This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Science-T2I: Addressing Scientific Illusions in Image Synthesis

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

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

Abstract

*We present a novel approach to integrating scientific knowledge into generative models, enhancing their realism and consistency in image synthesis. First, we introduce **Science-T2I**, an expert-annotated adversarial dataset comprising adversarial 20k image pairs with 9k prompts, covering wide distinct scientific knowledge categories. Leveraging **Science-T2I**, 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. Additionally, based on **Science-T2I**, 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 evaluators. Furthermore, by applying our proposed fine-tuning method to FLUX, we achieve a performance enhancement exceeding 50% 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 [7, 15, 17, 26, 60, 61]. As articulated by [9], "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 [43, 47, 49]. While these advancements enable the generation of higher resolution, more aesthetically pleasing images with superior Frechet Inception Distance (FID) scores [1, 3, 43, 58], these models often produce superficial imitations rather than authentic representations of the real visual world [4, 16, 39, 40]. 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 integrates scientific knowledge into image synthesis to bridge the gap between imagination and realism. We introduce **Science-T2I**, a comprehensive and expert-annotated dataset comprising over 20k 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 **Science-T2I**, 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 in Figure 1, and functions as a comprehensive end-to-end reward model – eliminating the dependence on language-guided inference processes, which can may fail 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 **Science-T2I**. 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.

