Science-T2I: Addressing Scientific Illusions in Image Synthesis

Supplementary Material

The supplementary material is structured as follows:

- Rational to prioritize rewriting capability in Section A.
- Detailed data curation process in Seciton B.
- Comparison with related benchmarks in Section C.
- The training settings for **SciScore** in Section D.
- Setup of **Science-T2I** S and **Science-T2I** C in Section E.
- Detailed baseline setup for **SciScore** in Section F.
- Additional results of **SciScore** in Section G.
- Additional analysis of **SciScore** in Section H.
- Qualitative analysis of IEE in Section I.
- Detailed benchmarking configuration in Section J.
- More results on benchmarking T2I model in Section K.
- More results and analysis of explicit prompt alignment in Section L.
- Details of two-stage training in Section M.
- Detailed two-stage training settings in Section N.
- Additional results of online fine-tuning in Section O.
- Additional observations in Section P.
- Limitations of **SciScore**in Section Q.
- Detailed task descriptions in Section R.

A. Rationale to Prioritize Rewriting Capability

In the development of our study, we considered incorporating additional reasoning tasks, such as reflection-based tasks [8] 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 [25] 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 [21], Stable Diffusion 3 [7], 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.

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 **Science-T2I**, 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 R9,R10,R11, 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. Comparison with Related Benchmarks

In order to provide a comprehensive understanding of how **Science-T2I** distinguishes itself from other existing datasets, we present a detailed comparison in Table C1. The key distinguishing features and advantages of **Science-T2I** are as follows: (1) it enables the direct utilization of **SciScore** for benchmark T2I generation, significantly enhancing efficiency compared to approaches based on LMMs; (2) its test set is uniquely designed to also serve as a benchmark for evaluating LMMs, offering dual functionality; and (3) it includes a large-scale training set that not only supports the training of generative models but also facilitates advancements in multimodal research.

Table C1. Comparison with related benchmarks.

Benchmark	Type	Category	Training Set	Evaluation	
				Generation	LMM
Commonsense-T2I [9]	Commonsense	5	✘	✔	✘
T2I-FactualBench [11]	Commonsense	8	✘	✔	✘
PhyBench [20]	Science	31	✘	✔	✘
Science-T2I (Ours)	Science	16	✔	✔	✔

D. 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 D2.

Table D2. 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 [19]
λ	0.25

E. Setup of Science-T2I S and Science-T2I 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:

- **Science-T2I 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 **Science-T2I** S is to assess the model's performance on data stylistically similar to its training set.
- **Science-T2I C**: This test set challenges the model in more complex scenarios, introducing contextual elements like explicit scene settings and diverse scenarios. Prompts in **Science-T2I** 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.

F. 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 [12] and BLIP-2 [16]. 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 to evaluate the alignment between the text and image pairs.

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.

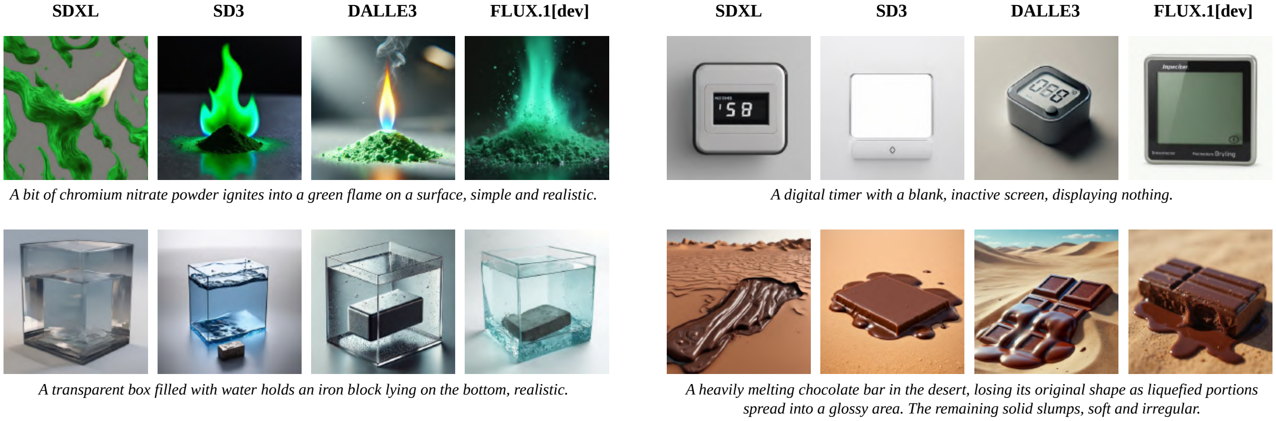
Superficial Prompt: Construct an explicit interpretation of the input prompt that disregards the underlying scientific reasoning or implied elements. Focus only on the superficial or literal descriptive aspects.

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].



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 [21], SD3 [7], 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.

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[26] and another incorporating CoT [26] 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 F3.

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.

