



# TIIF-Bench: How Does Your T2I Model Follow Your Instructions?

Xinyu Wei<sup>1,4</sup>, Jinrui Zhang<sup>1,4</sup>, Zeqing Wang<sup>2,4</sup>, Hongyang Wei<sup>3,4</sup>, Zhen Guo<sup>1,4</sup>, Lei Zhang<sup>1,4</sup>

<sup>1</sup>Hong Kong Polytechnic University    <sup>2</sup>Sun Yat-sen University

<sup>3</sup>Tsinghua University    <sup>4</sup>OPPO Research Institute

allen\_wei@stu.pku.edu.cn, cslzhang@comp.polyu.edu.hk

## Abstract

The rapid advancements of Text-to-Image (T2I) models have ushered in a new phase of AI-generated content, marked by their growing ability to interpret and follow user instructions. However, existing T2I model evaluation benchmarks fall short in limited prompt diversity and complexity, as well as coarse evaluation metrics, making it difficult to evaluate the fine-grained alignment performance between textual instructions and generated images. In this paper, we present **TIIF-Bench** (**T**ext-to-**I**mage **I**nstruction **F**ollowing **Benchmark), aiming to systematically assess T2I models' ability in interpreting and following intricate textual instructions. TIIF-Bench comprises a set of 5000 prompts organized along multiple dimensions, which are categorized into three levels of difficulties and complexities. To rigorously evaluate model robustness to varying prompt lengths, we provide a short and a long version for each prompt with identical core semantics. Two critical attributes, *i.e.*, text rendering and style control, are introduced to evaluate the precision of text synthesis and the aesthetic coherence of T2I models. In addition, we collect 100 high-quality designer level prompts that encompass various scenarios to comprehensively assess model performance. Leveraging the world knowledge encoded in large vision language models, we propose a novel computable framework to discern subtle variations in T2I model outputs. Through meticulous benchmarking of mainstream T2I models on TIIF-Bench, we analyze the pros and cons of current T2I models and reveal the limitations of current T2I benchmarks. Project Page: [https://a113n-w3i.github.io/TIIF\\_Bench/](https://a113n-w3i.github.io/TIIF_Bench/).**

## 1 Introduction

Text-to-Image (T2I) generation has emerged as a cornerstone of multimodal AI, enabling the translation of abstract textual concepts into detailed visual content, advancing applications from digital art to scientific visualization. Recent T2I models can be categorized into two main paradigms. Diffusion-based methods—exemplified by Stable Diffusion [1, 2], PixArt [3, 4, 5], Playground [6, 7], FLUX [8], SANA [9, 10, 11], and others[12, 13, 14, 15]—leverage U-Net or Diffusion-Transformer backbones to iteratively denoise Gaussian noises into photorealistic images, delivering superior visual fidelity and diversity. On the other hand, autoregressive (AR) approaches such as LlamaGen [16], Janus [17, 18], Infinity/VAR [19, 20] and other influential open-source efforts [21] treat images as token sequences, employing next-token prediction or scale-progressive generation to synthesize images. Recent works [22, 23, 24, 25] follow the same AR tokenization paradigm but additionally apply reinforcement-learning optimization to further enhance generation quality and controllability. Very recently, commercial T2I models such as GPT-4o [26], Imagen 3 [27] and MidJourney v7 [28] have propelled T2I to a new level. In particular, GPT-4o demonstrates powerful instruction-following

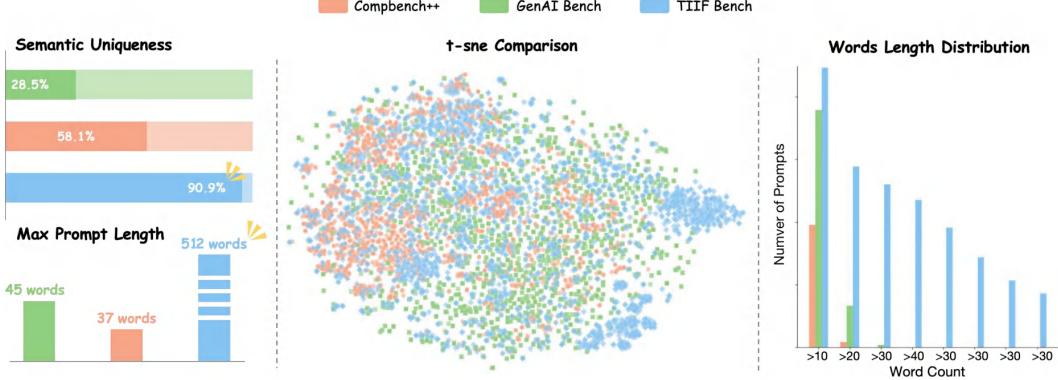


Figure 1: Prompt diversity/complexity of **TIIF-Bench** compared to prior benchmarks. (**Left**) Semantic uniqueness after de-duplication using a cosine similarity threshold of 0.85: less than 30% and 60% prompts in COMPBENCH++ and GENAI BENCH are unique, while more than 90% prompts in TIIF-Bench are unique. (**Middle**) t-SNE visualization of CLIP text embeddings shows that TIIF-Bench spans a much broader semantic space than existing benchmarks. (**Right**) Prompts length in COMPBENCH++ and GENAI BENCH are fixed and short, while TIIF-Bench covers a much wider range. As shown in the lower-left corner, the longest prompts in COMPBENCH++ and GENAI BENCH are less than 50 words, while TIIF-Bench contains complex prompts exceeding 500 words.

