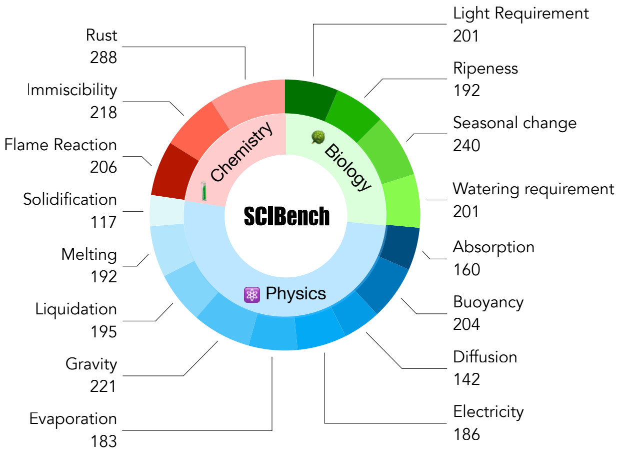


Figure 3. **Data statistics.** SciBench is organized into three primary scientific fields: Chemistry, Biology, and Physics. Each field is divided into specific categories, with the numbers indicating the volume of implicit prompts collected for each category.

2.3. Benchmarking Image Synthesis Models

Standard metrics like FID [23], IS [52], LPIPS [67], and CLIPScore [22] are commonly used to assess image synthesis models. With model advancements, newer methods emphasize human evaluation and multimodal LLM-based assessment. HPSv2 [61], PickScore [32] and ImageReward [62] provide human preference annotations, while VQAScore [35], TIFA [25], VIEScore [34], LLMscore [42], and DSG [9] utilize VQA-style evaluations. For object attributes and relationships, benchmarks like T2I-CompBench [26] and CLIP-R-Precision [46] have been introduced. However, there are few benchmarks focusing on the physical commonsense. PhyBench [44] establishes a set of grading criteria and employs vision-language models to discretely score images. In contrast, we introduce PhyScore, an end-to-end model designed to provide a more refined and continuous scoring mechanism for images.

3. Dataset: SciBench

We introduce SciBench, a novel dataset specifically designed to enhance text-to-image and multimodal models' understanding of underlying scientific principles. Unlike conventional datasets that focus on direct textual descriptions [11, 33, 38] and preference annotation [32, 61, 62], SciBench challenges models to perform implicit reasoning based on prompts that require scientific knowledge.

SciBench consists of 16 tasks that require the model to infer or visualize concepts not explicitly stated in the prompts but derived from underlying scientific principles. These tasks are inspired by existing research such as PhyBench [44] and Commonsense-T2I [18], as well as new concepts developed for this study. Each task is meticulously designed with the following objectives:

- **Rewriting Capability.** Tasks use prompts that allow flexible rephrasing, thereby enabling different expressions to effectively achieve the same visual meaning.
- **Scientific Knowledge Integration.** Tasks are based on established scientific principles in physics, chemistry, and biology, providing a clear and consistent framework. This approach reduces the ambiguity of commonsense knowledge, which can vary culturally or contextually. Examples include gravity, immiscibility, and flame reactions, where scientific laws offer a reliable reference.

We classify prompts into two types: those requiring inference from scientific knowledge and their rewritten versions that utilize rewriting capabilities. Additionally, PhyBench [44] reveals that models often ignore these principles, focusing instead on descriptive text, indicating a third category based on description rather than inference. To clarify these concepts, we introduce specific terminologies:

- **Implicit Prompt (IP).** It contains specific terms or phrases that imply certain visual characteristics or phenomena requiring interpretative reasoning based on scientific knowledge. For example, the prompt ”an unripe apple” suggests greenness without explicitly stating it.
- **Explicit Prompt (EP).** It reformulates the implicit prompt into a clear, descriptive statement that accurately reflects the intended image. For instance, the prompt ”a green apple” directly conveys the immaturity.
- **Superficial Prompt (SP).** It provides an explicit interpretation but neglects scientific reasoning, focusing only on surface descriptions and simplistic interpretations. For

example, interpreting ”an unripe apple” as ”a red apple” overlooks the implied maturity, leading to inaccuracies.

4. Method: SciScore

While CLIP [48] effectively aligns textual and visual data, it struggles to accurately match implicit prompts with their corresponding images. To address this limitation, we introduce SciScore, a reward model fine-tuned on SciBench that extends CLIP's architecture [48]. SciScore assesses the extent to which an image embodies the visual information derived from the scientific principles articulated within the prompt. In this section, we first define the reward mechanism for evaluating prompt-image compatibility (§4.1) and then detail the training methods used to optimize SciScore's performance (§4.2).

4.1. Reward Modeling

SciScore extends the CLIP architecture [48] by independently encoding a text prompt x and an image y into a shared high-dimensional vector space using separate transformer encoders [57], E_{txt} and E_{img}. The reward is computed based on the alignment between textual and visual modalities, quantified by the inner product of their respective encoded representations and subsequently scaled by a learnable temperature parameter T:

$$r(x, y) = T \cdot \frac{E_{\text{txt}}(x) \cdot E_{\text{img}}(y)}{\|E_{\text{txt}}(x)\| \, \|E_{\text{img}}(y)\|}. \tag{1}$$

4.2. Training Techniques

For developing SciScore we employed a fine-tuning approach on the CLIP [48] using SciBench. Each training instance is structured as a tuple $(x_i, x_e, x_s, y_e, y_s)$, where x_i is the implicit prompt, x_e and x_s are the explicit and superficial prompts, respectively. Correspondingly, y_e and y_s denote the explicit and superficial images.

Predicted Preferences Calculation. Following preference modeling approaches in language from prior work [45, 55], the predicted preference $\hat{p}_{\text{img}}(x_i \succ x_j; y)$ for prompt x_i over prompt x_j for a given image y is calculated as:

$$\hat{p}_{\text{img}}(x_i \succ x_j; y) = \frac{\exp(r(y, x_i))}{\exp(r(y, x_j)) + \exp(r(y, x_i))} \tag{2}$$

Similarly, for a given prompt x, the predicted preference $\hat{p}_{\text{txt}}(y_i \succ y_j; x)$ for image y_i over image y_j is given by:

$$\hat{p}_{\text{txt}}(y_i \succ y_j; x) = \frac{\exp(r(y_i, x))}{\exp(r(y_i, x)) + \exp(r(y_j, x))} \tag{3}$$

Implicit Prompt Alignment (IPA). Preliminary experiments revealed that the pretrained CLIP model [48] tends to embed the implicit prompt in a manner similarly to the corresponding superficial prompt. To address this issue, we minimize the KL divergence between the target preference $p_{\text{txt}} = [1, 0]$ and the predicted preference $\hat{p}_{\text{txt}} = [\hat{p}_{\text{txt}}(y_e \succ y_s; x_i), \hat{p}_{\text{txt}}(y_s \succ y_e; x_i)]$. This effectively aligns the implicit prompt with the explicit image over the superficial image. The loss function is defined as:

$$\mathcal{L}_{\text{IPA}} = \sum_{j=1}^{2} p_{\text{txt}_j} \left(\log p_{\text{txt}_j} - \log \hat{p}_{\text{txt}_j}\right) \tag{4}$$

Image Encoder Enhancement (IEE). To effectively handle reasoning tasks that involve fine-grained visual phenomena, it is imperative to enhance the capabilities of the image encoder. The objective of this enhancement is captured by the following loss function:

$$\mathcal{L}_{\text{IEE}} = \mathcal{L}^{+}_{\text{img}} + \mathcal{L}^{-}_{\text{img}}, \tag{5}$$

where $\mathcal{L}^{+}_{\text{img}}$ and $\mathcal{L}^{-}_{\text{img}}$ correspond to the losses associated with explicit and superficial image preferences, respectively. The explicit image loss $\mathcal{L}^{+}_{\text{img}}$ is defined as:

$$\mathcal{L}^{+}_{\text{img}} = \sum_{j=1}^{2} p^{+}_{\text{img}_j} \left(\log p^{+}_{\text{img}_j} - \log \hat{p}^{+}_{\text{img}_j}\right), \tag{6}$$

where $p^{+}_{\text{img}} = [1, 0]$ signifies a preference for the explicit image. The predicted probabilities are denoted by:

$$\hat{p}^{+}_{\text{img}} = [\hat{p}_{\text{img}}(x_e \succ x_s; y_e), \hat{p}_{\text{img}}(x_s \succ x_e; y_e)], \tag{7}$$

Similarly, the superficial image loss $\mathcal{L}^{-}_{\text{img}}$ is defined as:

$$\mathcal{L}^{-}_{\text{img}} = \sum_{j=1}^{2} p^{-}_{\text{img}_j} \left(\log p^{-}_{\text{img}_j} - \log \hat{p}^{-}_{\text{img}_j}\right), \tag{8}$$

where $p^{-}_{\text{img}} = [0, 1]$ indicates a preference for the superficial image. The predicted probabilities are given by:

$$\hat{p}^{-}_{\text{img}} = [\hat{p}_{\text{img}}(x_e \succ x_s; y_s), \hat{p}_{\text{img}}(x_s \succ x_e; y_s)], \tag{9}$$

The overall loss function integrates $\mathcal{L}_{\text{IPA}}$ with $\mathcal{L}_{\text{IEE}}$ as:

$$\mathcal{L} = \mathcal{L}_{\text{IPA}} + \lambda \mathcal{L}_{\text{IEE}}, \tag{10}$$

where λ is a hyperparameter that controls the relative weight of the image encoder enhancement loss in relation to the implicit prompt alignment loss.

5. T2I Model Fine-Tuning

5.1. Supervised Fine-tuning (SFT)

Current post-training algorithms for diffusion models, such as those utilizing PPO [7, 15] and DPO [58, 63], have significantly advanced model fine-tuning. However, these methods are constrained by the requirement that the optimization

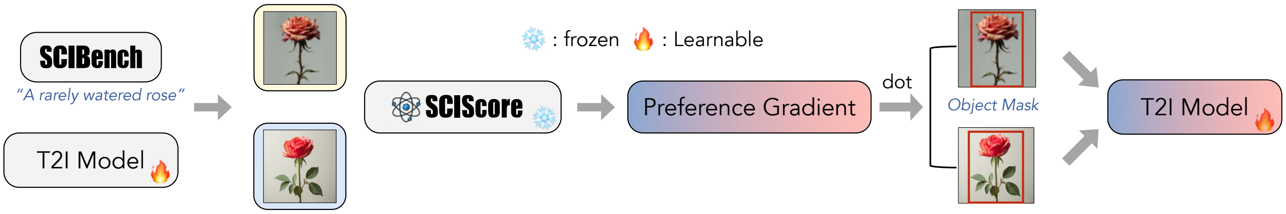


Figure 4. **Online fine-tuning Pipeline.** For each prompt, two images are generated to compute SCISCORE preference metric. Simultaneously, GroundingDINO [40] extracts segmentation masks from these images based on the prompts, which are then used to block gradient propagation in the corresponding spatial regions.

objectives remain within the distribution of the pre-trained model. While this limitation is acceptable for tasks like aesthetic enhancement, which involve preferences among generated images, it poses challenges for applications requiring scientific reasoning. Preliminary experiments demonstrate that pre-trained models lack an understanding of scientific principles, as they are primarily trained on descriptive prompts paired with images. This shortcoming presents a significant obstacle for post-training techniques aimed at embedding scientific comprehension into diffusion models.

Our methodology begins with the supervised fine-tuning of a pre-trained model to enhance its scientific understanding, utilizing the SCIBENCH. As illustrated by the experimental results in Table 3, FLUX [1] models consistently achieve superior performance in direct text-image alignment and exhibit a strong capacity for generating realistic styles, as evidenced by our preliminary experiments. Based on these observations, we adopt FLUX.1[dev][1] as our base model. Since FLUX [1] employs flow matching [39] framework, the SFT training objective is formulated as:

$$L_{SFT} = \mathbb{E}_{t,p_t(z|\epsilon),p(\epsilon)} \left\| v_\theta(z,t) - u_t(z|\epsilon) \right\|_2^2, \qquad (11)$$

In this formulation, we adopt the same mathematical notation as presented in [13] to ensure consistency.