G. Additional Results of SciScore

In this section, we extend Table 1 by providing detailed accuracy metrics for each category in Tables G4 and G3, which allows for a more nuanced evaluation of **SciS-**

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 F3. **Instruction For GPT Evaluation.** Text segments in red are specifically incorporated to facilitate CoT [26] reasoning.

Table G3. Performance comparison on **Science-T2I** 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 [12]	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 [15]	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 [26]	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**

core'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.

H. Additional Analysis

In this section, we present further in-depth analysis pertaining to the results and observations discussed in Section 6.

Performance of VLMs Approaches Random Guessing. Both CLIP-H [10] and BLIPScore [15] 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, Figure H4 provides a comparative analysis of the ROC curves for **SciScore**, CLIP-H [12], and BLIPScore [15], 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 [26]. 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 [26] framework suggest a fundamental challenge in maintaining logical consistency when processing and synthesizing information from visual sources, especially when dealing with com-

Table G4. Performance comparison on **Science-T2I** 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 [12]	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 [15]	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 [26]	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**

plex or ambiguous data. To substantiate these claims, we present qualitative results in Figure H5, 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.

I. Qualitative Analysis of IEE

Qualitative results, as shown in Figure I6, 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 subtle 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.

J. 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 J5 summarizes the configuration parameters used for each model, including the guidance scale and the number of inference steps.

Table J5. Configurations of each T2I model

T2I Model	Guidance Scale	Inference Step
SDv1.5 [24]	7.5	50
SDXL [21]	5.0	50
SD3 [7]	7.0	28
FLUX.1[schnell] [1]	0.0	4
FLUX.1[dev] [1]	0.0	30

K. More Results on Benchmarking T2I Model

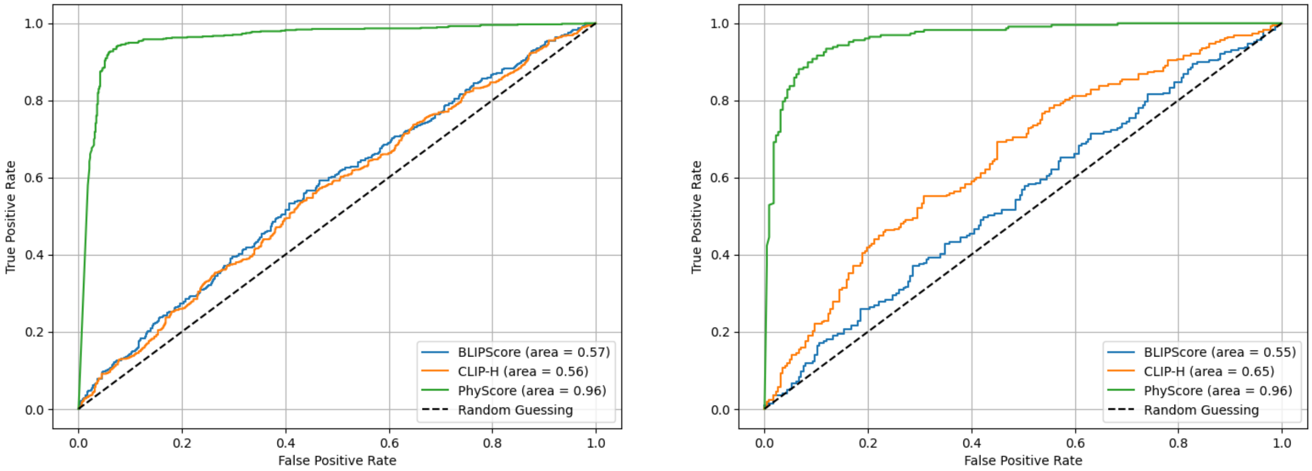
In this section, we further employ **SciScore** to benchmark additional state-of-the-art text-to-image models. Due to budgetary constraints, our evaluation is limited to open-source models. The results are presented in Table K6.

L. 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 [22]: 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 [20], our grading process is divided into two distinct aspects:

- **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 Q7a. After establishing the grading criteria, we selected all implicit prompts and their corresponding explicit prompts from **Science-T2I** S and **Science-T2I** C. Using



(a) **Science-T2I**S

(b) **Science-T2I**C

Figure H4. **ROC Curve Analysis.** The AUC scores for both BLIPScore [15] and CLIP-H [10] 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.

Table K6. Performance of T2I Models on **SciScore**.

T2I Model	Size	**Science-T2I** S				**Science-T2I** 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