capability, which can understand highly intricate prompts and return visually precise, stylistically coherent images in a single conversational loop.

With the rapid development of T2I techniques, how to comprehensively evaluate the performance, especially the instruction-following capability, of modern T2I models has become an important issue. Existing research on the performance evaluation of T2I models generally falls into two complementary categories. **Preference alignment approaches**, such as VQASCORE [29], HPSv2 [30], and VISIONREWARD [31], leverage learned reward models to align with human preferences. **Benchmark-driven approaches**, exemplified by COMPBENCH++ [32], GENEVAL [33] and GENAI BENCH [34], employ structured prompt sets across compositional dimensions (*e.g.*, object attributes, relations, numeracy, *etc.*) and CLIP-based metrics for evaluation.

While existing benchmarks have advanced T2I evaluation, they exhibit several limitations. First, prompts are typically short and of fixed length, as illustrated in Fig.1 (right). However, some T2I models are sensitive to prompt length (see Fig. 2), a factor that current benchmarks fail to account. Second, current benchmarks suffer from semantic redundancy, as shown in Fig. 1 (left), limiting their ability to test generalization across diverse concepts. Third, they often lack syntactic variety and exhibit narrow lexical coverage, as illustrated in Fig. 1 (middle). Finally, with regard to evaluation methodology, commonly used expert scorers, such as CLIP, are insufficient for capturing the fine-grained alignment between images and instructions (see Fig. 3). While some benchmarks [32] leverage large vision-language models (VLMs) as evaluators, their queries are typically too coarse-grained to fully exploit the rich semantic understanding embedded within VLMs (see Fig. 4).

To remedy these gaps, we introduce **TIIF-Bench** (Text-to-Image Instruction Following Benchmark), a benchmark built for the fine-grained assessment of T2I models. We extract ten concept pools from existing benchmarks and define 36 novel combinations of them with six compositional prompt dimensions. Each dimension incorporates multiple attributes, ensuring that every prompt is semantically distinct and exhibits diverse sentence structures. Additionally, two important dimensions previously overlooked, *text rendering* and *style control*, are introduced as dedicated categories in our TIIF-Bench. We also collect 100 real-world designer-level prompts that encode rich human priors and aesthetic judgment. For each prompt, we provide a concise version and an extended version to assess the sensitivity of T2I models to prompt length. In total, the benchmark offers  $6_{\text{combined\_dimension}} \times (300_{\text{short}} + 300_{\text{long}}) + 1_{\text{text\_generation}} \times (300_{\text{short}} + 300_{\text{long}}) + 1_{\text{style\_control}} \times (300_{\text{short}} + 300_{\text{long}}) + 1_{\text{designer\_level}} \times (100_{\text{short}} + 100_{\text{long}}) = 5000$  prompts. In evaluation, each prompt is accompanied by a set of attribute-specific yes/no questions, enabling VLMs to judge at a more granular level than a coarse score. Text rendering accuracy is further quantified by the proposed GNED metric (see Section 3.3 for details).