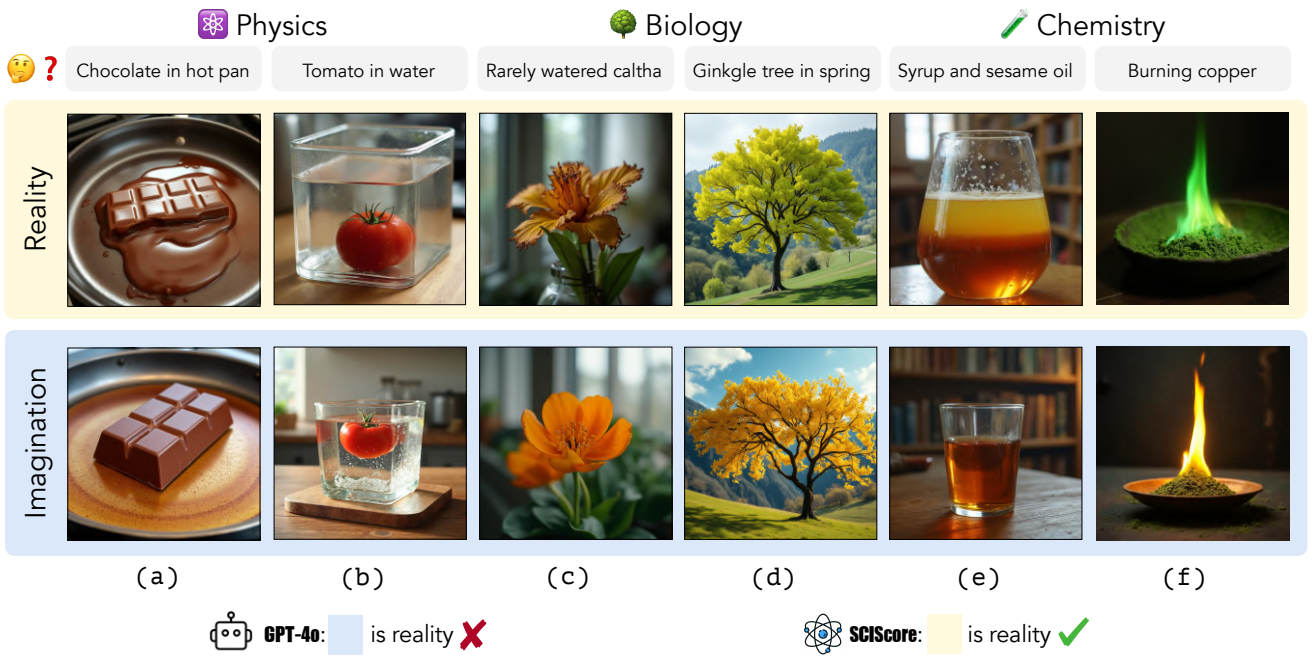

Figure 1. **Comparison between GPT-4o and SciScore.** Given a prompt (in grey) requiring scientific knowledge, FLUX [1] model generates imaginary images (lower row) that are far from reality (upper row). LMMs such as GPT-4o [2] fail to distinguish which image aligns better with reality. In contrast, 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 **Science-T2I** of over 9,000 prompts and 20,000 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 [35, 40], video generation [4, 6, 27, 39], and 3D modeling [19]. PhyBench [40] 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 [4] and PhyGenBench [39] evaluate whether current generative models can accurately simulate physical commonsense in real-world scenarios involving various material interactions. PhysComp [19] advances single-image 3D reconstruction by decomposing geometry into mechanical properties and enforcing static equilibrium. 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 [18] and DreamSync [51] simply use reject sampling. ReNO [13] focus on adapting a diffusion model during inference by purely optimizing the initial latent noise using a differentiable objective. Some other works [54, 59] leverages DPO [46] 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.

2.3. Benchmarking Image Synthesis Models

Standard metrics like FID [21], IS [48], LPIPS [62], and CLIPScore [20] are commonly used to assess image synthesis models. With model advancements, newer meth-

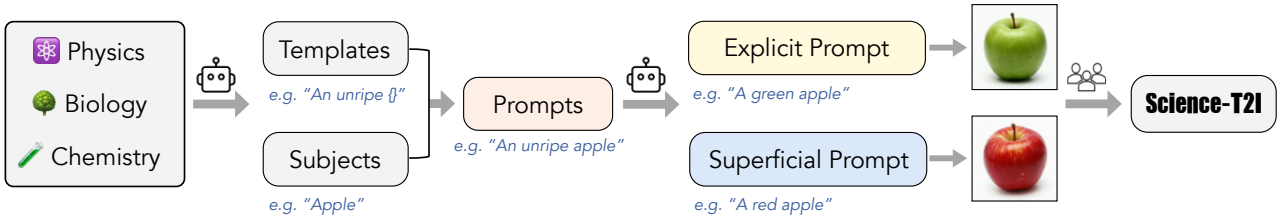

Figure 2. **Data curation pipeline**. For each task, GPT-4o [2] first generates structured templates that capture the scientific principles while allowing for variability in objects or substances. These templates are used to create implicit prompts, which GPT-4o [2] then expands into explicit and superficial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

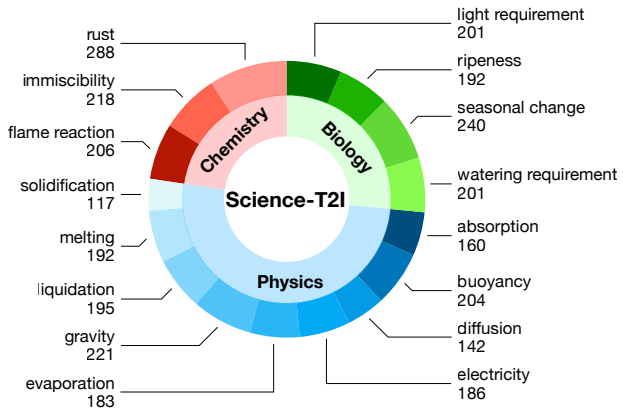

Figure 3. **Data statistics. Science-T2I** is organized into three primary scientific fields: Chemistry, Biology, and Physics. Each field is divided into specific categories, with the numbers indicating the volume of implicit prompts collected for each category.

ods emphasize human evaluation and multimodal LLM-based assessment. HPSv2 [56], PickScore [28] and ImageReward [57] provide human preference annotations, while VQAScore [31], TIFA [23], VIEScore [30], LLM-score [38], and DSG [8] utilize VQA-style evaluations. For object attributes and relationships, benchmarks like T2I-CompBench [24] and CLIP-R-Precision [42] have been introduced. However, there are few benchmarks focusing on the physical commonsense. PhyBench [40] establishes a set of grading criteria and employs vision-language models to discretely score images. In contrast, we introduce **SciScore**, an end-to-end model designed to provide a more refined and continuous scoring mechanism for images.

3. Dataset: Science-T2I

We introduce **Science-T2I**, 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 [10, 29, 34] and preference annotation [28, 56, 57], **Science-T2I** challenges models to perform implicit reasoning based on prompts that need scientific knowledge.