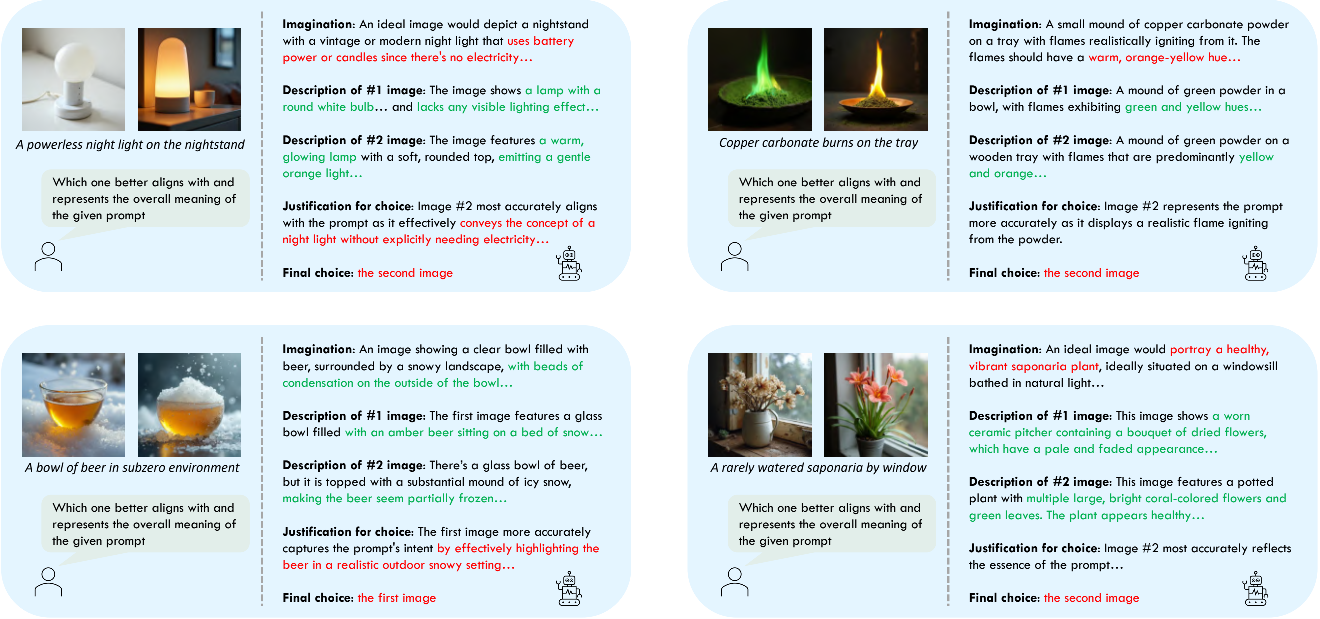
text-to-image models, we generated two images for each explicit prompt. These images were then evaluated by GPT-4o-mini [2] following the instructions detailed in Figure Q7b. This evaluation produced average scene scores and reality scores with the experimental results summarized in Table L8. To further 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 L8 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 Q7b, to **Science-T2I** S and **Science-T2I** 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:

$$HC_{SS} = \frac{SS}{2.0} \times 100 \tag{1}$$

where SS is the scene score assigned by GPT-4o-mini. For reality scores, we used two separate formulas to com-



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…

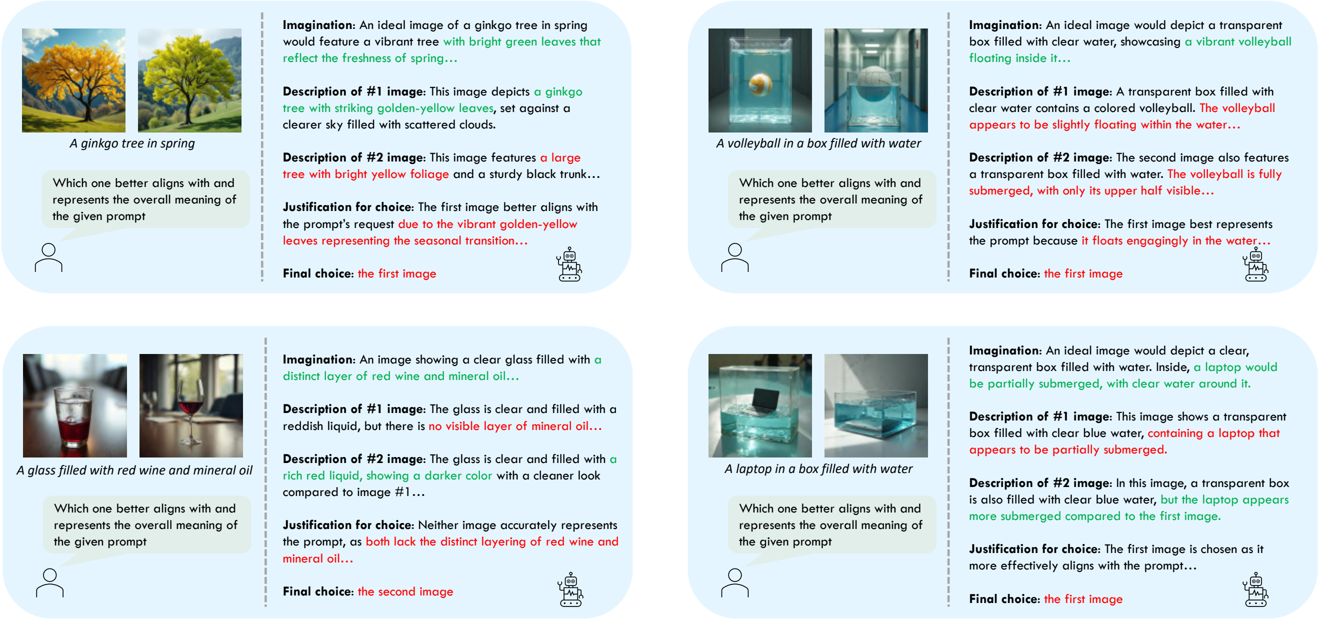
Description of #1 image: This image shows a worn ceramic pitcher containing a bouquet of dried flowers, which have a pale and faded appearance…