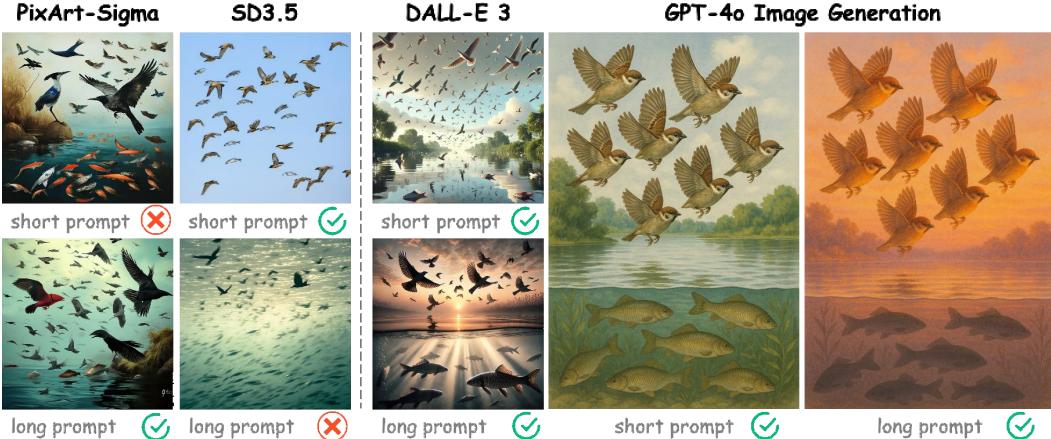


Figure 2: Prompt-length sensitivity on the *Numeracy* attribute (“more birds than fish”) from **TIIF Bench**. Short prompt: “*The birds are more numerous than the fish.*” Long prompt: “*The birds, with their feathers catching the gentle light of dawn, vastly outnumber their aquatic counterparts, the fish, which glide silently beneath the rippling surface of the water; their sleek forms moving like shadows in the depths below.*” We observe that PixArt-Sigma and SD 3.5 exhibit noticeable sensitivity to prompt length, with their ability to follow the core semantic instruction varying between the short and long versions of the same prompt. In contrast, DALL-E 3 and GPT-4o demonstrate strong robustness, maintaining consistent instruction-following performance regardless of prompt length.

We benchmark popular closed-source models, including GPT-4o [26], FLUX.1-Pro [8], DALLE-3 [35], MidJourney V6 and V7 [28], alongside leading open-source models, including Stable Diffusion (SD) series [1, 2], PixArt series [5, 4, 3], SANA series [10, 9], Playground series [6, 7], and autoregressive pipelines [16, 19, 20, 36, 37]. Our findings indicate that, with the exception of GPT-4o, both open-source and closed-source models perform relatively well on prompts involving object attributes (*e.g.*, color, texture, shape), yet consistently struggle with spatial tasks and reasoning tasks such as 2D/3D layouts and logical instructions. Moreover, models that achieve higher overall scores on TIIF-Bench tend to be more robust to prompt length, whereas lower-performing models exhibit greater sensitivity. This suggests a positive correlation between a model’s instruction comprehension and its image generation quality. Finally, although AR-based models generally produce lower-fidelity images, their instruction-following performance is comparable to that of advanced diffusion-based models, highlighting the inherent advantage of autoregressive architectures in semantic understanding.

The contributions of this paper are summarized as follows:

- (i) **Prompt-design methodology.** We identify critical limitations of existing benchmarks (fixed-length prompts, high semantic redundancy, and limited syntactic diversity) and propose a novel attribute-composition strategy to address these issues. Additionally, we introduce new evaluation dimensions and vary prompt length to test T2I model robustness.
- (ii) **Fine-grained evaluation protocol.** We propose an assessment framework that leverages the world knowledge of VLMs to pose attribute-specific yes/no queries, achieving finer alignment with human preference than previous metrics.
- (iii) **Important empirical insights.** Through extensive experiments on TIIF-Bench, we uncover consistent patterns in how T2I models of different architectures follow instructions during image generation, providing important insights for future research.

## 2 Related Work

Text-to-image (T2I) models have made remarkable strides in producing high-quality images. The recent debut of GPT-4o image generation, in particular, marks a major leap forward in a model’s ability to understand and follow complex, lengthy user instructions. However, objectively assessing this progress—especially a model’s **instruction-following capability**—remains challenging. Current evaluation efforts fall into two main categories: (i) scoring systems that measure how well generated



Figure 3: Illustration of the limitations of expert scorers widely used in T2I benchmarks such as COMPBENCH++ and GENAI BENCH. CLIP and BLIP can only assess image-level alignment and they struggle with evaluating the fine-grained instruction-following capability of T2I models, often producing scores misaligned with human judgments. UniDet fails in most cases, unable to detect objects in complex scenes produced by modern T2I models. VQAScore, a variant of CLIP fine-tuned for image quality assessment, remains limited in addressing the complexity of modern generations. As a result, the best models selected by these expert scorers are different from the one selected via human subjective evaluation (*i.e.*, GPT-4o).

images align with human visual preferences, and (ii) structured benchmarks that provide dimension-specific tests of model performance.

**Scoring Methods Aligned with Human Preferences.** CLIPSCORE [38], a widely adopted metric, fails to provide reliable judgments on complex prompts due to CLIP’s inherent bag-of-words behavior. To address this, VQASCORE [29] finetunes a CLIP-FlanT5 architecture specifically for the task of T2I image quality evaluation. Similarly, HPDV2 [30] mitigates the collection bias in human image preference data. When used to fine-tune CLIP, it results in HPSv2, a model with enhanced alignment to human judgments. However, these approaches remain fundamentally constrained by CLIP’s limitation, *i.e.*, it can assess high-level image–text alignment but struggle with fine-grained semantic correspondence. VISIONREWARD [31] attempts to address this by learning a fine-grained, multi-dimensional reward model via hierarchical visual assessment and interpretable weighting. Yet, its primary focus lies in video quality evaluation, limiting its applicability to static T2I tasks.