5.2. Masked Online Fine-tuning

After performing domain transfer using SFT, we apply an online fine-tuning approach for further model refinement with pipeline shown in Figure 4. Following the methodology proposed by DDPO [7], we conceptualize the denoising process within the diffusion model as a multi-step MDP:

$$\begin{aligned}
s_t &\triangleq (c, t, x_{1-t}), \quad a_t \triangleq x_{1-\Delta t-t} \\
P(s_{t+\Delta t} \mid s_t, a_t) &\triangleq (\delta_c, \delta_{t+\Delta t}, \delta_{x_{1-t-\Delta t}}) \\
\pi_\theta(a_t \mid s_t) &\triangleq p_\theta(x_{1-\Delta t-t} \mid c, t, x_{1-t}) \\
\rho_0(s_0) &\triangleq (p(c), \delta_0, \mathcal{N}(0, I)) \\
r(s_t, a_t) &\triangleq \begin{cases} r(x_0, c) & \text{if } t = 1 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

However, flow matching [39] is typically formulated as an Ordinary Differential Equation (ODE), resulting in a deterministic process. This deterministic formulation complicates the computation of the policy $\pi_\theta(a_t \mid s_t)$:

$$\pi_\theta(a_t \mid s_t) = \delta\left(x_{1-\Delta t-t} - (x_{1-t} - v_\theta(s_t)\Delta t)\right) \qquad (12)$$

In alignment with the discussion in [12], we can alternatively interpret flow matching [39] as a Stochastic Differential Equation (SDE), which is mathematically formulated as:

$$\mathrm{d}x_t = \left(v_\theta(x_t, t) + \frac{\sigma_t^2}{2\beta_t \eta_t} \lambda_t \right) \mathrm{d}t + \sigma_t \mathrm{d}B_t \qquad (13)$$

$$\eta_t = \left(\frac{\dot{\alpha}_t}{\alpha_t} \beta_t - \dot{\beta}_t \right), \quad \lambda_t = \left(v(x_t, t) - \frac{\dot{\alpha}_t}{\alpha_t} x_t \right) \qquad (14)$$

where B_t denotes Brownian motion. By discretizing this equation while leveraging the rectified flow employed by FLUX [1], where $\alpha_t = t$ and $\beta_t = 1 - t$, we obtain:

$$\pi_\theta(a_t \mid s_t) = \mathcal{N}\left(a_t; \mu_\theta(s_t), \sigma_t^2 I\right) \qquad (15)$$

$$\mu_\theta(s_t) = \frac{t\sigma_t^2 + 2(1-t)}{-2(1-t)} v_\theta(s_t)\Delta t + \frac{2(1-t) + \sigma_t^2 \Delta t}{2(1-t)} x_{1-t} \qquad (16)$$

In this framework, the parameter σ_t is subject to manual configuration. Notably, setting $\sigma_t = 0$ simplifies the formulation to the deterministic case, as delineated in Equation 12. For the training objective, we adopt DPO as introduced by [49]. Specifically, given a condition (typically a prompt) c, we randomly sample two trajectories:

$$\sigma_w = \{s_0^w, a_0^w, s_{\Delta t}^w, a_{\Delta t}^w, \ldots, s_1^w, a_1^w\} \qquad (17)$$

$$\sigma_l = \{s_0^l, a_0^l, s_{\Delta t}^l, a_{\Delta t}^l, \ldots, s_1^l, a_1^l\} \qquad (18)$$

Assuming that the reward satisfies $r(s_1^w, a_1^w) > r(s_1^l, a_1^l)$, the training objective is formulated as:

$$\mathbb{E}\left[\log \rho \left(\beta \log \frac{\pi_\theta(a_k^l | s_k^l)}{\pi_{\text{ref}}(a_k^l | s_k^l)} - \beta \log \frac{\pi_\theta(a_k^w | s_k^w)}{\pi_{\text{ref}}(a_k^w | s_k^w)} \right) \right] \qquad (19)$$

Table 1. Performance comparison of different models on SciBench S and SciBench C across different subjects. **Bold** values indicate the best performance.

Model	SciBench S				SciBench C			
	Physics	Chemistry	Biology	Avg.	Physics	Chemistry	Biology	Avg.
CLIP-H [27]	55.08	52.38	55.88	54.69	56.56	44.44	76.67	59.47
BLIPScore [36]	50.35	43.08	59.86	55.00	49.78	60.00	58.33	51.54
GPT-4o mini	61.97	73.81	86.76	70.83	69.29	70.00	90.00	74.78
+ CoT [60]	67.04	76.87	90.00	74.97	72.44	70.00	92.50	77.16
Human Eval	87.67	75.85	95.29	87.01	84.71	85.40	89.14	86.02
SciScore (ours)	**94.92**	**80.95**	**100.00**	**93.14**	**86.89**	**91.11**	**100.00**	**91.19**

Subject-Based Masking Strategy. Considering the subject-oriented characteristics inherent to our scientific reasoning tasks, we employed a subject-based masking strategy during training. Specifically, we extract the subject from the input prompt and utilize GroundingDINO [40] to identify the bounding box around the subject. Subsequently, only the content within this bounding box is used for gradient backpropagation. Define mask corresponding to the box as $\mathcal{M}$, then the final training objective:

$$\mathcal{L} = -\mathbb{E}\left[\log\rho\left(\beta\log\frac{\mathcal{M}^w \odot \pi_\theta(a_k^w \mid s_k^w)}{\mathcal{M}^w \odot \pi_{\text{ref}}(a_k^w \mid s_k^w)} - \beta\log\frac{\mathcal{M}^l \odot \pi_\theta(a_k^l \mid s_k^l)}{\mathcal{M}^l \odot \pi_{\text{ref}}(a_k^l \mid s_k^l)}\right)\right]. \tag{20}$$

6. Experiment: Reward Model

6.1. Implementation Details

Training Setting. We fine-tune the CLIP-H model [27] using our framework on SciBench training set with both text and image encoder learnable. The experiment completes within one hour on 8 A6000 GPUs.

Evaluation Setting. To evaluate the model's generalization, we introduce two manually annotated test sets:

- **SciBench S** (671 tuples): It matches the training set style, emphasizing simplicity and reasoning regions.
- **SciBench C** (227 tuples): It adds complexity through diverse scene settings in prompts and images.

We establish our baseline using three evaluation dimensions: VLMs, Large Multimodal Models (LMMs), and human assessments. For VLMs, we utilize CLIP-H [27] and BLIPScore [36, 37]. In the LMM category, we employ `GPT-4o-mini` [2] with CoT reasoning[60]. Human evaluations involved 10 experts with science or engineering degrees. Experimental results are presented in Table 1.

Table 2. Ablation study on different λ used in SciScore. **Bold** values indicate the best performance.

λ	SciBench S	SciBench C
0	**93.14**	88.99
0.1	92.85	90.75
0.5	92.85	**91.19**
0.75	**93.14**	88.99
0.25	**93.14**	**91.19**

6.2. Results

CLIP-H [48] and BLIPScore [36] demonstrate near-random accuracy (approximately 50%) across both test sets, underscoring their limitations in effectively distinguishing images when given implicit prompts. Even `GPT-4o-mini` [2], despite being equipped with a vast knowledge base, fails to deliver satisfactory performance in these tasks. Notably, the application of CoT prompting [60] does not yield significant improvements in this context. In contrast, SciScore not only achieves but surpasses human-level performance on both SciBench S and SciBench C, highlighting its superior generalization and efficacy in handling the tasks in a complex scenario. This result underscores the potential of SciScore to address the challenges inherent in understanding scientific knowledge, where other models struggle.

6.3. Ablation Study

Effect of IEE. To investigate the effect of IEE on the performance of the model, comparative experiments are conducted. As shown in Table 2, there exists a trade-off between increasing the IEE loss rate and maintaining IPA loss. A lower IEE loss rate fails to enhance the image encoder's ability to detect fine-grained details, whereas a higher IEE loss rate diminishes the focus on prompt alignment. We identified $\lambda = 0.25$ as the optimal for these objectives.

Table 3. **Performance of T2I Models on SciScore.** Here normalized difference (ND) represents that ND $= (\mathrm{IP} - \mathrm{SP})/(\mathrm{EP} - \mathrm{SP})$. **Bold** values indicate the best performance, while underlined values represent the second-best performance.

T2I Model	SciBench S				SciBench C			
	SP	EP	IP	ND	SP	EP	IP	ND
Stable Diffusion v1.5 [51]	19.35	26.88	22.37	40.11	22.45	28.19	23.40	16.55
Stable Diffusion XL [47]	21.80	31.90	**25.47**	36.34	26.21	34.22	**30.89**	**58.43**
Stable Diffusion 3 [13]	18.99	32.53	22.31	24.52	24.01	34.65	27.88	36.37
FLUX.1[schnell] [1]	18.45	**32.87**	24.43	**41.47**	25.12	**36.05**	29.66	41.54
FLUX.1[dev] [1]	17.69	32.85	23.56	38.72	23.78	34.70	27.26	31.87

Table 4. **SciScore on Various Methods.** Relative improvement (RI) is defined as the improvement in SciScore divided by the improvement achieved through generation based on explicit prompt. **Bold** values indicate the best performance.

Method	PS		PC	
	SciScore	RI	SciScore	RI
FLUX.1[dev]	24.47	/	27.62	/
+EP	35.70	/	35.03	/
+SFT	29.87	48.09	31.70	55.06
+SFT+OFT	30.52	**53.87**	32.15	**61.13**

6.4. Benchmarking Text-To-Image Generation.

By leveraging the superior performance of SciScore, rather than relying on VLM that require complex prompting techniques and demonstrate comparatively inferior performance, we propose an end-to-end utilization of SciScore for benchmarking current text-to-image models.

Three-Dimensional Evaluation. We assessed the scientific reasoning capabilities of current state-of-the-art text-to-image models through a three-dimensional evaluation. Specifically, we evaluated: the alignment between implicit prompts and (1) images generated from implicit prompts, (2) images generated from explicit prompts, and (3) images generated from superficial prompts. For each alignment evaluation, we selected one implicit, explicit, and superficial prompt forming a tuple from the SciBench S and SciBench C, respectively. We generated two images per prompt using the text-to-image models and calculated the average SciScore. The results are in Table 3.

Analysis: Explicit Prompt Alignment. The experiment results in Table 3 reveals that the FLUX series models [1] consistently outperform the Stable Diffusion series on explicit prompt alignment. In particular, SDv1.5 [51] exhibits a significant performance gap when compared to the other models in the study. Further detailed analysis and discussion can be found in the appendix.

Analysis: Reasoning Capability. Based on the data presented in Table 3, it is evident that current text-to-image models demonstrate notable limitations in interpreting implicit meanings within prompts. These models are more likely to generate images that align with the literal aspects of the prompts, rather than inferring or representing deeper, implicit meanings. This limitation is reflected in the models' normalized difference (ND) scores, where the majority fall below 50, with an average around 35.

7. Experiment: T2I Model Fine-Tuning

7.1. Implementation Details

Training Setting. We first fine-tune FLUX.1[dev] [1] on SciBench using SFT in conjunction with LoRA [24] for 2,000 steps. This process generates LoRA weights intended for subsequent online fine-tuning. For the online fine-tuning phase, we randomly select 300 implicit prompts from SciBench to serve as the training set. During each epoch, 32 prompts are sampled, with each prompt paired with two images, and their corresponding SciScore is computed. Subject masks are extracted from the images using GroundingDINO [40]. The model is subsequently fine-tuned for approximately 100 steps.

Evaluation Setting. We design two prompt sets to assess the generative model: one emphasizing simplicity consists of 100 implicit prompts selected from SciBench S, and the other incorporating environmental settings including 100 implicit prompts sourced from both SciBench C and the training set, enriched with additional environmental contexts. For evaluation, we generate five images per prompt and calculate their average SciScore.

7.2. Results

The results in Table 4 demonstrate that both SFT and online fine-tuning enhance SciScore's performance. To further investigate the factors driving these enhancements, we employed explicit prompts corresponding to all implicit prompts in PS and PC. This approach allowed us to calcu-

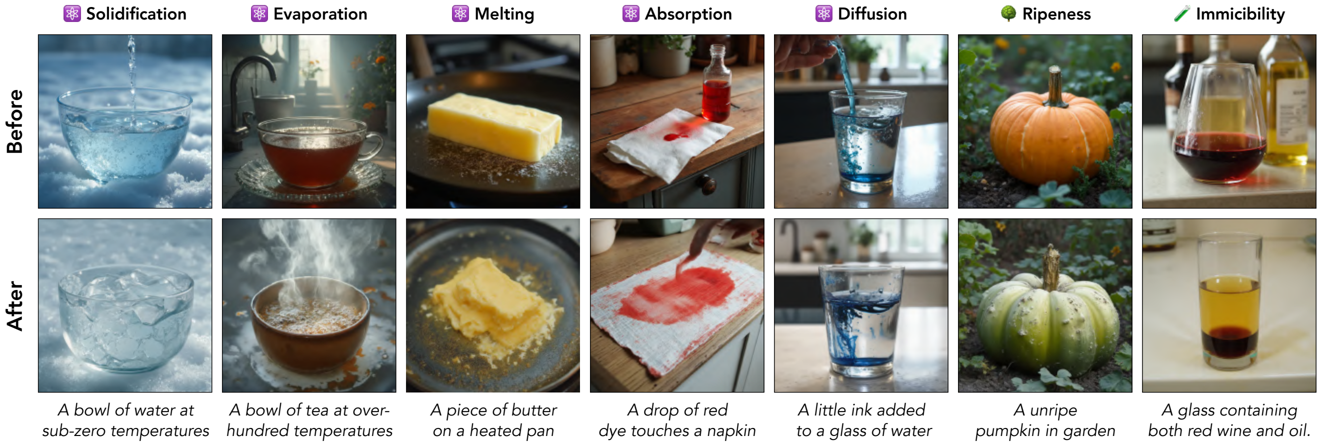

Solidification
Evaporation
Melting
Absorption
Diffusion
Ripeness
Immicibility
Before
After
A bowl of water at sub-zero temperatures
A bowl of tea at over-hundred temperatures
A piece of butter on a heated pan
A drop of red dye touches a napkin
A little ink added to a glass of water
A unripe pumpkin in garden
A glass containing both red wine and oil.

Figure 5. **Case study.** The upper images are generated using the baseline FLUX.1[dev] [1], whereas the lower images are produced with our fine-tuning method. Each image pair utilizes an identical random seed to ensure consistency in comparison. Note that the displayed prompts are summaries of the original prompts used for illustration purposes.

late the average performance of SciScore, which serves as an upper bound for our method. The findings reveal that our proposed technique achieves an impressive performance increase, surpassing the baseline by over 50%. Comparative examples are provided in Figure 5.