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 H5. **Qualitative Failure Cases of GPT**. In both cases, the CoT [26] reasoning approach from Figure F3 is applied, but errors in either interpretation or visual comprehension impact the final decision. Green text indicates correct inference, while red text marks errors.

pute the correspondence for explicit and superficial images. Specifically, for superficial images (SI), the human correspondence for reality score is calculated as:

$$HC_{RS}^{SI} = \left(1 - \frac{RS}{3.0}\right) \times 100 \tag{2}$$

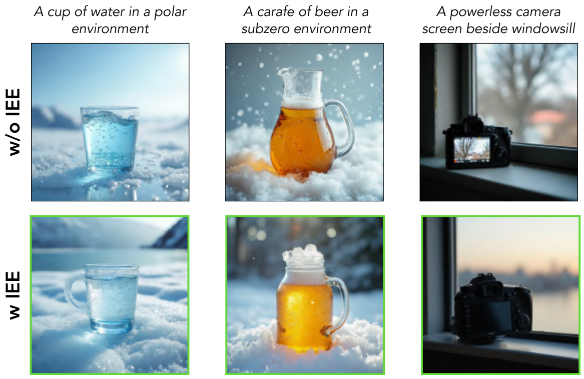

A cup of water in a polar environment
A carafe of beer in a subzero environment
A powerless camera screen beside windowsill
w/o IEE
w IEE

Figure I6. **Qualitative Analysis of IEE.** Images enclosed by green borders denote the correct selection in each pair.

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

$$HC_{RS}^{EI} = \frac{RS}{3.0} \times 100 \tag{3}$$

The comparative results are shown in Table L7.

Table L7. **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$_{\text{SS}}$	RS	HC$_{\text{RS}}$
Science-T2I S	EI	1.827	91.13	2.731	91.03
	SI	1.635	81.74	0.476	84.13
Science-T2I C	EI	1.855	92.73	2.490	83.00
	SI	1.630	81.50	0.636	78.79

M. 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 [17] 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):

$$\frac{dx_t}{dt} = v(x_t, t), \tag{4}$$

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

$$x_t = \alpha_t x_0 + \beta_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{5}$$

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:

$$v(x, t) = \dot{\alpha}_t \mathbb{E}[x_* | x_t = x] + \dot{\beta}_t \mathbb{E}[\epsilon | x_t = x], \tag{6}$$

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

$$\mathcal{L}_{SFT}(\theta) := \mathbb{E}_{x_*, \boldsymbol{\epsilon}, t}\left[\left\|v_\theta(x_t, t) - \dot{\alpha}_t x_* - \dot{\beta}_t \epsilon\right\|^2\right] \tag{7}$$

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:

$$\max_{p_\theta} \mathbb{E}_{c, x_0 \sim p_\theta(x_0|c)}\left[r(c, x_0)\right] - \beta \mathcal{D}_{\text{KL}}\left[p_\theta(x_0|c) \| p_{\text{ref}}(x_0|c)\right] \tag{8}$$

where the hyper-parameter β controls regularization. According to [23], the unique global optimal solution p_θ^* to this optimization problem is given by:

$$p_\theta^*(x_0|c) = p_{\text{ref}}(x_0|c) \exp\left(\frac{r(c, x_0)}{\beta}\right) / Z(c) \tag{9}$$

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:

$$r(c, x_0) = \beta \log \frac{p_\theta^*(x_0|c)}{p_{\text{ref}}(x_0|c)} + \beta \log Z(c) \tag{10}$$

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

$$p_{BT}(x_0^w \succ x_0^l | c) = \sigma(r(c, x_0^w) - r(c, x_0^l)) \tag{11}$$

where σ is the sigmoid function, x_0^w is the preferred outcome, and x_0^l 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:

$$L_{BT}(\phi) = \mathbb{E}_{c, x_0^w, x_0^l}\left[\log \sigma\left(r_\phi(c, x_0^l) - r_\phi(c, x_0^w)\right)\right] \tag{12}$$

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

$$\begin{aligned} \mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{c, x_0^w, x_0^l}\Bigg[&\log\sigma\Bigg(\beta \log \frac{p_\theta(x_0^w|c)}{p_{\text{ref}}(x_0^w|c)} \\ &- \beta \log \frac{p_\theta(x_0^l|c)}{p_{\text{ref}}(x_0^l|c)}\Bigg)\Bigg] \end{aligned} \tag{13}$$

Table L8. **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	**Science-T2I** S				**Science-T2I** C			
	SS	PoE	FS	PoE	SS	PoE	FS	PoE
Stable Diffusion v1.5 [24]	1.298	64.90	2.470	49.40	1.261	63.05	2.446	48.92
Stable Diffusion XL [21]	1.718	85.90	3.510	70.20	1.679	83.95	3.360	67.20
Stable Diffusion 3 [7]	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