**Benchmark-Driven Evaluation Frameworks.** COMPBENCH++ [32] provides an 8k prompts suite across four compositional themes (attribute binding, object relations, numeracy, complex scenes) and augments it with detection-based 2D-/3D-spatial/numeracy metrics plus VLM-assisted scoring. GENAI-BENCH [29] adds 1.6k prompts aiming for broader reasoning coverage, while GENEVAL [33] uses object-detection signals to dissect co-occurrence, layout, counts and colors. Despite their scope, these benchmarks [39] share key drawbacks: many prompts are short, templated, and highly repetitive, limiting semantic and linguistic diversity; and their evaluation protocols, including rule-based object detection and single-score VLM judgments, struggle to capture semantic fidelity and often correlate poorly with human preference. Motivated by these observations, we introduce TIIF-Bench, a hierarchically structured benchmark featuring semantically rich prompts, broader prompt

According to the image and your previous answer, evaluate how well the image aligns with the text prompt: {xxx}. Give a score from 0 to 100, according the criteria:  
...  
 Provide your analysis and explanation in JSON format with the following keys: score (e.g., 85), explanation (within 20 words).



a backpack, three apples, two strawberries, and a bowl were on the kitchen counter

The black camera was next to the white tripod.

eight bottles

"score": 90  
 "explanation": "The text correctly identifies all objects with an accurate spatial layout but misses multiple strawberries on counter."

"score": 90  
 "explanation": "Camera is black and next to a white tripod, matching text description accurately."

"score": 85  
 "explanation": "Image shows ten bottles, not eight as mentioned in the text."

Figure 4: Failure cases of current T2I evaluation methods under coarse prompt queries. Since the image caption is used as part of the prompt, these methods can easily induce hallucinations of VLMs such as GPT-4o and consequently lead to highly optimistic scores.

length variation, an expanded lexical scope, and a suite of fine-grained evaluation metrics, offering a more comprehensive and reliable assessment of T2I models’ instruction-following capabilities.

### 3 Construction of TIIF-Bench

#### 3.1 Limitations of Current T2I Evaluation Benchmarks

As T2I models evolve, how to accurately evaluate their ability to follow natural language instructions has become an increasingly important problem. The existing main benchmarks, including COMPBENCH++ [32], GENEVAL [33] and GENAI BENCH [34], decompose the evaluation of instruction-following into interpretable dimensions, including object attributes, inter-object relations, and a spectrum of reasoning skills (*e.g.*, counting, comparative, and logical inferences). However, we observe that these benchmarks still suffer from several key limitations, which can be categorized as **prompt-design flaws** and **evaluation-protocol flaws**.

**Prompt Design Flaws.** The prompt-design flaws can be classified into three main types. First, *fixed and short prompt lengths*. In COMPBENCH++, the 2D and 3D dimensions include a total of 2,000 prompts; however, these prompts span only four lengths (5, 6, 7, and 8 words). As illustrated in Fig. 2, some T2I models show sensitivity to prompt length. Second, *high semantic redundancy*. We extract CLIP text embeddings for all prompts and compute their pairwise cosine similarity. Prompts with similarity above a threshold of 0.85 are considered semantically duplicated. As shown in Fig. 1 (right), less than 30% of the prompts in COMPBENCH++ and 60% in GENAI BENCH are unique after removal of duplication. This indicates a substantial degree of semantic redundancy, significantly limiting their coverage of instruction semantic space. Third, *poor lexical diversity*. Fig. 1 (middle) visualizes prompt embeddings projected into  $\mathbb{R}^2$  via t-SNE. Due to the widespread use of templated phrasing in COMPBENCH++, for example, the 2D relationship dimension consistently uses fixed expressions such as “on the left of”, exhibiting low syntactic diversity. As a result, their embeddings form a compact cluster with limited dispersion.