7.3. Ablation Study

Necessity of SFT. In Figure 6, the blue line shows SFT performed before online fine-tuning, while the purple line illustrates the case without initial SFT. Both scenarios use identical configurations for online fine-tuning. The results demonstrate that initiating online fine-tuning with SFT leads to a more stable increase in SciScore. In contrast, online fine-tuning without preceding SFT does not improve SciScore. This discrepancy is likely due to the model's limited ability to effectively learn from two suboptimal samples when SFT is not first applied. These observations highlight the critical role of starting with SFT to ensure the model trains within the distribution defined by the objective, facilitating effective online fine-tuning.

Masking Strategy As A Denoiser. Starting from the checkpoint obtained by SFT, we conducted two additional experiments to evaluate the masking strategy's effect on model performance. The results revealed that SciScore curve for the model without the masking strategy was unstable, and the generated images showed signs of collapse. To further explore this issue, we halved the learning rate in an attempt to stabilize training. While this adjustment prevented the collapse of the generated images, it did not lead to an increase of SciScore . This observation suggests that, without the masking strategy, the model tends to indiscriminately consider all features from the preferred images as equally important, effectively treating all features as 'preferred'. However, only the visual features pertinent to the scientific principles contained in the prompt are truly relevant. This indiscriminate preference introduces substantial noise into the training process, hindering the model's ability to learn effectively. In contrast, the model employing the masking strategy demonstrated a more stable increase on SciScore throughout training.

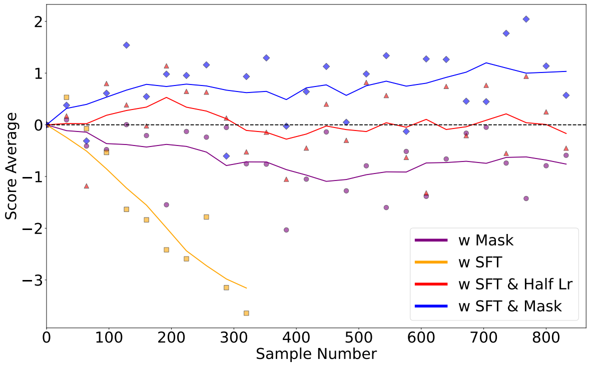


Figure 6. Ablation study of the two-stage training process. At each step, all prompts in PS are employed to generate two images per prompt, followed by the calculation of the average SciScore. The result illustrates the deviation from the initial baseline.

8. Conclusion

We present SciScore, a reward model aimed at integrating scientific knowledge into image synthesis models. Utilizing our expert-annotated dataset, SciBench, with over 60,000 images and 9,000 prompt in total, we established a framework for evaluating and improving image realism. Our two-stage training approach, featuring supervised fine-tuning and masked online fine-tuning, led to significant performance enhancements. We demonstrate that SciScore achieves human-level performance in aligning outputs with scientific knowledge.

References

[1] Flux. https://blackforestlabs.ai/. 1, 2, 5, 7, 8, 3, 9

[2] Gpt-4o. https://openai.com/index/hello-gpt-4o/. 2, 3, 6, 1, 4, 7

[3] Ic-light. https://openreview.net/pdf?id=u1cQYxRI1H. 1

[4] Laion-aesthetics. https://laion.ai/blog/laion-aesthetics/. 9, 10

[5] Dalle-3. https://openai.com/index/dall-e-3/. 1, 3

[6] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024. 1, 2

[7] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning, 2024. 4, 5, 9

[8] Wengling Chen and James Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[9] 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-image generation. *arXiv preprint arXiv:2310.18235*, 2023. 3

[10] Michael F Cohen and John R Wallace. *Radiosity and realistic image synthesis*. Morgan Kaufmann, 1993. 1

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[12] Carles Domingo-Enrich, Michal Drozdzal, Brian Karrer, and Ricky T. Q. Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control, 2024. 5

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 5, 7, 1, 3, 8

[14] Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *arXiv preprint arXiv:2406.04312*, 2024. 2

[15] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models, 2023. 4

[16] Hany Farid. Perspective (in)consistency of paint by text, 2022. 1

[17] James A Ferwerda, Sumanta N Pattanaik, Peter Shirley, and Donald P Greenberg. A model of visual adaptation for realistic image synthesis. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 249–258, 1996. 1

[18] Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. Commonsense-t2i challenge: Can text-to-image generation models understand commonsense?, 2024. 1, 3

[19] Donald P Greenberg, Kenneth E Torrance, Peter Shirley, James Arvo, Eric Lafortune, James A Ferwerda, Bruce Walter, Ben Trumbore, Sumanta Pattanaik, and Sing-Choong Foo. A framework for realistic image synthesis. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 477–494, 1997. 1

[20] Jianshu Guo, Wenhao Chai, Jie Deng, Hsiang-Wei Huang, Tian Ye, Yichen Xu, Jiawei Zhang, Jenq-Neng Hwang, and Gaoang Wang. Versat2i: Improving text-to-image models with versatile reward. *arXiv preprint arXiv:2403.18493*, 2024. 2

[21] Minghao Guo, Bohan Wang, Pingchuan Ma, Tianyuan Zhang, Crystal Elaine Owens, Chuang Gan, Joshua B Tenenbaum, Kaiming He, and Wojciech Matusik. Physically compatible 3d object modeling from a single image. *arXiv preprint arXiv:2405.20510*, 2024. 2

[22] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 3, 5

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3

[24] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 7

[25] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023. 3

[26] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 3

[27] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 6, 3, 4, 5

[28] Henrik Wann Jensen. *Realistic image synthesis using photon mapping*. AK Peters/crc Press, 2001. 1

[29] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video

generation from world model: A physical law perspective. 2024. 2

[30] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. 8

[31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 9

[32] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023. 3

[33] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. 3

[34] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*, 2023. 3

[35] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Evaluating and improving compositional text-to-visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5290–5301, 2024. 3

[36] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 6, 3, 4, 5

[37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 6, 3

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 3

[39] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. 5, 8

[40] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. 5, 6, 7, 9

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 2, 9

[42] 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 preprint arXiv:2305.11116*, 2023. 3

[43] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024. 1, 2

[44] Fanqing Meng, Wenqi Shao, Lixin Luo, Yahong Wang, Yiran Chen, Quanfeng Lu, Yue Yang, Tianshuo Yang, Kaipeng Zhang, Yu Qiao, et al. Phybench: A physical commonsense benchmark for evaluating text-to-image models. *arXiv preprint arXiv:2406.11802*, 2024. 1, 2, 3, 5, 11

[45] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. 4

[46] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 3

[47] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 1, 7, 3, 5, 8

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 4, 6, 5

[49] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. 5, 8

[50] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 7, 5, 8, 9

[52] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 3

[53] Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, D. A. Forsyth, and Anand Bhattad. Shadows don't lie and lines can't bend! generative models don't know projective geometry...for now, 2024. 1

[54] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 1

[55] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022. 4

[56] Jiao Sun, Deqing Fu, Yushi Hu, Su Wang, Royi Rassin, Da-Cheng Juan, Dana Alon, Charles Herrmann, Sjoerd van Steenkiste, Ranjay Krishna, et al. Dreamsync: Aligning text-to-image generation with image understanding feedback. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2023. 2

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 4

[58] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization, 2023. 4

[59] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 2

[60] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. 6, 3, 4, 5

[61] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 3

[62] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023. 3, 9, 10

[63] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Qimai Li, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model, 2024. 1, 4

[64] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8941–8951, 2024. 2

[65] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[66] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018. 1

[67] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3

SciBench: Addressing Scientific Illusions in Image Synthesis

Supplementary Material

The supplementary material is structured as follows:

A. Rationale to Prioritize Rewriting Capability

In the development of our study, we considered incorporating additional reasoning tasks, such as reflection-based tasks [16] that evaluate the consistency between objects and their reflections. However, these tasks present unique challenges that influenced our decision to exclude them.

Reflection-based tasks require the representation of precise geometric details to capture the relationships between objects and their reflections. Such intricate geometric information cannot be fully conveyed through textual descriptions alone. Consequently, current text-to-image generation models face difficulties in producing both correct and incorrect images for these tasks. This limitation hampers the creation of a consistent and valid dataset necessary for evaluating generative models on reflection-based generation.

Given these constraints, we prioritized tasks that can be effectively rephrased and consistently described using language-based prompts. This ensures generative models can interpret and generate the required images more reliably, thereby facilitating robust data collection and analysis. By focusing on linguistically describable tasks, we enhance the reproducibility and validity of our findings.

Tasks that lack this flexibility, particularly those requiring detailed geometric representation [53] and those subtle light-related features [3] beyond the capacity of textual prompts, are reserved for future exploration.

B. Detailed Data Curation Process

In this section, we provide a detailed overview of our data curation process. We describe the methods used for generating subject-based prompts, synthesizing images, and establishing criteria for image selection.

Subject-Based Prompt. For each task, we first employ `GPT-4o` [2] to define a comprehensive set of templates for the implicit prompts. These templates act as structured frameworks that capture the essence of the reasoning required while allowing for variability in the objects or substances involved. Using the templates, `GPT-4o` [2] generated a variety of implicit prompts by inserting appropriate objects or substances into the placeholders. Then for each implicit prompt, we used `GPT-4o` [2] to generate the corresponding explicit prompt and superficial prompt. An illustration of this instruction process is provided in Figure B1.

Synthetic Image Generation. The limited availability of images relevant to our specific scientific reasoning tasks within existing datasets and online resources necessitated the generation of synthetic images. However, we could not arbitrarily select a text-to-image model, as this choice directly affects both the quality of the generated data and the efficiency of data acquisition. Among the numerous advanced models available, our choice was informed by a comprehensive evaluation of several key factors. Below, we outline the primary considerations that guided our decision:

- **Descriptive Text-Image Alignment**: The core objective involves generating images that accurately reflect both explicit and superficial prompts. This necessitates a model with a robust capability to align textual descriptions with corresponding visual elements. Meanwhile, effective text-image alignment is also paramount for efficient data collection.
- **Realistic Style Consistency**: Our reasoning-based tasks are fundamentally grounded in scientific principles and real-world phenomena. Consequently, it is imperative that the generated images exhibit a style that reflects realism rather than abstract or cartoonish representations.

Based on these criteria, we conducted a qualitative evaluation of several state-of-the-art text-to-image models, including Stable Diffusion XL [47], Stable Diffusion 3 [13], DALLE 3 [5], and FLUX.1[dev] [1]. As illustrated in Figure B2, FLUX.1[dev] [1] consistently outperformed the other models in both text-image alignment and realistic style consistency. Therefore, FLUX.1[dev] [1] was selected as the model for synthetic image generation.

> User Prompt
>
> Assume you are an experienced scientist. Your task is to generate both a explicit prompt and a superficial prompt based on a given input prompt. The input prompt is formulated with scientific principles and will serve as input for a text-to-image generative model. It may include terminology or phrases that are not overtly descriptive but imply certain visual characteristics or phenomena, requiring interpretative scientific reasoning to convey their meaning.
>
> Explicit Prompt: Reformulate the input prompt into a precise, explicit, and descriptively accurate statement that aligns with the intended visual outcome, incorporating the implied scientific nuances and characteristics.
>
> Superficial Prompt: Construct an explicit interpretation of the input prompt that disregards the underlying scientific reasoning or implied elements. Focus only on the superficial or literal descriptive aspects.
>
> Example: {“input prompt”: “an unripe apple”, “explicit prompt”: “a green apple”, “superficial prompt”: “a red apple”}
>
> Here is the input prompt: [Your Input] and please output in the following format:
>
> {“explicit prompt”: , “superficial prompt”: }

Figure B1. **Framework For Prompt Collection.** This figure presents a detailed workflow for generating explicit and superficial prompts from implicit input prompts using `GPT-4o` [2].

Criteria For Image Curation. As outlined in Section 3, the scientific principles inherent in the implicit prompt confer distinct visual features to the subject matter. During the image generation process for SciBench, particular emphasis was placed on the regions where these visual features are manifested. Our primary objective was to ensure that these regions accurately represent the concepts in alignment with the underlying scientific principles specified in the prompts. To achieve this, we established stringent criteria for the images, specifically: (1) minimizing noise and (2) preventing the introduction of irrelevant semantic information. As illustrated in Figure Q9,Q10,Q11, we accomplish this by selecting data with the simplest possible backgrounds, such as solid colors. Additionally, we filter the data to ensure that the regions of interest are as large as possible, thereby maximizing the prominence of the visual features.

C. Detailed Training Settings for SciScore

This section provides a overview of the hyper-parameter settings utilized during the training of SciScore. The key parameters, including batch size, learning rate, and optimizer configurations, are summarized in Table C1.

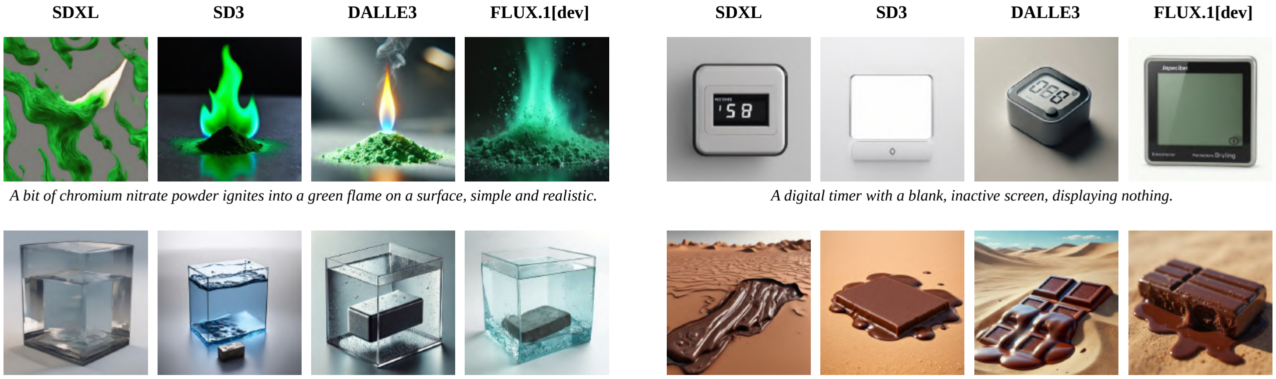
Table C1. Hyper-parameter settings used for training SciScore.