Choice of σ_t. We determine the value of σ_t by adhering to the methodology presented in [13]. Initially, we define the hyperparameters S_{churn}, $S_{\min}$, $S_{\max}$, and S_{noise}. Subsequently, we define γ_t as follows:

$$\gamma_t = \begin{cases} \min\left(S_{\text{churn}} \cdot \Delta t, \sqrt{2} - 1\right) & \text{if } t \in [S_{\min}, S_{\max}] \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

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

$$\sigma_t = S_{\text{noise}} \cdot \sqrt{\gamma_t^2 + 2\gamma_t} \cdot (1 - t). \tag{15}$$

Pre-Training Subject Extraction. We integrate GroundingDINO [18] 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 [18] during training to generate corresponding masks. These masks are then utilized to apply gradient masking.

Gradient Masking. The mask generated by GroundingDINO [18] is derived from the resolution of the RGB image. However, gradients are computed within the model's latent space, as detailed in LDM [24]. The connection between the RGB image and the latent space is facilitated by a pretrained Variational Autoencoder (VAE) [14], 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:

$$\left(\frac{x_1}{H} \cdot H_l,\ \frac{y_1}{W} \cdot W_l,\ \frac{x_2}{H} \cdot H_l,\ \frac{y_2}{W} \cdot W_l\right) \tag{16}$$

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.

N. 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 N9.

Table N9. 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 [19]	AdamW [19]
gradient accumulation	8	2
LoRA rank	16	16
S_{churn}	/	0.1
$S_{\min}, S_{\max}$	/	$0, \infty$
S_{noise}	/	1.0
β	/	10

O. 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 [27] 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 N9, 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 [27] 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 [6], 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 O10.

Table O10. **Comparison of Average Rewards.** The online fine-tuning approach consistently outperforms the baseline, demonstrating the effectiveness of the proposed algorithm.

Method	LAION [4]	ImageReward [27]
FLUX.1[schnell]	5.855	0.949
+OFT	**6.074**	**1.023**

P. 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 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.”

Q. Limitations

Building upon the concepts introduced in Section P, 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 G4 and Table G3.

R. 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 **Science-T2I**. Additionally, illustrative examples from **Science-T2I** are presented in Figure R9,R10,R11 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.

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 [20], 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.