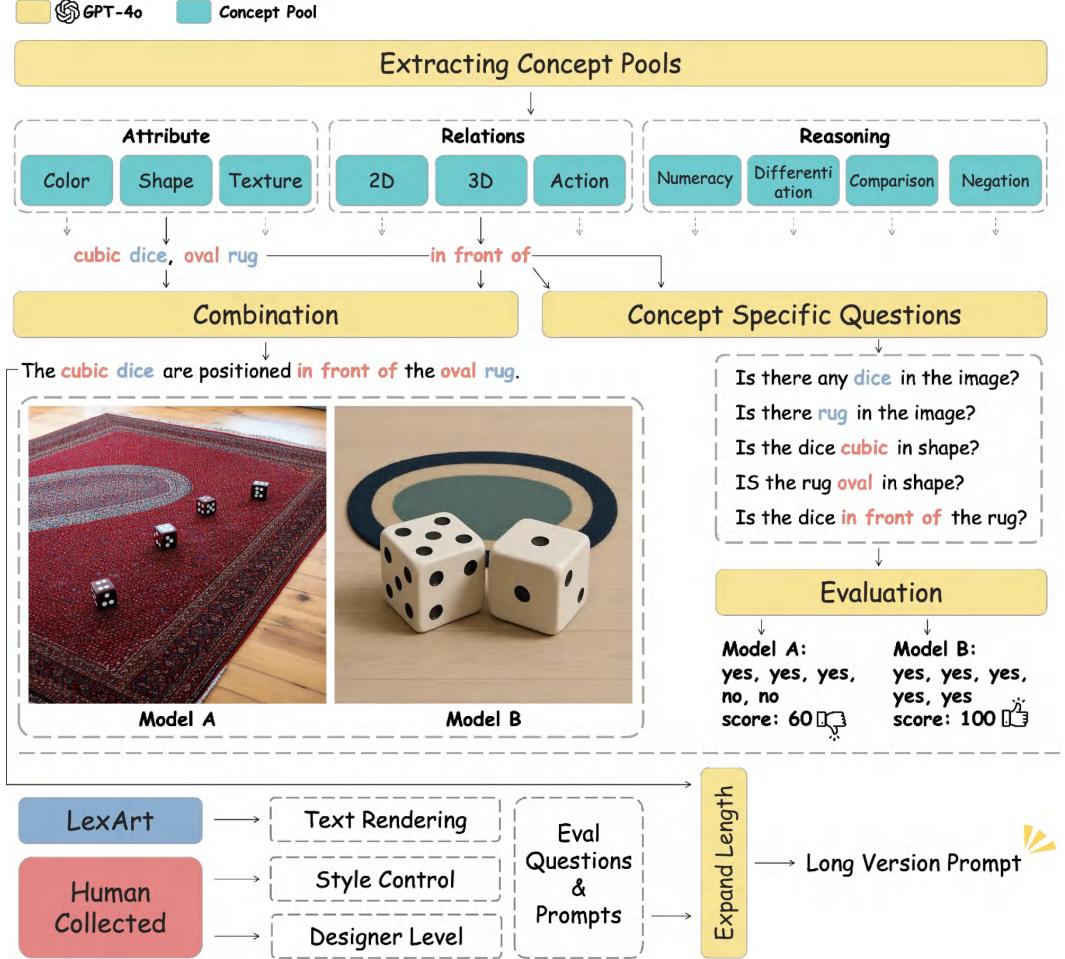


Figure 5: **(Top)**: Prompt-construction and evaluation pipeline of TIIF-Bench. The example depicts a “Color + 3D Perspective” pairing, one of the 36 distinct attribute-pool combinations defined in our framework. **(Bottom)**: In addition to systematically reusing prompts from existing benchmarks, we gather text-generation prompts from Lex-Art [40] and curate style control and real-world designer-level prompts. All prompts are further expanded by GPT-4o to produce long-form counterparts.

**Evaluation Protocol Flaws.** The evaluation-protocol flaws can be classified into two main types. First, *reliance on weak expert models*. Most benchmarks use CLIP to score image–text alignment, but CLIP is known to behave like a “bag of words” [29], conflating semantically distinct prompts like “a boy is on the bottom of a bee” and “a bee is on the bottom of a boy” (see Fig. 3). Other expert models like BLIP offer no significant improvement in fine-grained alignment, while object detectors such as UniDet often fail for complex generated images. Second, *coarse use of VLMs*. Some benchmarks query strong VLMs to assess image–text alignment. However, these evaluations are often based on a single generic prompt such as “*According to the image and your previous answer, evaluate how well the image aligns with the text prompt: {xxx}.*” Such coarse queries fail to decompose the rich, multi-attribute semantics present in modern T2I prompts. Moreover, explicitly including the full T2I prompt (*i.e.*, the image caption) in the question will induce VLM hallucinations. In such cases, the model may overfit to the linguistic input rather than grounding its judgment in the visual evidence, resulting in overly optimistic evaluations, as illustrated in Fig. 4.

### 3.2 Prompts of TIIF-BENCH

To address the limitations of existing T2I evaluation benchmarks, we build upon the hierarchical prompt taxonomies of prior work by introducing a two-stage process: **concept pool construction** followed by **attribute composition**. This process generates new prompts that target a T2I model’s

Table 1: The 10 dimensions used to build the TIIF-Bench prompt set. We first mine these dimensions from existing benchmarks and regroup them into three broad categories: attributes, relations, and reasoning. For concepts mined from COMPBENCH++, we leverage GPT-4o to separate the core **objects** from their **attributes/relations**; the two components are stored independently for later composition (*i.e.*, the **red** components and **blue** components are stored separately). In contrast, concepts extracted from GENAI BENCH are saved as whole phrases because their action links and logical relations—such as *differentiation*, *comparison*, and *negation*—cannot be cleanly factorized.