Hyper-parameters	SciScore
batch size	128
learning rate	2×10^{-6}
learning rate schedule	cosine
weight decay	0.3
training steps	600
warmup steps	150
optimizer	AdamW [41]
λ	0.25

D. Setup of SciBench S and SciBench C

For the evaluation of SciScore, two meticulously curated test sets are employed, each manually annotated and subjected to a stringent iterative review process by domain experts. This process involved cross-referencing the annotators’ specialized knowledge with authoritative online sources to ensure accuracy and consistency. The validation procedure was repeated until unanimous consensus was achieved among all annotators, thereby enhancing the reliability of the test sets. These sets are strategically designed to evaluate the model’s performance across varying levels of complexity and are characterized as follows:

- **SciBench S**: This test set closely replicates the stylistic and structural attributes of the training data. It emphasizes simplicity by focusing on specific regions and strictly adhering to the annotation criteria in Section B. The goal of SciBench S is to assess the model’s performance on data stylistically similar to its training set.
- **SciBench C**: This test set challenges the model in more complex scenarios, introducing contextual elements like explicit scene settings and diverse scenarios. Prompts in SciBench C may include phrases such as ”in a bedroom” or ”on the street,” adding spatial and contextual variability. This complexity evaluates the model’s ability to adapt to nuanced, less constrained environments.



A bit of chromium nitrate powder ignites into a green flame on a surface, simple and realistic.

A digital timer with a blank, inactive screen, displaying nothing.

A transparent box filled with water holds an iron block lying on the bottom, realistic.

A heavily melting chocolate bar in the desert, losing its original shape as liquefied portions spread into a glossy area. The remaining solid slumps, soft and irregular.

Figure B2. **Comparative Data Analysis.** Models such as SDXL [47], SD3 [13], and DALLE3 [5] occasionally failed to align generated images accurately with the provided textual descriptions. Meanwhile, FLUX.1[dev] [1] demonstrated superior performance, producing the most realistic images among all evaluated models.

E. Detailed Baseline Setup for SciScore

This section provides detailed descriptions of the baseline setups employed to evaluate the performance of SciScore.

Vision-Language Models (VLMs). We employ two VLMs as baseline models, CLIP-H [27] and BLIP-2 [37]. The reward computation involves encoding the implicit input prompt and the input image using their respective text and image encoders. Subsequently, we apply the scoring mechanism described in Section 4.1 to evaluate the alignment between the text and image pairs.

Language Multimodal Models (LMMs). As a baseline for LMMs, we leverage `GPT-4o-mini`[2]. To assess its performance, we conduct evaluations under two different settings: one without employing the Chain-of-Thought (CoT) reasoning approach[60] and another incorporating CoT [60] to facilitate step-by-step reasoning. Specifically, we prompt `GPT-4o-mini`[2] to choose between two images by selecting either "the first" or "the second." Recognizing that the model may exhibit insensitivity to the order of image presentation, we mitigate this potential bias by conducting the evaluation twice, each time with the order of the input images reversed. We then compute the average accuracy across these two evaluations to obtain a more robust and reliable performance measure. The complete instruction set is detailed comprehensively in Figure E3.

Human Evaluation. To provide a human performance baseline, we collected data from 10 human evaluators, all of whom hold at least a college degree, primarily in science or engineering disciplines. This selection criterion ensures that the evaluators possess foundational scientific knowledge necessary to perform inference tasks.

F. Additional Results of SciScore

In this section, we extend Table 1 by providing detailed accuracy metrics for each category in Tables F3 and F2, which allows for a more nuanced evaluation of SciScore's performance across different categories: light requirement (LR),watering requirement (WR), ripeness (RI), seasonal change (SC), flame reaction (FR), immiscibility (IM), rust (RU), absorption (AB), buoyancy (BU), diffusion (DI), electricity (EL), evaporation (EV), gravity (GA), liquidation (LI), melting (ME), solidification (SO).

The extended results demonstrate that SciScore consistently outperforms baseline models across the majority of tasks. Furthermore, SciScore achieves perfect accuracy (100%) on several specific tasks, underscoring its effectiveness and robustness in diverse scenarios.

G. Additional Analysis

In this section, we present further in-depth analysis pertaining to the results and observations discussed in Section 6.2.

Performance of VLMs Approaches Random Guessing. Both CLIP-H [22] and BLIPScore [36] demonstrate low accuracy, hovering around 50, across both test sets. This suboptimal performance is primarily attributable to the pretraining phase, where the majority of textual data are highly descriptive and explicitly reference their corresponding visual content. As a result, during inference, the text encoder predominantly relies on these descriptive terms within the prompt. When a test prompt is associated with two images that both contain the main elements described in the prompt, the model struggles to differentiate between them effectively. This ambiguity leads to performance that is comparable to random guessing, highlighting a significant limitation in the current pretrained multimodal model. Furthermore,

User Prompt

You will be presented with a textual prompt followed by two visual images. Your task is to critically analyze and compare both images, selecting the one that most accurately aligns with and represents the overall meaning of the given prompt. First, you should imagine how an ideal image would look based on the prompt, and then describe both images in detail. Finally, combining your initial visualization with the descriptions of the two images, you should select the image that most effectively conveys the intended meaning of the prompt, providing a reasoned justification for your choice.

Here is the input: {"prompt": [Your Input Prompt], "image-1": [Your Input Image], "image-2": [Your Input Image]}

Please output in the following format:

{'imagination': , 'description of image-1': , 'description of image-2': , 'justification for choice': , 'final choice': }

Figure E3. **Instruction For GPT Evaluation.** Text segments in red are specifically incorporated to facilitate CoT [60] reasoning.

Table F2. Performance comparison on SCIBENCH S and across different categories. **Bold** values indicate the best performance.

Model	ME	DI	EL	SO	IM	EV	AB	LI	FR	SC	RI	RU	LR	WR	BU	GR
CLIP-H [27]	25.00	71.43	47.62	40.48	54.17	26.67	57.14	77.78	73.33	81.48	34.62	16.67	62.22	31.11	63.89	78.33
BLIPScore [36]	56.94	50.00	52.38	44.05	53.12	20.00	38.10	33.33	76.67	58.33	38.46	42.86	76.67	38.89	50.00	47.50
GPT-4o mini	36.11	77.38	82.14	35.71	65.63	**100.00**	33.33	76.39	58.89	97.22	53.85	95.24	96.67	83.33	56.94	71.31
+ CoT [60]	36.11	85.71	86.90	45.24	68.75	**100.00**	33.33	81.94	56.67	98.15	61.54	97.62	96.67	88.89	52.78	80.33
Human Eval	98.15	65.87	95.63	86.11	**77.78**	**100.00**	66.67	82.08	80.95	90.74	94.62	92.86	96.89	99.56	**74.55**	92.99
SCISCORE (ours)	**100.00**	**97.62**	**100.00**	**90.48**	68.75	**100.00**	**71.43**	**100.00**	**97.78**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	66.67	**98.33**

Figure G4 provides a comparative analysis of the ROC curves for SCISCORE, CLIP-H [27], and BLIPScore [36], illustrating the relative performance of each model.

Limitations of LMMs in Vision-Based Scientific Reasoning. Despite being equipped with an extensive knowledge base, GPT-4o-mini [2] fails to achieve satisfactory performance in vision-based scientific reasoning tasks, even when incorporating advanced techniques such as Chain-of-Thought (CoT) prompting [60]. We posit that the primary reasons for this inadequate performance are twofold. First, the model exhibits a limited capacity to accurately capture and interpret the complex visual features inherent in scientific data, such as intricate diagrams, graphs, and microscopic images, which are crucial for tasks that rely heavily on visual information. This limitation hampers the model's ability to effectively integrate visual inputs with its existing knowledge base, leading to superficial or incorrect interpretations. Second, during the inference process, the model tends to generate reasoning chains that contain internal contradictions and inconsistencies, undermining the overall reliability and coherence of its scientific reasoning. These contradictory reasoning patterns within the CoT [60] framework suggest a fundamental challenge in maintaining logical consistency when processing and synthesizing information from visual sources, especially when dealing with complex or ambiguous data. To substantiate these claims, we present qualitative results in Figure G5, which illustrate specific instances where GPT-4o-mini [2] fails to accurately interpret visual data and produces reasoning sequences that are internally conflicting and logically flawed.

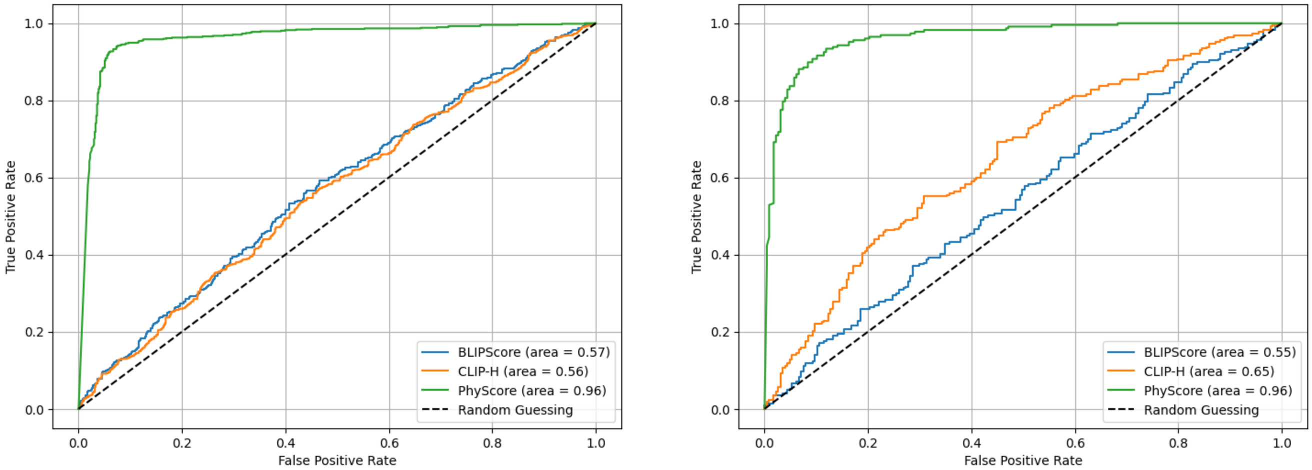
SCISCORE Achieves Human-Level Performance. This enhanced efficacy can be primarily attributed to the inherent limitations in the specialized expertise of human evaluators. Although these evaluators typically possess undergraduate or advanced degrees and maintain a foundational understanding of relevant scientific domains, their knowledge bases are finite and often constrained by the boundaries of their specific areas of expertise. Such limitations can impede their ability to accurately and comprehensively assess all instances within diverse and extensive test sets, particularly when confronted with novel or interdisciplinary examples that lie outside their immediate knowledge scope. In contrast, SCISCORE leverages extensive contextual knowledge acquired from the training data, enabling it to generalize effectively and maintain consistent performance across diverse and challenging test scenarios.

H. Qualitative Analysis of IEE

Qualitative results, as shown in Figure H6, demonstrate the effectiveness of incorporating IEE loss at an appropriate rate. The examples presented focus on the model's ability to capture fine-grained and nuanced details. In the first two pairs, the task involves distinguishing between the frozen and liquid states of various liquids, which relies on sub-

Table F3. Performance comparison on SciBench C across different categories. **Bold** values indicate the best performance.

Model	ME	DI	EL	SO	IM	EV	AB	LI	FR	SC	RI	RU	LR	WR	BU	GR
CLIP-H [27]	66.67	78.57	21.43	57.14	50.00	0.00	64.29	66.67	46.67	88.89	75.00	35.71	80.00	60.00	58.33	75.00
BLIPScore [36]	58.33	50.00	28.57	42.86	62.50	50.00	50.00	29.17	60.00	75.00	54.17	57.14	53.33	46.67	62.50	40.00
GPT-4o mini	67.65	67.86	64.29	50.00	68.75	90.00	50.00	75.00	53.33	88.89	87.50	89.29	**100.00**	83.33	54.17	97.50
+ CoT [60]	67.65	**85.71**	85.71	57.14	68.75	95.00	32.14	79.17	50.00	88.89	87.50	**92.86**	**100.00**	93.33	41.67	**100.00**
Human Eval	91.03	66.75	**90.87**	77.55	**86.61**	95.71	**78.57**	76.79	77.14	96.83	83.78	**92.86**	88.57	84.76	**83.33**	98.57
SciScore (ours)	**100.00**	**85.71**	85.71	**92.86**	81.25	**100.00**	71.43	**100.00**	**100.00**	**100.00**	**100.00**	**92.86**	**100.00**	**100.00**	41.67	**100.00**



(a) SciBenchS

(b) SciBenchC

Figure G4. **ROC Curve Analysis.** The AUC scores for both BLIPScore [36] and CLIP-H [22] are relatively low, implying that these models exhibit only marginally better performance than a random classifier. In contrast, SciScore demonstrates superior efficacy, with a nearly optimal AUC score, indicating a high level of discriminative power and robustness in classification performance.

tle differences in transparency—frozen water exhibits lower transparency compared to liquid water. The third example pertains to a localized region within the image, where the model must determine whether the screen within this small region displays meaningful content. By incorporating IEE loss, the model enhances its visual discrimination and contextual analysis capabilities, enabling it to make more accurate and context-aware predictions.