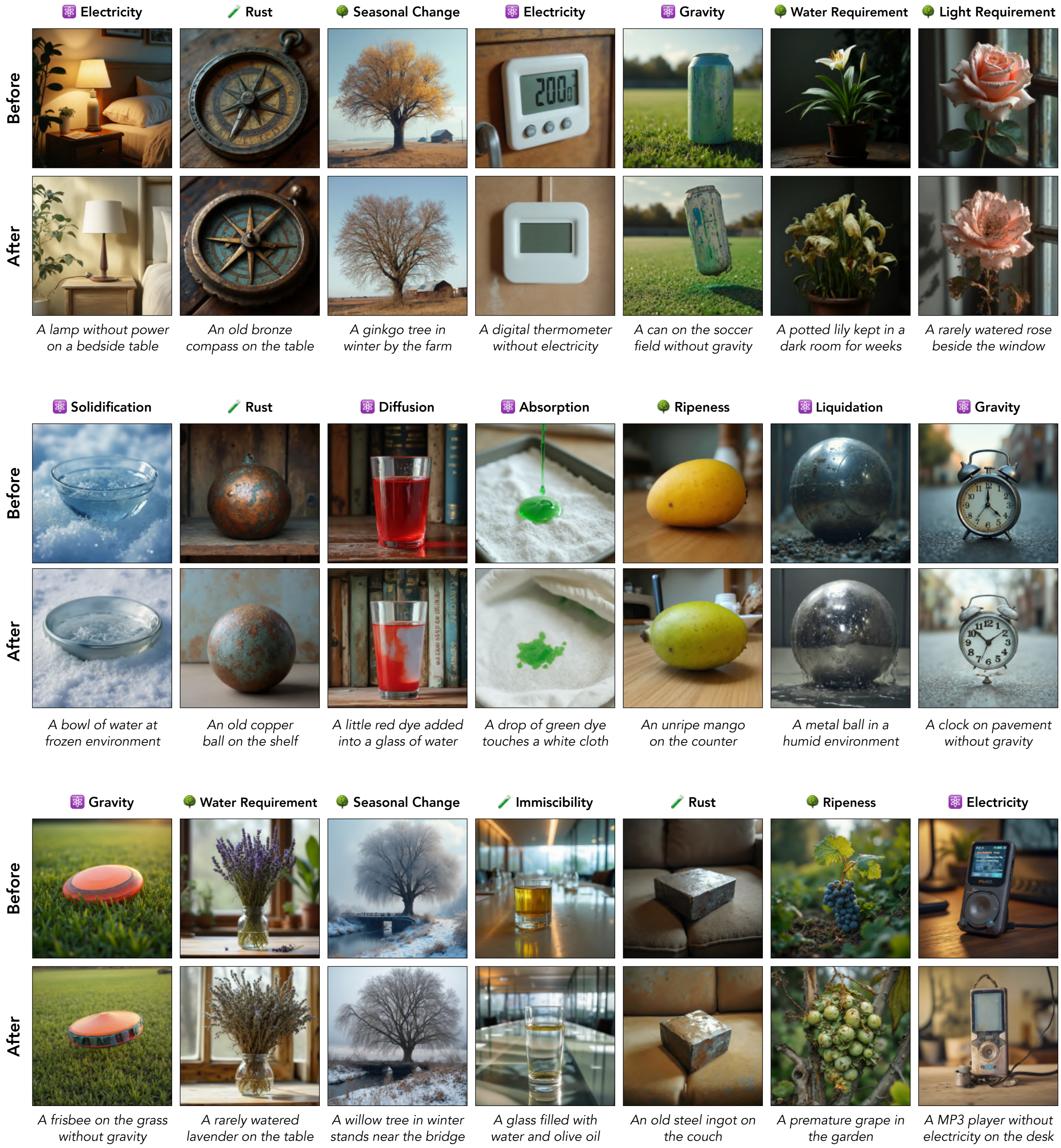
User Prompt

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 Q7. 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 Q8. **Additional Generated Samples.** Each pair of images is produced using the same random seed to ensure consistency.

- **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.

References

[1] Flux. https://blackforestlabs.ai/. 2, 3, 5, 9, 10

[2] Gpt-4o. https://openai.com/index/hello-gpt-4o/. 1, 3, 4, 5, 6, 7

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

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

[5] Dalle-3. https://openai.com/index/dall-e-3/. 2, 3

[6] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning, 2024. 10

[7] 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. 1, 3, 5, 9

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

[9] Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. Commonsense-t2i challenge: Can text-to-image generation models understand commonsense?, 2024. 2

[10] 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. 4, 6

[11] Ziwei Huang, Wanggui He, Quanyu Long, Yandi Wang, Haoyuan Li, Zhelun Yu, Fangxun Shu, Long Chan, Hao Jiang, Leilei Gan, and Fei Wu. T2i-factualbench: Benchmarking the factuality of text-to-image models with knowledge-intensive concepts, 2024. 2

[12] 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. 2, 4, 5

[13] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. 9

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

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

[16] 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. 2

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

[18] 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. 9

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

[20] 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. 2, 5, 11

[21] 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, 3, 5, 9

[22] 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. 5

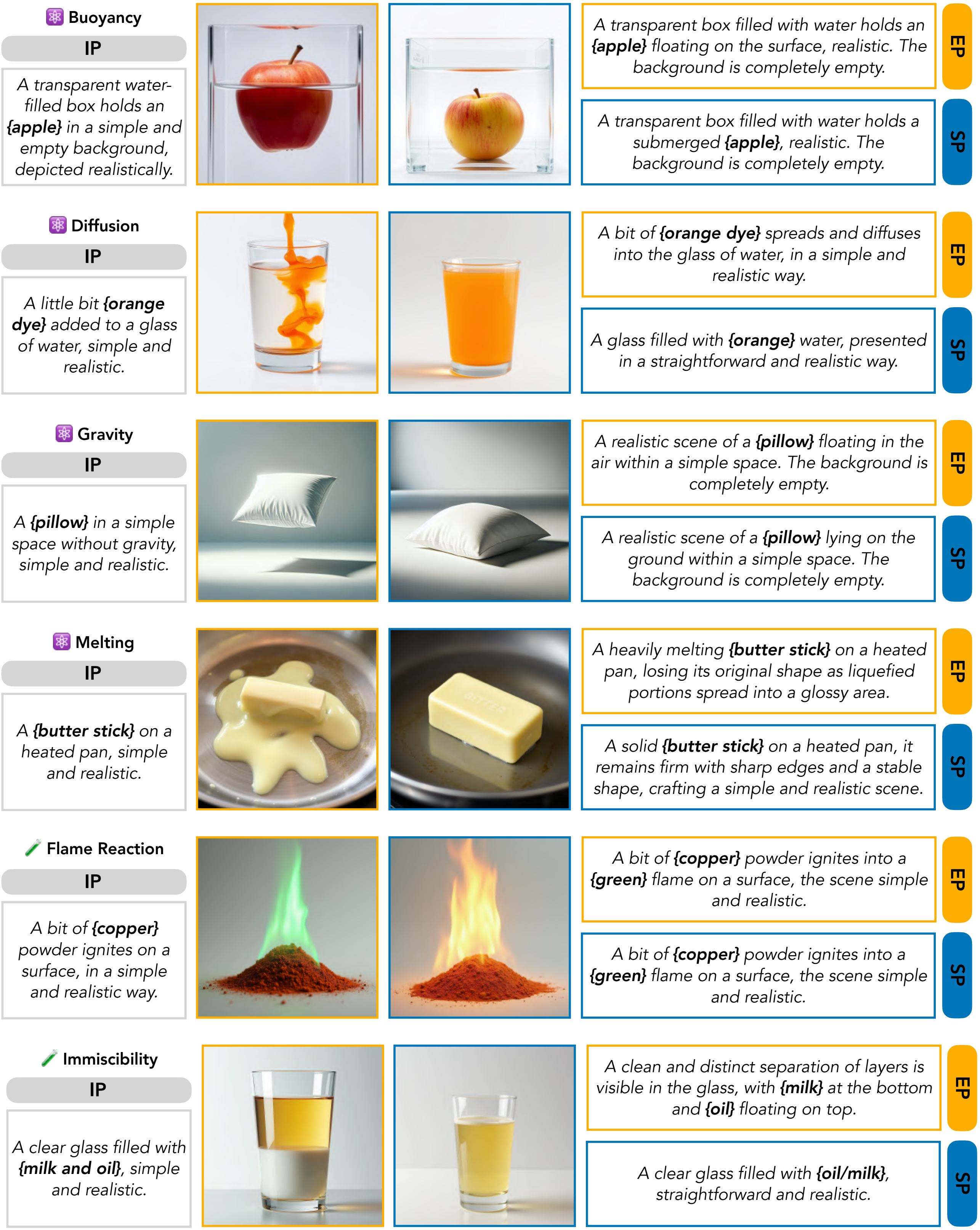
[23] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. 8