Category	Dimension	Source	Example Prompt	Extracted Attributes
Attributes	Color	Comp++	A tan dog with sky blue eyes posing for a picture with a man sitting on a chair in the background.	tan dog, sky blue eyes
	Shape	Comp++	The conical salt and pepper shakers with their cylindrical bases and spherical tops seasoned food in the square dining room.	conical salt and pepper shakers, cylindrical bases, spherical tops, square dining room
	Texture	Comp++	The wooden table is covered with a fabric tablecloth and adorned with a glass vase.	wooden table, fabric tablecloth, glass vase
Relations	2D Spatial	Comp++	A butterfly on the top of a desk.	butterfly on top of desk
	3D Perspective	Comp++	A key in front of a girl.	key in front of girl
	Action	GenAI	A dragon perched majestically on a craggy, smoke-wreathed mountain.	dragon perched on mountain
Reasoning	Numeracy	Comp++	Three bottles stood next to three printers on the shelf.	three bottles, three printers
	Differentiation	GenAI	A person in uniform pointing out landmarks to a person in a window seat wearing noise-canceling headphones.	a person in uniform, another person wearing headphones
	Comparison	GenAI	In a mysterious swamp, the flowers are taller than the trees.	flowers are taller than the trees
	Negation	GenAI	The girl with glasses is drawing, and the girl without glasses is singing.	girl without glasses is singing

core capabilities such as object-attribute binding and spatial layout. To more comprehensively assess instruction-following performance, we further introduce three **new evaluation dimensions**: text rendering, style control, and real-world designer-level prompts. Each prompt is also processed by a **length augmentation** module that generates both concise and verbose versions, enabling evaluation across varying linguistic lengths. The full data generation pipeline is illustrated in Fig. 5.

**Concept Pool Construction.** We first group the prompts in existing benchmarks based on their semantics and leverage GPT-4o to extract the underlying *object–attribute/relation pairs*, forming a set of dimension-specific concept pools. In total, we construct **10 concept pools** from existing benchmarks, categorized them into **three groups**, as summarized in Tab. 1.

**Attribute Composition.** Building upon concept pools, we generate prompts by randomly combining attributes from each pool, leveraging GPT-4o to compose them into natural instructions. As illustrated in Tab. 2, we define **36 distinct combinations**, each paired with a dedicated meta-prompt to guide GPT-4o to assemble instructions. This sampling strategy ensures that the resulting prompts are both semantically unique and compositionally diverse. Prompts that combine elements drawn from a *single* concept-pool group are classified as **Basic Following**. In contrast, **Advanced Following** prompts intertwine elements taken from *different* concept-pool groups, yielding more intricate compositions.

**New Evaluation Dimensions.** To extend evaluation beyond conventional instruction-following skills, we introduce three novel dimensions: text rendering, style control, and real-world scenario prompts. (i) **Text rendering** evaluates a model’s ability to accurately reproduce complex typographic elements, using prompts sourced from the *Lex-Art* corpus [40]. (ii) **Style control** assesses the model’s capacity to adhere to high-level artistic directives, with prompts manually curated from leading AIGC creator communities. (iii) **Designer-level** prompts involve complex instructions that incorporate practical constraints and domain-specific knowledge, also collected through manual annotation. The text rendering and style control dimensions are included in the **Advanced Following** set, while the designer-level prompts constitute the **Designer Level Following** set.

Table 2: Based on difficulty, we define **36 attribute-pool combinations**, grouped into three levels: **Basic Following**, **Advanced Following**, and **Designer Level Following**. **Basic Following** prompts are constructed by combining attributes/relations and objects from a single attribute group, while **Advanced Following** prompts involve cross-group composition, as well as two specialized dimensions—*text rendering* and *style control*—to evaluate text rendering and aesthetic coherence. Finally, 100 real-world **Designer Level Following** prompts are manually curated. The rightmost column reports the number of prompts per subclass, all of which are further expanded into long-form variants.