As illustrated in Figure 3, **Science-T2I** consists of 16 tasks that require the model to infer or visualize concepts not explicitly stated in the prompts but rooted in underlying scientific principles. These tasks are inspired by existing research such as PhyBench [40] and Commonsense-T2I [16], 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 [40] 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.

We leverage GPT-4o to generate templates and prompts during data curation. These outputs are then used to drive T2I models for image generation, as shown in Figure 2.

4. Method: SciScore

While CLIP [44] 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 **Science-T2I** that extends CLIP's architecture [44]. **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 [44] by independently encoding a text prompt x and an image y into a shared high-dimensional vector space using separate transformer encoders [52], 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:

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

4.2. Training Techniques

For developing **SciScore** we employed a fine-tuning approach on the CLIP [44] using **Science-T2I**. 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 [41, 50], the predicted preference $\hat{p}_{\text{img}}(x_a \succ x_b; y)$ for prompt x_a over prompt x_b for a given image y is calculated as:

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

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

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

Implicit Prompt Alignment (IPA). Preliminary experiments revealed that the pretrained CLIP model [44] 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 hyper-parameter that controls the relative weight of the image encoder enhancement loss in relation to the implicit prompt alignment loss.

5. Two-Stage T2I Model Fine-Tuning

5.1. Supervised Fine-tuning (SFT)

Current post-training algorithms for diffusion models, such as those utilizing PPO [5, 14] and DPO [53, 58], have significantly advanced model fine-tuning. However, these methods are constrained by the requirement that the optimization objectives remain within the distribution of the pre-trained

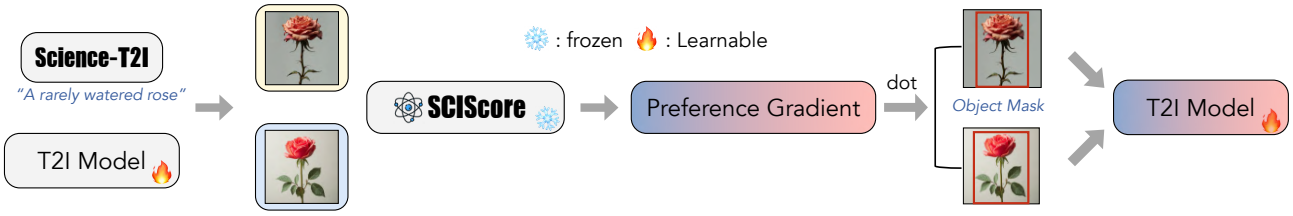

Figure 4. **Online fine-tuning Pipeline.** For each prompt, two images are generated to compute **SciScore** preference metric. Simultaneously, GroundingDINO [37] extracts segmentation masks from these images based on the prompts, which are then used to block gradient propagation in the corresponding spatial regions.

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 **Science-T2I**. 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 [36] framework, the SFT training objective is formulated as:

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

In this formulation, we adopt the same mathematical notation as presented in [12] to ensure consistency.

5.2. Online Fine-tuning (OFT)

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 [5], 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 [36] is typically formulated as an Ordinary Differential Equation (ODE), resulting in a deterministic process. This deterministic formulation complicates the computation of the policy $\pi_\theta(a_t \mid s_t)$:

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

In alignment with the discussion in [11], we can alternatively interpret flow matching [36] as a Stochastic Differential Equation (SDE), which is mathematically formulated as:

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

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

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:

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

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

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 [45]. Specifically, given a condition (typically a prompt) c, we randomly sample two trajectories:

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

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

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:

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

Table 1. Performance comparison of different models on **Science-T2I** S and **Science-T2I** C across different subjects, measured by accuracy in two-choice selection task. **Bold** values indicate the best performance.

Model	**Science-T2I** S				**Science-T2I** C			
	Physics	Chemistry	Biology	Avg.	Physics	Chemistry	Biology	Avg.
CLIP-H [25]	55.08	52.38	55.88	54.69	56.56	44.44	76.67	59.47
BLIPScore [32]	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 [55]	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 [37] 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:

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

6. Experiment: SciScore

6.1. Implementation Details

Training Setting. We fine-tune the CLIP-H model [25] using our framework on **Science-T2I** 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:

- **Science-T2I S** (671 tuples): It matches the training set style, emphasizing simplicity and reasoning regions.
- **Science-T2I C** (227 tuples): It adds complexity through diverse scene settings in prompts and images.