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

[25] 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

[26] 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. 3, 4, 5, 7

[27] 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. 10



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

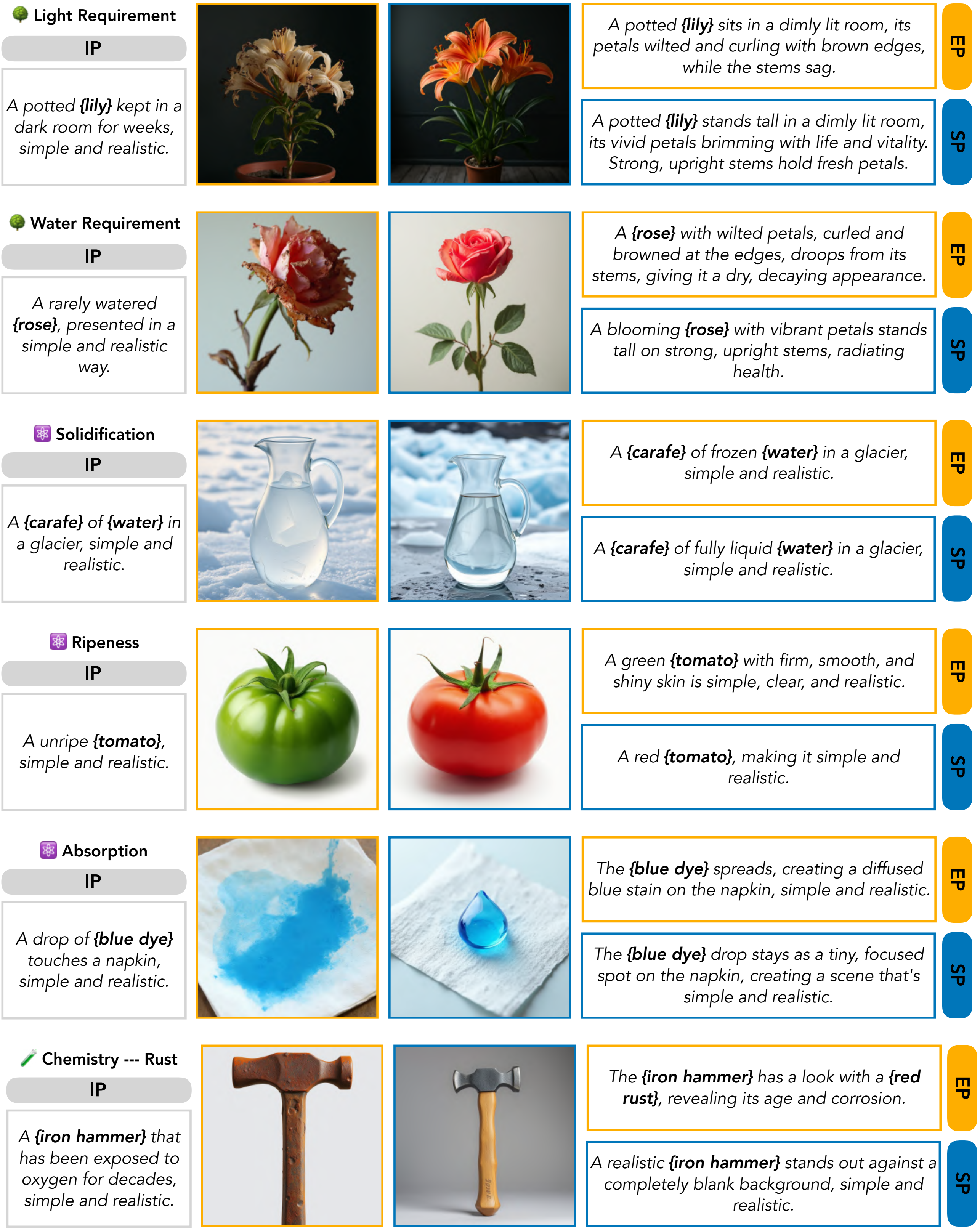
IP

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 R9. **Several examples from Science-T2I.** 'EP' denotes explicit prompts (yellow blocks), 'SP' denotes superficial prompts (blue blocks), and 'IP' denotes implicit prompts (grey blocks).