I. Detailed Benchmarking Configuration

To facilitate equitable comparisons among the different T2I models, we standardized the output image resolution to 1024×1024 pixels for all models. Table I4 summarizes the configuration parameters used for each model, including the guidance scale and the number of inference steps.

J. More Results on Benchmarking T2I Model

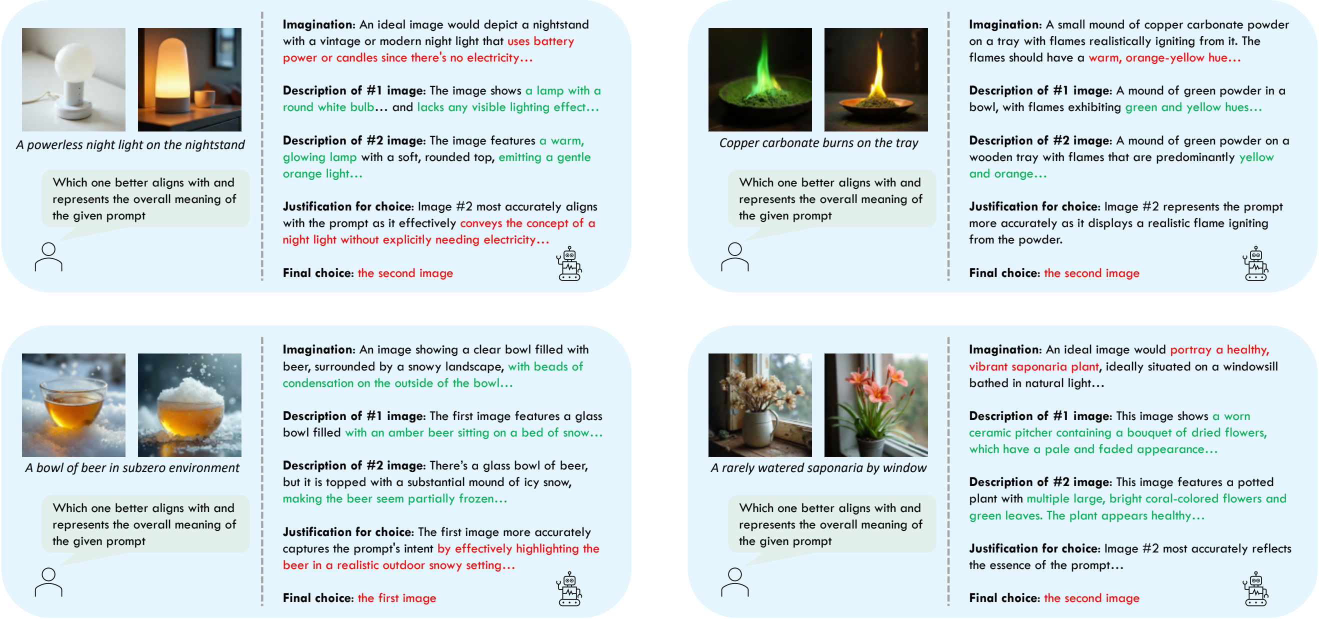
In this section, we further employ SciScore to benchmark additional state-of-the-art text-to-image models. Due to budgetary constraints, our evaluation is limited to open-source models. The results are presented in Table 3.

Table I4. Configurations of each T2I model

T2I Model	Guidance Scale	Inference Step
SDv1.5 [51]	7.5	50
SDXL [47]	5.0	50
SD3 [13]	7.0	28
FLUX.1[schnell] [1]	0.0	4
FLUX.1[dev] [1]	0.0	30

K. More Results on Explicit Prompt Alignment

While SciScore effectively evaluates the alignment between an implicit prompt and an image, it shares a common limitation inherent to all CLIP-based models [48]: the scores are only meaningful when comparing different pairs. In other words, SciScore can indicate that one prompt-image pair has better alignment than another but does not provide an absolute measure. To overcome this limitation in the context of the Explicit Prompt Alignment evaluation, we have developed a systematic grading criterion to assess alignment comprehensively. Inspired by PhyBench [44], our grading process is divided into two distinct aspects:



A powerless night light on the nightstand

Which one better aligns with and represents the overall meaning of the given prompt

Imagination: An ideal image would depict a nightstand with a vintage or modern night light that uses battery power or candles since there's no electricity…

Description of #1 image: The image shows a lamp with a round white bulb… and lacks any visible lighting effect…

Description of #2 image: The image features a warm, glowing lamp with a soft, rounded top, emitting a gentle orange light…

Justification for choice: Image #2 most accurately aligns with the prompt as it effectively conveys the concept of a night light without explicitly needing electricity…

Final choice: the second image

Copper carbonate burns on the tray

Which one better aligns with and represents the overall meaning of the given prompt

Imagination: A small mound of copper carbonate powder on a tray with flames realistically igniting from it. The flames should have a warm, orange-yellow hue…

Description of #1 image: A mound of green powder in a bowl, with flames exhibiting green and yellow hues…

Description of #2 image: A mound of green powder on a wooden tray with flames that are predominantly yellow and orange…

Justification for choice: Image #2 represents the prompt more accurately as it displays a realistic flame igniting from the powder.

Final choice: the second image

A bowl of beer in subzero environment

Which one better aligns with and represents the overall meaning of the given prompt

Imagination: An image showing a clear bowl filled with beer, surrounded by a snowy landscape, with beads of condensation on the outside of the bowl…

Description of #1 image: The first image features a glass bowl filled with an amber beer sitting on a bed of snow…

Description of #2 image: There's a glass bowl of beer, but it is topped with a substantial mound of icy snow, making the beer seem partially frozen…

Justification for choice: The first image more accurately captures the prompt's intent by effectively highlighting the beer in a realistic outdoor snowy setting…

Final choice: the first image

A rarely watered saponaria by window

Which one better aligns with and represents the overall meaning of the given prompt

Imagination: An ideal image would portray a healthy, vibrant saponaria plant, ideally situated on a windowsill bathed in natural light…

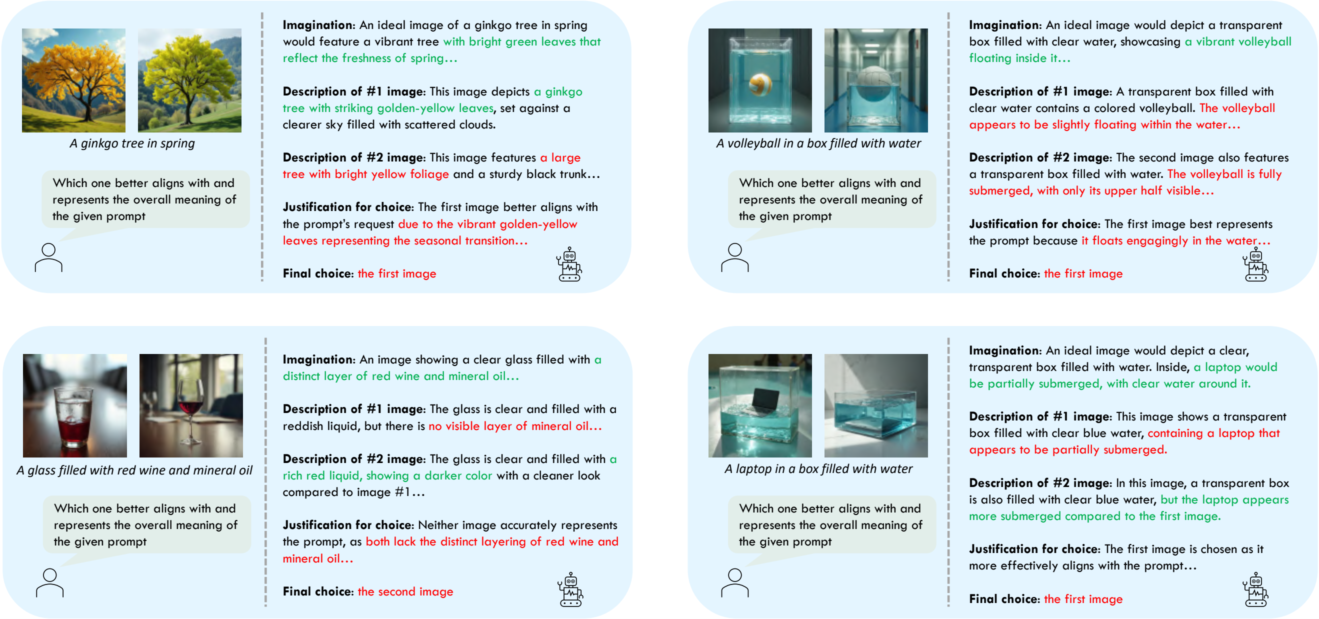
Description of #1 image: This image shows a worn ceramic pitcher containing a bouquet of dried flowers, which have a pale and faded appearance…

Description of #2 image: This image features a potted plant with multiple large, bright coral-colored flowers and green leaves. The plant appears healthy…

Justification for choice: Image #2 most accurately reflects the essence of the prompt…

Final choice: the second image

(a) **Reasoning Failure.** `GPT-4o-mini` [2] inaccurately infers the target image by misinterpreting the input prompt and neglecting the underlying scientific principles embedded within it. Instead of employing a systematic reasoning process, it relies predominantly on intuitive imagination.



A ginkgo tree in spring

Which one better aligns with and represents the overall meaning of the given prompt

Imagination: An ideal image of a ginkgo tree in spring would feature a vibrant tree with bright green leaves that reflect the freshness of spring…

Description of #1 image: This image depicts a ginkgo tree with striking golden-yellow leaves, set against a clearer sky filled with scattered clouds.

Description of #2 image: This image features a large tree with bright yellow foliage and a sturdy black trunk…

Justification for choice: The first image better aligns with the prompt's request due to the vibrant golden-yellow leaves representing the seasonal transition…

Final choice: the first image

A volleyball in a box filled with water

Which one better aligns with and represents the overall meaning of the given prompt

Imagination: An ideal image would depict a transparent box filled with clear water, showcasing a vibrant volleyball floating inside it…

Description of #1 image: A transparent box filled with clear water contains a colored volleyball. The volleyball appears to be slightly floating within the water…

Description of #2 image: The second image also features a transparent box filled with water. The volleyball is fully submerged, with only its upper half visible…

Justification for choice: The first image best represents the prompt because it floats engagingly in the water…

Final choice: the first image

A glass filled with red wine and mineral oil

Which one better aligns with and represents the overall meaning of the given prompt

Imagination: An image showing a clear glass filled with a distinct layer of red wine and mineral oil…

Description of #1 image: The glass is clear and filled with a reddish liquid, but there is no visible layer of mineral oil…

Description of #2 image: The glass is clear and filled with a rich red liquid, showing a darker color with a cleaner look compared to image #1…

Justification for choice: Neither image accurately represents the prompt, as both lack the distinct layering of red wine and mineral oil…

Final choice: the second image

A laptop in a box filled with water

Which one better aligns with and represents the overall meaning of the given prompt

Imagination: An ideal image would depict a clear, transparent box filled with water. Inside, a laptop would be partially submerged, with clear water around it.

Description of #1 image: This image shows a transparent box filled with clear blue water, containing a laptop that appears to be partially submerged.

Description of #2 image: In this image, a transparent box is also filled with clear blue water, but the laptop appears more submerged compared to the first image.

Justification for choice: The first image is chosen as it more effectively aligns with the prompt…

Final choice: the first image

(b) **Visual Limitation.** `GPT-4o-mini` [2] inaccurately describes the image, thereby impeding the reasoning process. Specifically, for tasks involving spatial relationships, it fails to make correct judgments, resulting in erroneous interpretations of positional dynamics within the visual content.

Figure G5. **Qualitative Failure Cases of GPT**. In both cases, the CoT [60] reasoning approach from Figure E3 is applied, but errors in either interpretation or visual comprehension impact the final decision. Green text indicates correct inference, while red text marks errors.

- **Main Subject Alignment (Scene Score, SS)**: This aspect evaluates whether all descriptive visual content specified in the prompt is present in the corresponding image.
- **Implicit Visual Alignment (Reality Score, RS)**: This aspect assesses whether the implicit visual elements, derived from underlying scientific principles present in the implicit prompt, are accurately represented in the image.

For illustrative purposes, we present examples in Figure P7a. After establishing the grading criteria, we selected all implicit prompts and their corresponding ex-

Table J5. Performance of T2I Models on SciScore.

T2I Model	Size	SciBench S				SciBench C			
		SP	EP	IP	ND	SP	EP	IP	ND
Stable Diffusion 3.5	medium	20.40	34.29	24.11	26.71	24.67	36.14	29.32	40.54
	large	20.11	33.72	24.39	31.45	24.68	35.11	29.28	44.10
	turbo	19.37	31.57	22.71	27.38	23.86	33.49	27.38	36.55

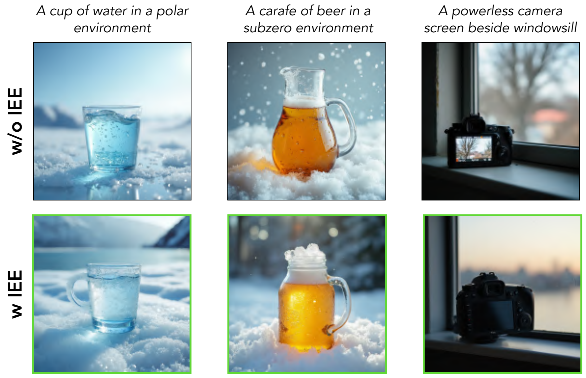

A cup of water in a polar environment
A carafe of beer in a subzero environment
A powerless camera screen beside windowsill
w/o IEE
w IEE

Figure H6. **Qualitative Analysis of IEE.** Images enclosed by green borders denote the correct selection in each pair.

plicit prompts from SciBench S and SciBench C. Using text-to-image models, we generated two images for each explicit prompt. These images were then evaluated by `GPT-4o-mini` [2] following the instructions detailed in Figure P7b. This evaluation produced average scene scores and reality scores with the experimental results summarized in Table K7. To further substantiate the effectiveness of the GPT-based evaluation, we examine the concordance between GPT assessments and human evaluations.