Level	Dimension	Combination Policy	Count
Basic Following	Attribute	Shape+Color, Color+Texture, Texture+Color, Shape+Texture	{100,50,50,100}
	Relation	2D Spatial, 3D Perspective, Action+2D, Action+3D	{75,75,75,75}
	Reasoning	Numeracy, Negation, Differentiation, Comparison	{75,75,75,75}
Advanced Following	Attribute + Relation	Action+Color, Action+Texture, Color+2D, Color+3D, Shape+2D, Shape+3D, Texture+2D, Texture+3D	{50,50,33,33,33,33,33,33}
	Attribute + Reasoning	Numeracy+{Color, Texture}, Comparision+{Color, Texture}, Differentiation+{Color, Texture}, Negation+{Color, Texture}	{40,40,40,40,40,40,40,40}
	Relation + Reasoning	Numeracy+{2D, 3D}, Comparision+{2D, 3D}, Differentiation+{2D, 3D}, Negation+{2D, 3D}	{40,40,40,40,40,40,40,40}
	Text Generation	—	{150}
	Style Control	—	{150}
Designer Level Following	Complex	—	{100}

**Length Augmentation.** Finally, for each generated prompt, we leverage GPT-4o to construct a corresponding long-form variant by expanding the content through natural language paraphrasing and stylistic elaboration, while faithfully preserving its original semantics. The meta-prompt that guides GPT-4o to expand the original instruction can be found in Sec. A1.1.

### 3.3 Evaluation Method of TIIF-Bench

To overcome the limitations of expert scorers, such as CLIP, and open-set object detectors in evaluating the generated images, we propose an attribute-specific, fine-grained evaluation protocol. Our method leverages the vast world knowledge encoded in VLMs to assess the alignment between textual instructions and generated images. As illustrated in Fig. 5, given an input prompt, we first extract its  $N$  core concepts  $C = \{c_i\}_{i=1}^N$  from the constructed concept pools. For each concept  $c_i$ , we employ an LLM (e.g., GPT-4o) to generate a corresponding yes/no question  $q_i$ , resulting in a set of evaluation questions  $Q = \{q_i\}_{i=1}^N$  and the associated ground-truth answers  $A = \{\hat{a}_i\}_{i=1}^N$ . The generated image, along with the questions  $Q$ , is then input into a VLM (e.g., GPT-4o, Qwen2.5-VL-72B), which produces  $N$  predicted answers. The final evaluation score  $s$  for the generated image is computed as:

$$s = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[a_i = \hat{a}_i],$$

where  $\mathbb{I}[\cdot]$  denotes the indicator function, and  $a_i$  represents the VLM’s answer to question  $q_i$ . By leveraging attribute-specific questions, our approach does not require the full T2I prompt during evaluation, thus mitigating hallucinations induced by language bias in previous methods.

For the newly introduced **style control** dimension, we directly utilize “Is this image in the {Style} style?” as the evaluation question. As this query is independent of specific image objects, it further mitigates the risk of hallucination. For **designer-level** prompts, which are inherently complex and encode deep human priors, we manually construct a tailored set of evaluation questions for each prompt to ensure a reliable assessment.

For **text rendering** evaluation, we propose a **Global Normalized Edit Distance (GNED)** metric. Let  $P = \{p_1, \dots, p_m\}$  be the set of text-rendering words from the prompt and  $G = \{g_1, \dots, g_n\}$  the set of OCR-extracted words from the generated image. GNED calculates the minimal character-level normalized edit distance (NED) between words in  $P$  and  $G$  via the Hungarian algorithm for optimal bipartite matching  $\mathcal{M}$ , then adds a penalty  $|m - n|$  for unmatched words (e.g., missing/hallucinated

text) and normalizes the total cost by  $\max(m, n)$ :

$$\text{GNED}(P, G) = \left( \sum_{(i,j) \in \mathcal{M}} \text{NED}(p_i, g_j) + |m - n| \right) / \max(m, n),$$

where  $\text{NED}(p_i, g_j)$  quantifies character-level discrepancies between matched word pairs. GNED is bounded within  $[0, 1]$ , where 0 indicates perfect alignment and 1 indicates maximal deviation. Unlike metrics such as OCR Recall and PNED [40], which are sensitive to text length and word count imbalance, GNED robustly penalizes both over-generation and omission, enabling consistent and comparable evaluation across samples of varying lengths. This makes GNED particularly suitable for glyph-agnostic and content-sensitive T2I model evaluation. Illustrative examples of the GEND metric are provided in Fig. 6 for better understanding.

Prompt	Janus-Pro	SD 3.5	GPT-4o
<p><i>A picture of a charming illustration showcasing a teacup and a slice of cake in warm, inviting tones, with the text on it:</i>  <i>"drink", "tea", "eat", "cake", "relax", "enjoy", "delight".</i></p>  <p>Recall: 0.00    GNED: 0.95</p>	 <p>Recall: 1.00    GNED: 0.16</p>	 <p>Recall: 0.86    GNED: 0.14</p>	
<p><i>prompt: A picture of a woman in a green dress standing near a glass door, with the text on it:</i>  <i>"Corporate", "DISC", "Workshop", "THE", "BIZ", "STUDIO".</i></p>  <p>Recall: 0.00    GNED: 0.81</p>  <p>Recall: 0.83    GNED: 0.14</p>  <p>Recall: 1.00    GNED: 0.00</p>			

Figure 6: Visualization examples for Recall and GNED. **Top row:** Janus-Pro fails to generate any of the required words completely, resulting in a Recall of 0. Additionally, the overall quality of the generated text is poor, leading to a high GNED score. SD 3.5 successfully generates all target words from the prompt, yielding a Recall of 1.00; however, the word “enjoy” contains minor distortions, resulting in a GNED of 0.16. GPT-4o omits the word “cake,” achieving a Recall of 6/7. Despite the omission, the visual quality of the generated words is higher than that of SD 3.5, thus resulting in a lower GNED score. **Bottom row:** GPT-4o generates all required words with excellent typographic fidelity, achieving both the highest Recall score of 1.00 and the best GNED score of 0.00.