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

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

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

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

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

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

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

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

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

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

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

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

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

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

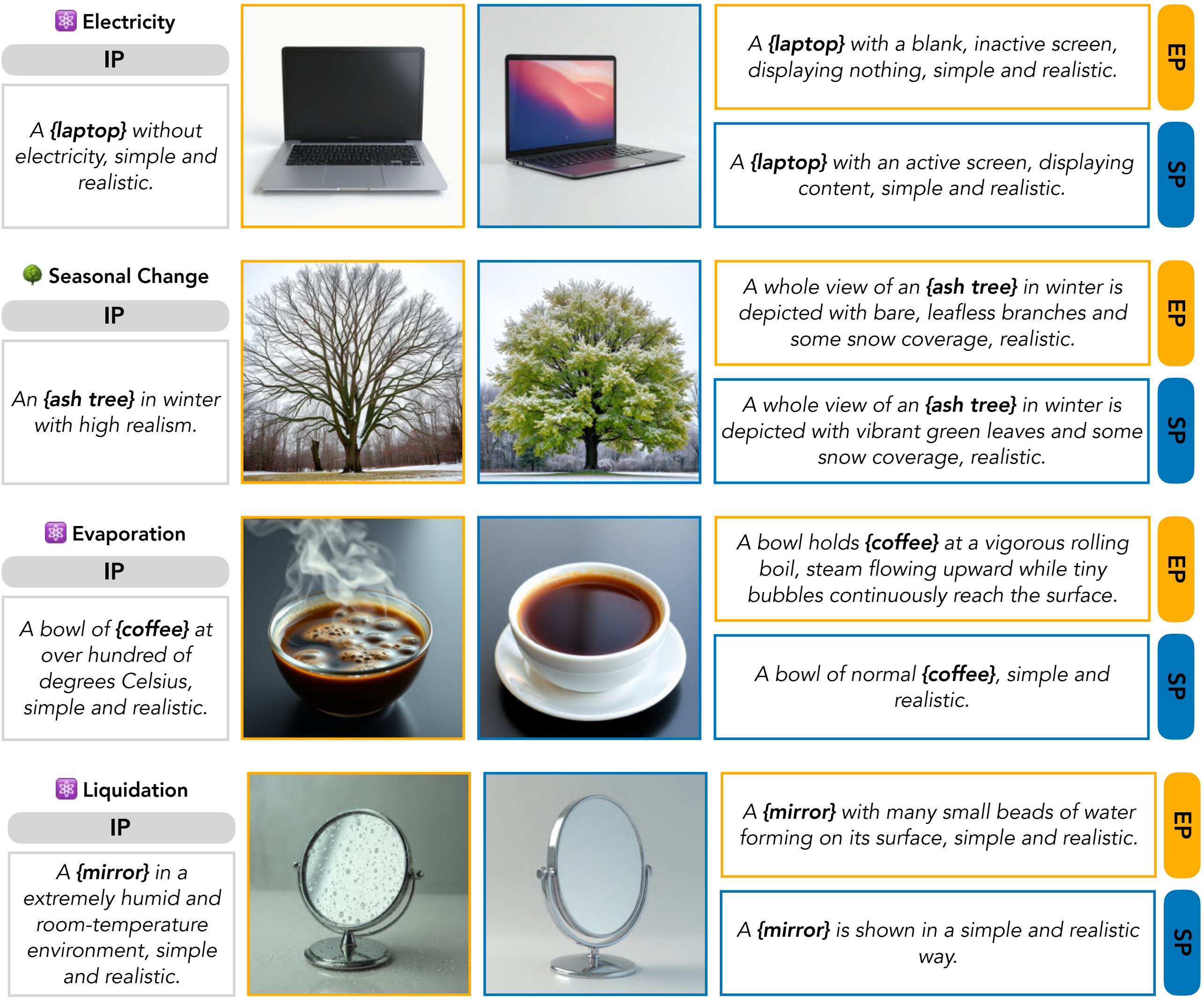
The ***{blue dye}*** *drop stays as a tiny, focused spot on the napkin, creating a scene that's simple and realistic.*

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

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

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

Figure R10. **Several examples from Science-T2I.** 'EP' denotes explicit prompts (yellow blocks), 'SP' denotes superficial prompts (blue blocks), and 'IP' denotes implicit prompts (grey blocks).



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

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

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

*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.*

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

*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.*

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

*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.*

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

Figure R11. **Several examples from Science-T2I.** ’EP’ denotes explicit prompts (yellow blocks), ’SP’ denotes superficial prompts (blue blocks), and ’IP’ denotes implicit prompts (grey blocks).