Relative Weakness in Scientific Scene Generation. The results presented in Table K7 indicate that the average full score (FS = SS + RS) is consistently lower than the scene score across all models. This suggests that the models exhibit weaker performance when generating outputs related to complex scientific phenomena compared to simpler subjects within prompts. A plausible explanation is that these phenomena often involve intricate features such as spatial relationships or uncommon object states (e.g., melting chocolate, a cup of frozen water), which are underrepresented in the models' pretraining data.

Concordance Between GPT and Human. To assess the effectiveness of GPT-based evaluation methods, we designed an experiment aimed at demonstrating the alignment between GPT's judgments and those of human experts. Utilizing an established evaluation framework, we applied the same scoring methodology, which is detailed in Figure P7b, to SciBench S and SciBench C. Human experts assigned scores based on these criteria, and after reaching consensus, we calculated the average scores. For explicit images, the human experts assigned an average scene score of 2 and an average reality score of 0; for superficial images, the average scene score was 2 and the average reality score was 3. Subsequently, we performed the same evaluation using `GPT-4o-mini` [2]. To quantify the correspondence between `GPT-4o-mini`'s evaluations and those of the human experts, we calculated the human correspondence (HC) for both scene and reality scores. The human correspondence for the scene score is computed as:

$$HC_{SS} = \frac{SS}{2.0} \times 100 \tag{21}$$

where SS is the scene score assigned by `GPT-4o-mini`. For reality scores, we used two separate formulas to compute the correspondence for explicit and superficial images. Specifically, for superficial images (SI), the human correspondence for reality score is calculated as:

$$HC_{RS}^{SI} = \left(1 - \frac{RS}{3.0}\right) \times 100 \tag{22}$$

For explicit images (EI), the human correspondence is:

$$HC_{RS}^{EI} = \frac{RS}{3.0} \times 100 \tag{23}$$

The comparative results are shown in Table K6.

Table K6. **Concordance Between `GPT-4o-mini` and Human Experts.** The average agreement rate of over 80% demonstrates `GPT-4o`'s strong alignment with human expert assessments of scene and reality aspects, highlighting its reliability.

Dataset	IT	SS	HC_{SS}	RS	HC_{RS}
SciBench S	EI	1.827	91.13	2.731	91.03
	SI	1.635	81.74	0.476	84.13
SciBench C	EI	1.855	92.73	2.490	83.00
	SI	1.630	81.50	0.636	78.79

Table K7. **Performance of T2I Models on Explicit Prompt Alignment.** The Full Score (FS) is the sum of the Scene Score (SS) and the Reality Score (RS): FS = SS + RS. The Percentage of Expectation (PoE) is calculated by dividing the score by its expected value.

T2I Model	SciBench S				SciBench C			
	SS	PoE	FS	PoE	SS	PoE	FS	PoE
Stable Diffusion v1.5 [51]	1.298	64.90	2.470	49.40	1.261	63.05	2.446	48.92
Stable Diffusion XL [47]	1.718	85.90	3.510	70.20	1.679	83.95	3.360	67.20
Stable Diffusion 3 [13]	1.786	89.30	3.898	77.96	1.780	89.00	3.836	76.72
FLUX.1[schnell] [1]	1.730	86.50	3.730	74.60	1.772	88.60	3.825	76.50
FLUX.1[dev] [1]	1.720	86.00	3.641	72.82	1.702	85.10	3.676	73.52
Expectation	2.000	100.00	5.000	100.00	2.000	100.00	5.000	100.00

L. Details of Two-Stage Training

In this section, we present a detailed overview of our two-stage training framework, which integrates SFT and masked online fine-tuning to enhance flow matching models.

Supervised Fine-tuning (SFT). Flow matching models [39] are continuous-time generative models that define a time-dependent velocity field $v(x_t, t)$ to transport samples from a noise distribution p_1 to data distribution p_0 over a time interval $t \in [0, 1]$. The transformation is governed by the ordinary differential equation (ODE):

$$\frac{dx_t}{dt} = v(x_t, t), \tag{24}$$

with the initial condition $x_1 \sim p_1$. The forward process is constructed as:

$$x_t = \alpha_t x_0 + \beta_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{25}$$

where $\alpha_0 = 1$, $\beta_0 = 0$, $\alpha_1 = 0$, and $\beta_1 = 1$, ensuring the consistency of the marginal distributions with the initial and terminal conditions. The velocity field $v(x_t, t)$ is represented as the sum of two conditional expectations:

$$v(x, t) = \dot{\alpha}_t \mathbb{E}[x_* | x_t = x] + \dot{\beta}_t \mathbb{E}[\epsilon | x_t = x], \tag{26}$$

which can be approximated by the model $v_\theta(x, t)$ by minimizing the following training objective:

$$\mathcal{L}_{SFT}(\theta) := \mathbb{E}_{x_*, \epsilon, t}\left[\|v_\theta(x_t, t) - \dot{\alpha}_t x_* - \dot{\beta}_t \epsilon\|^2\right] \tag{27}$$

Direct Preference Optimization (DPO). RLHF aims to optimize a conditional distribution $p_\theta(x_0|c)$ such that the expected reward $r(c, x_0)$ is maximized, while simultaneously regularizing the KL-divergence from a reference distribution p_{ref}. This objective is formulated as:

$$\max_{p_\theta} \mathbb{E}_{c, x_0 \sim p_\theta(x_0|c)} \left[r(c, x_0)\right] - \beta \mathcal{D}_{\text{KL}} \left[p_\theta(x_0|c) \| p_{\text{ref}}(x_0|c)\right] \tag{28}$$

where the hyper-parameter β controls regularization. According to [49], the unique global optimal solution p^*_θ to this optimization problem is given by:

$$p^*_\theta(x_0|c) = p_{\text{ref}}(x_0|c) \exp\left(\frac{r(c, x_0)}{\beta}\right) / Z(c) \tag{29}$$

where $Z(c) = \sum_{x_0} p_{\text{ref}}(x_0|c) \exp\left(\frac{r(c,x_0)}{\beta}\right)$ is partition function. Then the reward function can be expressed as:

$$r(c, x_0) = \beta \log \frac{p^*_\theta(x_0|c)}{p_{\text{ref}}(x_0|c)} + \beta \log Z(c) \tag{30}$$

To model human preferences, the Bradley-Terry (BT) model is employed, which represents the probability of one outcome being preferred over another as:

$$p_{BT}(x^w_0 \succ x^l_0 | c) = \sigma(r(c, x^w_0) - r(c, x^l_0)) \tag{31}$$

where σ is the sigmoid function, x^w_0 is the preferred outcome, and x^l_0 is the less preferred one.. $r(c, x_0)$ can be parameterized by a neural network ϕ and estimated via maximum likelihood training for binary classification:

$$L_{BT}(\phi) = \mathbb{E}_{c, x^w_0, x^l_0}\left[\log \sigma\left(r_\phi(c, x^l_0) - r_\phi(c, x^w_0)\right)\right] \tag{32}$$

By leveraging the relationship between the reward function and the optimal policy p^*_θ, the DPO objective is derived as:

$$\begin{aligned} \mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{c, x^w_0, x^l_0}\Bigg[\log\sigma\Bigg(\beta \log \frac{p_\theta(x^w_0|c)}{p_{\text{ref}}(x^w_0|c)} \\ - \beta \log \frac{p_\theta(x^l_0|c)}{p_{\text{ref}}(x^l_0|c)}\Bigg)\Bigg] \end{aligned} \tag{33}$$

Choice of σ_t. We determine the value of σ_t by adhering to the methodology presented in [30]. Initially, we define the hyperparameters S_{churn}, $S_{\min}$, $S_{\max}$, and S_{noise}. Subsequently, we define γ_t as follows:

$$\gamma_t = \begin{cases} \min\left(S_{\text{churn}} \cdot \Delta t, \sqrt{2} - 1\right) & \text{if } t \in [S_{\min}, S_{\max}] \\ 0 & \text{otherwise,} \end{cases} \tag{34}$$

where Δt represents the timestep difference between consecutive sampling steps. Following this, we define σ_t by

$$\sigma_t = S_{\text{noise}} \cdot \sqrt{\gamma_t^2 + 2\gamma_t} \cdot (1 - t). \tag{35}$$

Pre-Training Subject Extraction. We integrate GroundingDINO [40] to facilitate the extraction of masks from images. To streamline the process, we initially employ LLM to identify and extract the relevant subjects from the training prompt set prior to the training phase. The extracted subjects are subsequently provided to GroundingDINO [40] during training to generate corresponding masks. These masks are then utilized to apply gradient masking.

Gradient Masking. The mask generated by GroundingDINO [40] is derived from the resolution of the RGB image. However, gradients are computed within the model's latent space, as detailed in LDM [51]. The connection between the RGB image and the latent space is facilitated by a pretrained Variational Autoencoder (VAE) [31], which inherently exhibits localist properties. Specifically, let the latent representation have dimensions (H_l, W_l, C_l) and the corresponding decoded image have dimensions (H, W, C). If the mask extracted from the image is defined by the bounding box coordinates (x_1, y_1, x_2, y_2), then the corresponding mask in the latent space is computed as:

$$\left(\frac{x_1}{H} \cdot H_l,\ \frac{y_1}{W} \cdot W_l,\ \frac{x_2}{H} \cdot H_l,\ \frac{y_2}{W} \cdot W_l\right) \tag{36}$$

This latent-space mask is subsequently applied to the gradients of the model to modulate the training process.

Padding Technique. Certain tasks require the careful consideration of positional relationships rather than solely the object's internal state. For example, in the *gravity* task, the object's position relative to the ground is critically important, making the use of the object mask alone insufficient for accurate analysis. To address this limitation, we extend the height and width dimensions of the mask by an additional 10%. This strategic padding ensures that the surrounding positional context is adequately captured, improving task performance and contextual understanding.

M. Two-Stage Training Settings

In this section, we detail the hyper-parameter configurations employed in our two-stage training framework for the T2I model, which is presented in Table M8.

Table M8. Hyper-parameter settings for T2I Model fine-tuning.

Hyper-parameters	SFT	Online FT
batch size	16	8
learning rate	1×10^{-4}	6×10^{-4}
training steps	2456	103
optimizer	AdamW [41]	AdamW [41]
gradient accumulation	8	2
LoRA rank	16	16
S_{churn}	/	0.1
$S_{\min}, S_{\max}$	/	$0, \infty$
S_{noise}	/	1.0
β	/	10

N. Additional Results of Online Fine-tuning

To further assess the effectiveness of the proposed algorithm, we conducted additional experiments utilizing different reward models. Specifically, we employed the LAION aesthetic predictor [4] and ImageReward [62] as the reward functions for our comprehensive evaluations. It is important to note that, in these experiments, we did not implement the masking strategy described in the main text.

Training Setting. All configurations align with those presented in Table M8, except for the specific settings detailed below. We fine-tuned the FLUX.1[schnell] [1] using four inference steps. For training with the LAION aesthetic predictor [4], each training step involved sampling 64 images, employing a learning rate of 3×10^{-4}, and conducting training over 164 steps. When utilizing ImageReward [62] for training, we similarly sampled 64 images per step, applied a learning rate of 1×10^{-4}, implemented gradient accumulation step of 8, and trained for a total of 550 steps. Adhering to the configuration outlined in DDPO [7], the training prompt set comprised 45 distinct animal categories.

Evaluation Setting. The test prompt set consisted of an additional 10 animal categories not present in the training set. For each prompt, we generated 100 images and calculated the average reward assigned by the respective reward model, which served as our performance metric. The final experimental results, showcasing the average rewards achieved on the test set, are presented in Table N9.

O. Additional Observations

During our investigation, particularly in the data curation phase, we observed that all the scientific phenomena involved can be uniformly represented using a **subject + condition** framework. Specifically, all tasks involve implicit prompts structured in this manner. For example, the

Table N9. **Comparison of Average Rewards.** The online fine-tuning approach consistently outperforms the baseline, demonstrating the effectiveness of the proposed algorithm.

Method	LAION [4]	ImageReward [62]
FLUX.1[schnell]	5.855	0.949
+OFT	**6.074**	**1.023**

prompt *an unripe apple* comprises the subject *apple* and the condition *unripe*; similarly, *a laptop without electricity* includes the subject *laptop* and the condition *without electricity*. Building on this observation, we identified that, for each task, the component requiring scientific reasoning can be closely associated either with the subject or with the condition. We classify these tasks as *subject-oriented* tasks and *condition-oriented* tasks, depending on the reasoning focus.

Subject-Oriented Tasks. In subject-oriented tasks, the necessity for scientific reasoning arises primarily from the subject's properties. In these tasks, different subjects under the same condition exhibit different visual features due to their inherent characteristics. For example, the *buoyancy* task is subject-oriented because different objects placed in water either float or sink depending on their densities relative to water, which is an intrinsic property of the subjects.