We establish our baseline using three evaluation dimensions: VLMs, LMMs, and human assessments. For VLMs, we utilize CLIP-H [25] and BLIPScore [32, 33]. In the LMM category, we employ `GPT-4o-mini` [2] with CoT reasoning[55]. Human evaluations involved 10 experts with science or engineering degrees. The evaluation involves presenting one implicit prompt alongside two images: one aligned with the corresponding explicit prompt and the other with the superficial prompt. Models and humans are tasked with selecting the image that best corresponds to implicit prompt. Experimental results are presented in Table 1.

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

λ	**Science-T2I** S	**Science-T2I** 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 [44] and BLIPScore [32] 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 [55] does not yield significant improvements in this context. In contrast, **SciScore** not only achieves but surpasses human-level performance on both **Science-T2I** S and **Science-T2I** C, highlighting its superior generalization and efficacy in handling the tasks in a complex scenario. This result underscores the potential of **SciScore** to address 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

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

T2I Model	**Science-T2I** S				**Science-T2I** C			
	SP	EP	IP	ND	SP	EP	IP	ND
Stable Diffusion v1.5 [47]	19.35	26.88	22.37	40.11	22.45	28.19	23.40	16.55
Stable Diffusion XL [43]	21.80	31.90	**25.47**	36.34	26.21	34.22	**30.89**	**58.43**
Stable Diffusion 3 [12]	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	**Science-T2I** S		**Science-T2I** C	
	SciScore	RI	**SciScore**	RI
FLUX.1[dev]	23.56	/	27.26	/
+EP	32.85	/	34.70	/
+SFT	30.00	69.32	31.44	56.18
+SFT+OFT	31.18	**82.02**	32.15	**65.73**

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.

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 **Science-T2I** S and **Science-T2I** C, respectively. We generated two images per prompt using the text-to-image models and calculated the average **SciScore**. The results are illustrated 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 [47] 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 **Science-T2I** using SFT in conjunction with LoRA [22] for 2,000 steps. This process generates LoRA weights intended for subsequent OFT. For the OFT phase, we randomly select 300 implicit prompts from **Science-T2I** 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 [37]. The model is then fine-tuned for approximately 100 steps.

Evaluation Setting. We construct two distinct prompt sets by extracting all implicit prompts from **Science-T2I** S and **Science-T2I** C. For evaluation, we generate five distinct images for each prompt and compute the average **SciScore** across these images to ensure robust results.

7.2. Results

The results in Table 4 demonstrate that both SFT and OFT enhance **SciScore**'s performance. To further investigate the factors driving these enhancements, we employed explicit prompts corresponding to all implicit prompts in **Science-T2I** S and **Science-T2I** C. This approach allowed us to cal-