Condition-Oriented Tasks. In condition-oriented tasks, scientific reasoning is predominantly associated with the condition applied to the subject. In these tasks, varying conditions applied to the same subject result in different visual features. For instance, the *gravity* task is condition-oriented since a subject exhibits different behaviors under different gravitational conditions: it floats in the air under ”without gravity” and rests on the ground under ”normal gravity.”

P. Limitations

Building upon the concepts introduced in Section O, we have observed that SciScore performs well on condition-oriented tasks following training, which is anticipated. However, our observations indicate that SciScore does not handle subject-oriented tasks effectively. A notable example is its weaker performance compared to humans on the *buoyancy* task, as illustrated in Table F3 and Table F2.

Q. Detailed Task Descriptions

In this section, we provide detailed descriptions of the tasks incorporated into our study. These tasks are designed to evaluate various biological, chemical, and physical phenomena presented in SciBench. Additionally, illustrative examples from SciBench are presented in Figure Q9,Q10,Q11 to demonstrate the tasks.

- **Light requirement**: Plants change color and leaf size based on adequate or insufficient light exposure.
- **Watering requirement**: Plants exhibit differences in foliage health, wilting, and growth when receiving sufficient or inadequate water, leading to reduced growth.
- **Ripeness**: Fruits alter their color and texture significantly when they are ripe compared to when they are unripe.
- **Seasonal change**: Plants display variations in leaf color, size, and blooming patterns across different seasons.
- **Flame reaction**: Chemical substances naturally produce their distinct flame colors vividly when burned.
- **Immiscibility**: Two liquids either mix uniformly or separate into layers based on their chemical properties.
- **Rust**: Metals appear shiny, smooth, and reflective before oxidation, and corroded, flaky, and brittle after rusting.
- **Absorption**: A solid either soaks up a liquid or repels it, depending on their material properties.
- **Buoyancy**: Substances either float on or sink in water based on their density relative to water.
- **Diffusion**: When a small amount of liquid is added, it either disperses uniformly or remains separate.
- **Electricity**: Electronic products change appearance, such as glowing or sparking, when electric current is applied.
- **Evaporation**: Liquids boil and produce vapor when reaching boiling points; otherwise, they remain calm.
- **Gravity**: Objects appear differently positioned when influenced by gravity versus in a gravity-free environment.
- **Liquidation**: Air condenses into water droplets on surfaces cooled below room temperature.
- **Melting**: Objects transition from solid to liquid, changing shape and structure upon reaching melting points.
- **Solidification**: Liquids become solids, altering their form and texture when cooled below solidification points.

Sampled Examples

Example1: [
“Prompt”: “A transparent water-filled box holds a basketball, depicted realistically.”,
“Scene Grading”: {
- 0 point: The picture does not feature a basketball inside a transparent box filled with water in any capacity.
- 1 point: The picture shows a basketball, but it is not inside a transparent box. Alternatively, the basketball are in a transparent box, but there is no water present.
- 2 points: The picture accurately depicts a basketball inside a transparent box filled with water.
},
“Reality Grading”: {
- 0 point: The basketball is completely sinking to the bottom of the water.
- 1-2 point: The basketball is completely submerging in the water, but doesn’t reach the bottom. Less mistakes will earn a higher score.
- 3 points: The picture shows basketball floating on the surface of the water.
}]

Example2: [
“Prompt”: “A clear glass filled with water and oil, simple and realistic.”,
“Scene Grading”: {
- 0 point: There is no glass or no liquid in the glass, or the scene is irrelevant (e.g., the focus is not on the glass or liquid at all).
- 1 point: The glass contains liquid, but the focus on the liquid or the glass is unclear, or there are distracting elements in the scene.
- 2 points: The glass is clearly depicted with some liquid in it, with no distractions, offering a simple, clear, and realistic depiction.”
},
“Reality Grading”:{
- 0 points: Liquids are mixed or incorrectly positioned (e.g., water and oil blended or misplaced).
- 1 point: Water and oil are present but with partial inaccuracies in separation or positioning (e.g., water floating on oil, blurred boundaries).
- 2 points: Liquids are correctly positioned with visible separation (oil atop water), but minor deviations from realism exist (e.g., slight issues with clarity or texture).
- 3 points: Fully realistic depiction with correct positioning (oil floating on water) and clear separation.
}]

(a) **Representative Samples.** Inspired by PhyBench [44], We present a two-tiered grading framework comprising ”Scene Grading” and ”Reality Grading.” The first level, Scene Grading, assesses fundamental alignment by verifying whether the primary subjects specified in the prompt are accurately depicted in the generated image. The second level, Reality Grading, evaluates the degree to which the generated image aligns with the implicit physical realities or expectations inherent in the implicit prompt.

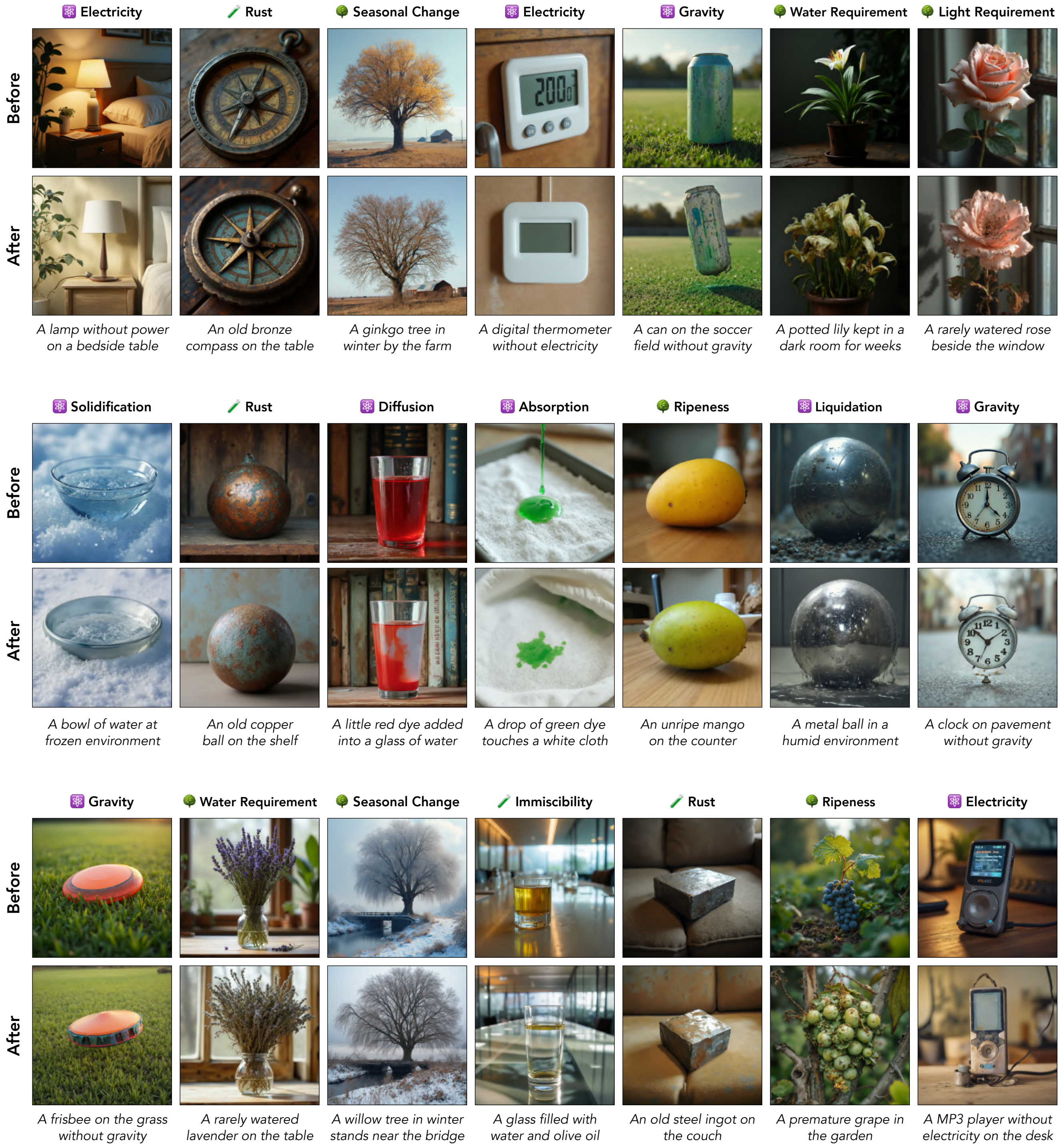
User Prompt

Imagine you are an experienced scientist. Begin by evaluating the provided image using the specified scene composition criteria. If the image does not fully satisfy these criteria, assign a reality score of 0. However, if the scene meets all the criteria, proceed to assess its realism based on the given reality scoring guidelines, disregarding stylistic aspects and minor background details. Please first describe the image in detail and then adhere strictly to these criteria to ensure an accurate scoring of the image.

Here is the input: {“Prompt”: [Your Input Prompt], “Scene Grading”: [Your Input Scene Grading], “Reality Grading”: [Your Input Reality Grading], “Image”: [Your Input Image]}. Please present your evaluation in the following format: {“description”:, “scene score”: , “reality score”: }

(b) **Image Evaluation Instruction.** In the context of the two-tiered grading framework, it is unnecessary to assess reality grading when an image fails to achieve a full score in scene grading. This is because reality grading presupposes that the main subject specified in the prompt is present in the image. Therefore, we assign a reality grading score of zero to any image that does not attain a full score in scene grading.

Figure P7. Sample prompts accompanied by corresponding evaluation criteria and instructions for image assessment



A lamp without power on a bedside table

An old bronze compass on the table

A ginkgo tree in winter by the farm

A digital thermometer without electricity

A can on the soccer field without gravity

A potted lily kept in a dark room for weeks

A rarely watered rose beside the window

A bowl of water at frozen environment

An old copper ball on the shelf

A little red dye added into a glass of water

A drop of green dye touches a white cloth

An unripe mango on the counter

A metal ball in a humid environment

A clock on pavement without gravity

A frisbee on the grass without gravity

A rarely watered lavender on the table

A willow tree in winter stands near the bridge

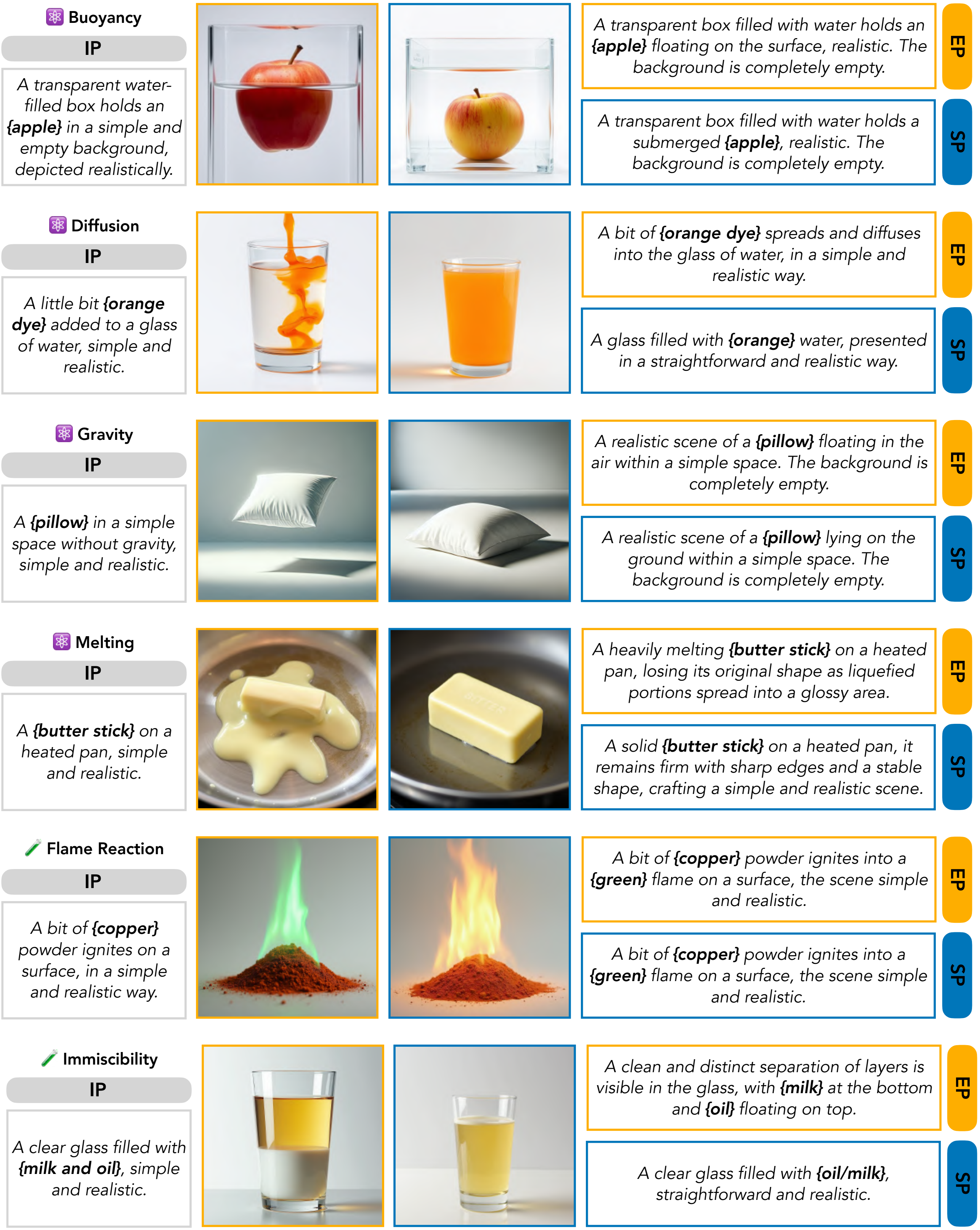
A glass filled with water and olive oil

An old steel ingot on the couch

A premature grape in the garden

A MP3 player without electricity on the desk

Figure P8. **Additional Generated Samples.** Each pair of images is produced using the same random seed to ensure consistency.