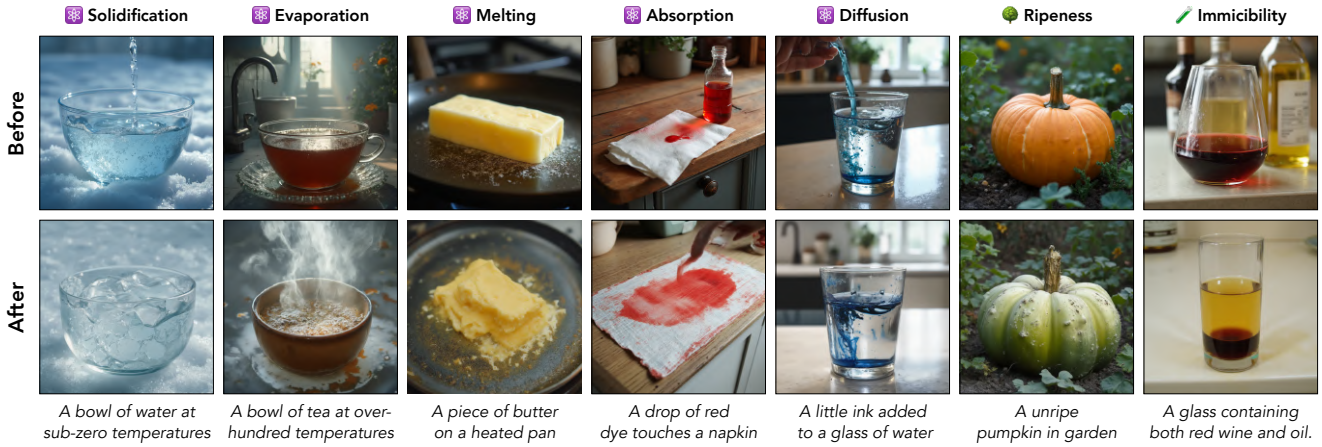

Figure 5. **Case study.** The upper images are generated using the baseline FLUX.1[dev] [1], whereas the lower images are produced with our fine-tuning method. Each image pair utilizes an identical random seed to ensure consistency in comparison. Note that the displayed prompts are summaries of the original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

culate 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 OFT, while the purple line illustrates the case without initial SFT. Both scenarios use identical configurations for OFT. The results demonstrate that initiating OFT with SFT leads to a more stable increase in **SciScore**. In contrast, OFT 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 OFT.

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

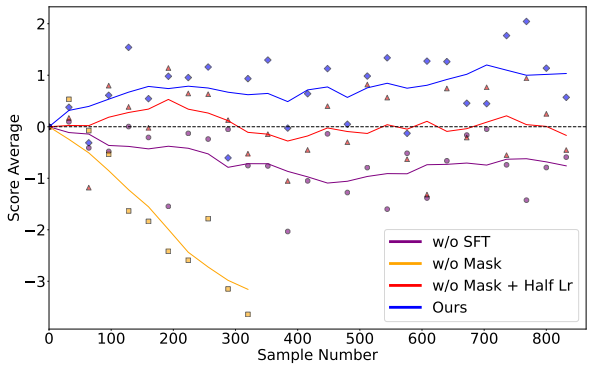

Figure 6. Ablation study of two-stage training. At each step, all prompts in **Science-T2I** S are employed to generate two images per prompt, followed by the calculation of the average **SciScore**. The result illustrates the deviation from the initial baseline.

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.

8. Conclusion

We present **SciScore**, a reward model aimed at integrating scientific knowledge into image synthesis models. Utilizing our expert-annotated dataset, **Science-T2I**, with over 20,000 image pairs and 9,000 prompts in total, we established a framework for evaluating and improving image realism. Our two-stage training approach, featuring supervised fine-tuning and online fine-tuning, led to significant performance improvement. We show that **SciScore** achieves human-level performance in aligning with scientific knowledge.

Acknowledgments

We thank the reviewers for their constructive feedback. We thank Alistair King for sharing insightful code, which was instrumental in our fine-tuning process. SX also acknowledges support from Open Path AI Foundation, Intel AI SRS, IITP grant funded by the Korean Government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project), Amazon Research Award, and NSF Award IIS-2443404.

References

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

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

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

[4] 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

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

[6] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023. 2

[7] 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

[8] 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

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

[10] 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

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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

[19] 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

[20] 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. 2

[21] 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. 2

[22] 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

[23] 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

[24] 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

[25] 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

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

[27] 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

[28] 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

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

[30] 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

[31] 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

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

[33] 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

[34] 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

[35] Yunlong Lin, Tian Ye, Sixiang Chen, Zhenqi Fu, Yingying Wang, Wenhao Chai, Zhaohu Xing, Lei Zhu, and Xinghao Ding. Aglldiff: Guiding diffusion models towards unsupervised training-free real-world low-light image enhancement. *arXiv preprint arXiv:2407.14900*, 2024. 2

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

[37] 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

[38] 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

[39] 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

[40] 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

[41] 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

[42] 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

[43] 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

[44] 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

[45] 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

[46] 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

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

[48] 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. 2

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

[50] 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

[51] 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

[52] 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

[53] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caim-

ing Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization, 2023. 4

[54] 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

[55] 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

[56] 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

[57] 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

[58] 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

[59] 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

[60] 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

[61] 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

[62] 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. 2