⚛ **Buoyancy**

IP

A transparent water-filled box holds an **{apple}** in a simple and empty background, depicted realistically.

EP: A transparent box filled with water holds an **{apple}** floating on the surface, realistic. The background is completely empty.

SP: A transparent box filled with water holds a submerged **{apple}**, realistic. The background is completely empty.

⚛ **Diffusion**

IP

A little bit **{orange dye}** added to a glass of water, simple and realistic.

EP: A bit of **{orange dye}** spreads and diffuses into the glass of water, in a simple and realistic way.

SP: A glass filled with **{orange}** water, presented in a straightforward and realistic way.

⚛ **Gravity**

IP

A **{pillow}** in a simple space without gravity, simple and realistic.

EP: A realistic scene of a **{pillow}** floating in the air within a simple space. The background is completely empty.

SP: A realistic scene of a **{pillow}** lying on the ground within a simple space. The background is completely empty.

⚛ **Melting**

IP

A **{butter stick}** on a heated pan, simple and realistic.

EP: A heavily melting **{butter stick}** on a heated pan, losing its original shape as liquefied portions spread into a glossy area.

SP: A solid **{butter stick}** on a heated pan, it remains firm with sharp edges and a stable shape, crafting a simple and realistic scene.

🧪 **Flame Reaction**

IP

A bit of **{copper}** powder ignites on a surface, in a simple and realistic way.

EP: A bit of **{copper}** powder ignites into a **{green}** flame on a surface, the scene simple and realistic.

SP: A bit of **{copper}** powder ignites into a **{green}** flame on a surface, the scene simple and realistic.

🧪 **Immiscibility**

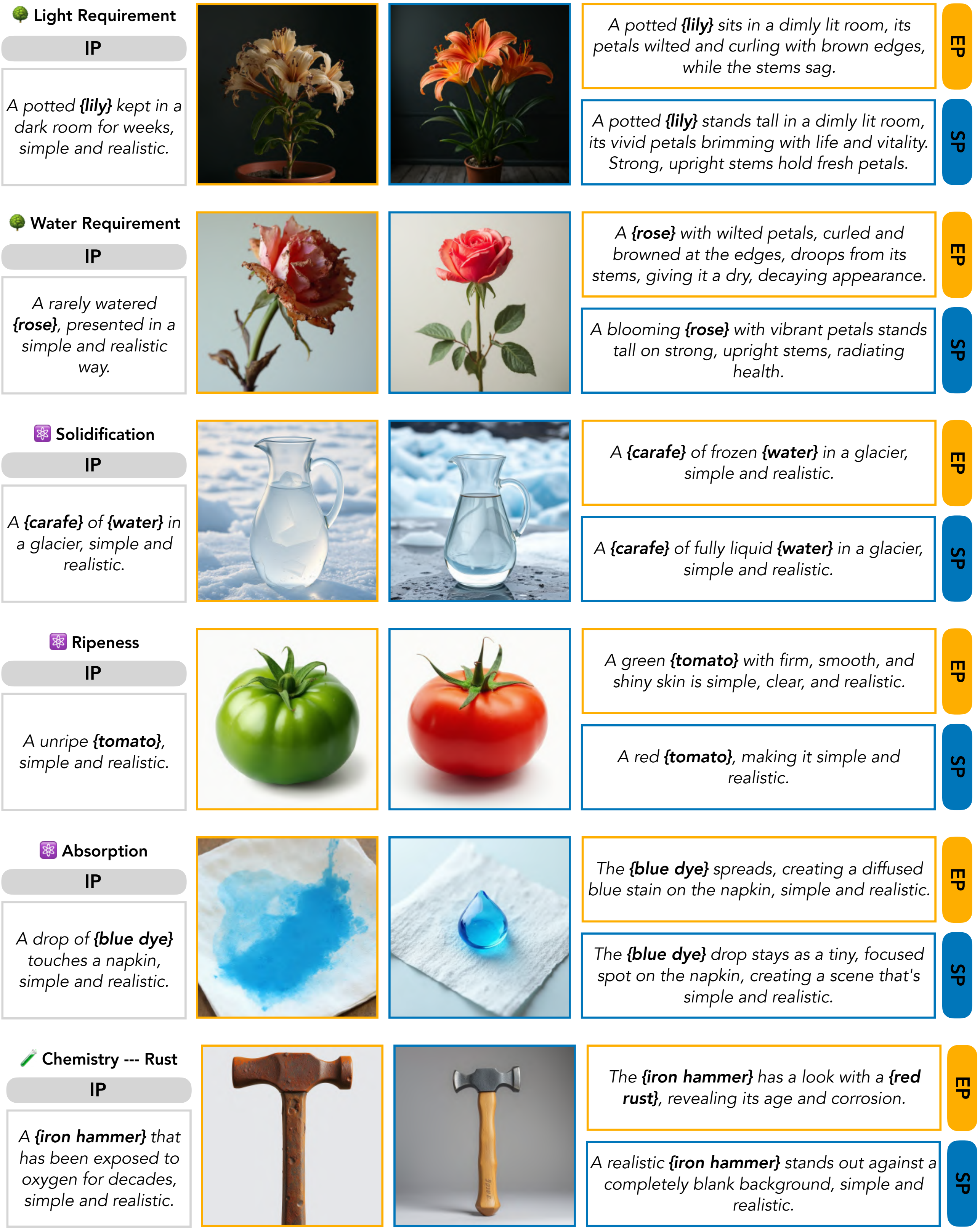
IP

A clear glass filled with **{milk and oil}**, simple and realistic.

EP: A clean and distinct separation of layers is visible in the glass, with **{milk}** at the bottom and **{oil}** floating on top.

SP: A clear glass filled with **{oil/milk}**, straightforward and realistic.

Figure Q9. **Several examples from SciBench.** 'EP' denotes explicit prompts (yellow blocks), 'SP' denotes superficial prompts (blue blocks), and 'IP' denotes implicit prompts (grey blocks).



🌳 Light Requirement

IP

*A potted **{lily}** kept in a dark room for weeks, simple and realistic.*

EP

*A potted **{lily}** sits in a dimly lit room, its petals wilted and curling with brown edges, while the stems sag.*

SP

*A potted **{lily}** stands tall in a dimly lit room, its vivid petals brimming with life and vitality. Strong, upright stems hold fresh petals.*

🌳 Water Requirement

IP

*A rarely watered **{rose}**, presented in a simple and realistic way.*

EP

*A **{rose}** with wilted petals, curled and browned at the edges, droops from its stems, giving it a dry, decaying appearance.*

SP

*A blooming **{rose}** with vibrant petals stands tall on strong, upright stems, radiating health.*

⚛️ Solidification

IP

*A **{carafe}** of **{water}** in a glacier, simple and realistic.*

EP

*A **{carafe}** of frozen **{water}** in a glacier, simple and realistic.*

SP

*A **{carafe}** of fully liquid **{water}** in a glacier, simple and realistic.*

⚛️ Ripeness

IP

*A unripe **{tomato}**, simple and realistic.*

EP

*A green **{tomato}** with firm, smooth, and shiny skin is simple, clear, and realistic.*

SP

*A red **{tomato}**, making it simple and realistic.*

⚛️ Absorption

IP

*A drop of **{blue dye}** touches a napkin, simple and realistic.*

EP

*The **{blue dye}** spreads, creating a diffused blue stain on the napkin, simple and realistic.*

SP

*The **{blue dye}** drop stays as a tiny, focused spot on the napkin, creating a scene that's simple and realistic.*

🧪 Chemistry --- Rust

IP

*A **{iron hammer}** that has been exposed to oxygen for decades, simple and realistic.*

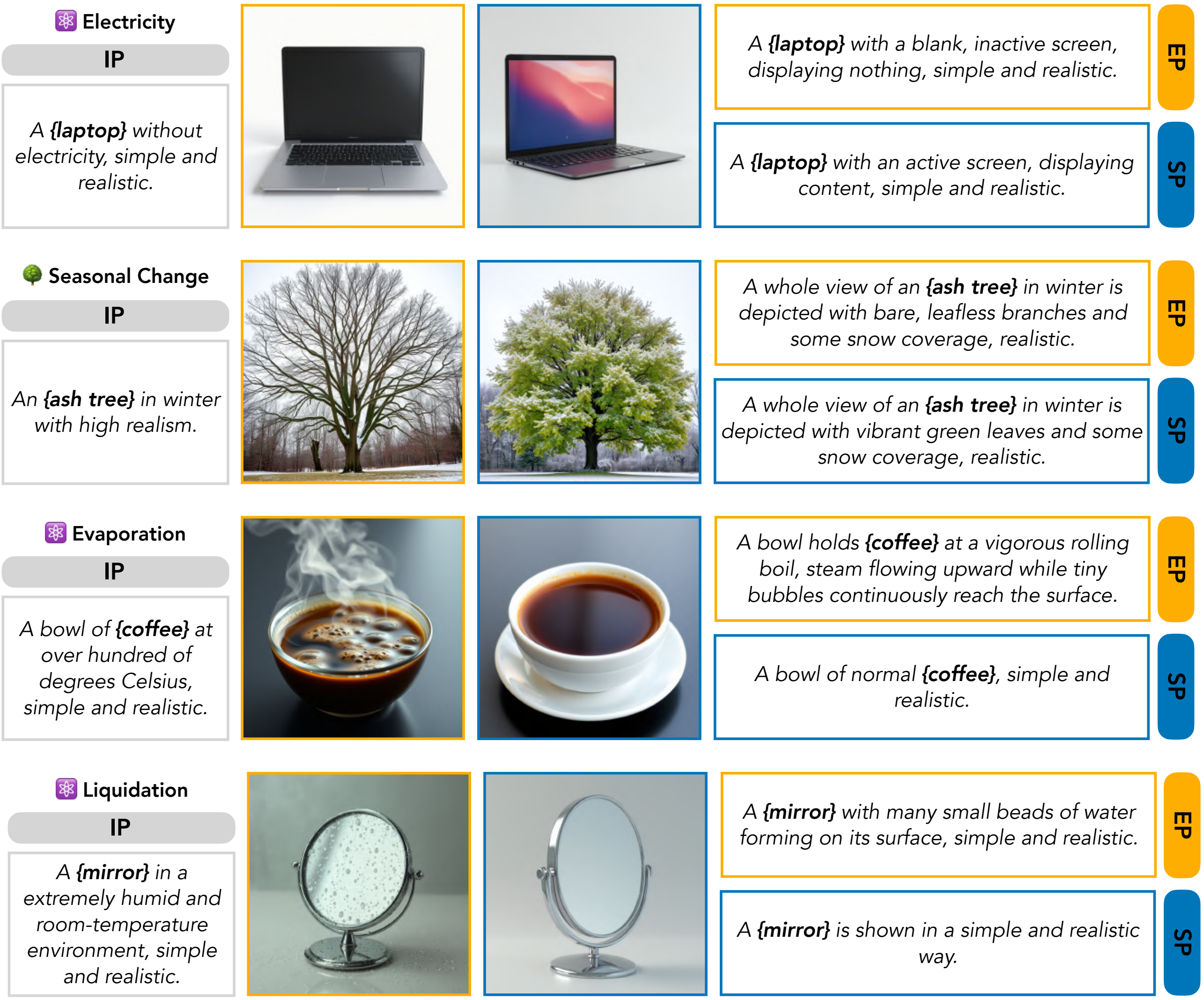
EP

*The **{iron hammer}** has a look with a **{red rust}**, revealing its age and corrosion.*

SP

*A realistic **{iron hammer}** stands out against a completely blank background, simple and realistic.*

Figure Q10. **Several examples from SciBench.** 'EP' denotes explicit prompts (yellow blocks), 'SP' denotes superficial prompts (blue blocks), and 'IP' denotes implicit prompts (grey blocks).



⚛ Electricity

IP

*A **{laptop}** without electricity, simple and realistic.*

*A **{laptop}** with a blank, inactive screen, displaying nothing, simple and realistic.*

EP

*A **{laptop}** with an active screen, displaying content, simple and realistic.*

SP

🌳 Seasonal Change

IP

*An **{ash tree}** in winter with high realism.*

*A whole view of an **{ash tree}** in winter is depicted with bare, leafless branches and some snow coverage, realistic.*

EP

*A whole view of an **{ash tree}** in winter is depicted with vibrant green leaves and some snow coverage, realistic.*

SP

⚛ Evaporation

IP

*A bowl of **{coffee}** at over hundred of degrees Celsius, simple and realistic.*

*A bowl holds **{coffee}** at a vigorous rolling boil, steam flowing upward while tiny bubbles continuously reach the surface.*

EP

*A bowl of normal **{coffee}**, simple and realistic.*

SP

⚛ Liquidation

IP

*A **{mirror}** in a extremely humid and room-temperature environment, simple and realistic.*

*A **{mirror}** with many small beads of water forming on its surface, simple and realistic.*

EP

*A **{mirror}** is shown in a simple and realistic way.*

SP

Figure Q11. **Several examples from SciBench.** 'EP' denotes explicit prompts (yellow blocks), 'SP' denotes superficial prompts (blue blocks), and 'IP' denotes implicit prompts (grey blocks).