## 4 Experiments

### 4.1 Performance of T2I Models on TIIF-Bench

We selected 10% of prompts from each dimension (and carefully chose 25 challenging prompts from the designer dimension) to create a **testmini** subset<sup>1</sup>, on which we evaluate a suite of widely-used *closed-source* and leading *open-source* models, including GPT-4o [26], FLUX.1-Pro [8], DALLE-3 [35], MidJourney V6 and V7 [28], Stable Diffusion series [1, 2], PixArt series [5, 4, 3], SANA series [10, 9], Playground series [6, 7] and AR pipelines [16, 19, 20, 36, 37]. The results on testmini subset are presented in Tab. 3.

<sup>1</sup>Closed-source T2I models incur significant computational cost and time for image generation. Following previous benchmarks such as CompBench++ [32], we carefully curated a testmini subset to facilitate efficient evaluation and validation of existing and coming T2I models on TIIF-Bench. The performance of open-source models on the ENTIRE TIIF-Bench is detailed in Appendix A3.2.

Table 3: Performance of closed-source models and state-of-the-art open-source models on TIIF-Bench **testmini** subset. **GPT-4o is used as the VLM for evaluation.** Evaluated systems are grouped into (i) **diffusion-based** open-source models, (ii) **autoregressive** open-source models, and (iii) **closed-source** models. Within each group, the highest score is shown in bold and color-coded for that group.

Model	Overall		Basic Following						Advanced Following						Designer											
			Avg			Attribute		Relation		Reasoning			Avg		Attribute +Relation		Attribute +Reasoning		Relation +Reasoning		Style		Text		Real World	
	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long		
<b>Diffusion based Open-Source Models</b>																										
FLUX.1 dev	71.09	<b>71.78</b>	<b>83.12</b>	78.65	87.05	83.17	<b>87.25</b>	80.39	<b>75.01</b>	72.39	65.79	<b>68.54</b>	67.07	<b>73.69</b>	<b>73.84</b>	73.34	<b>69.09</b>	<b>71.59</b>	66.67	66.67	43.83	<b>52.83</b>	70.72	<b>71.47</b>		
SD XL	54.96	42.13	65.72	53.28	59.33	50.83	77.57	62.57	60.32	46.57	49.73	36.22	47.82	35.57	56.22	45.34	52.59	36.09	73.33	60.00	16.83	0.83	50.92	41.59		
SD 3	67.46	66.09	78.32	77.75	83.33	79.83	82.07	78.82	71.07	74.07	61.46	59.56	61.07	64.07	68.84	70.34	50.96	57.84	66.67	76.67	59.83	20.83	63.23	67.34		
SD 3.5 L	<b>71.15</b>	66.96	78.34	79.56	79.50	76.50	80.96	83.21	72.46	<b>78.71</b>	<b>67.67</b>	61.18	66.46	61.89	73.53	<b>74.15</b>	60.03	61.53	73.33	63.33	<b>70.52</b>	42.52	64.43	66.39		
SD3.5 M	70.17	66.19	80.20	75.20	84.50	77.00	78.90	77.51	77.21	71.08	62.33	66.49	66.54	77.51	57.92	62.32	61.46	63.53	80.00	73.33	53.39	28.51	<b>71.64</b>	64.93		
FlowGRPO	69.90	69.67	80.15	<b>79.90</b>	<b>89.00</b>	<b>85.00</b>	78.80	78.50	72.64	76.21	63.28	67.38	<b>70.83</b>	72.37	58.37	65.66	59.50	65.85	73.33	66.67	62.44	52.94	64.18	63.81		
SANA Sprint	63.68	58.50	76.58	71.00	75.33	71.33	81.82	72.07	72.57	69.57	57.67	51.80	55.32	54.94	68.46	66.72	62.59	63.46	80.00	60.00	8.82	5.83	66.96	58.01		
SANA 1.5	67.15	65.73	79.66	77.08	79.83	77.83	85.57	<b>83.57</b>	73.57	69.82	61.50	60.67	65.32	56.57	69.96	73.09	62.96	65.84	80.00	<b>80.00</b>	17.83	15.83	71.07	68.83		
Playground v2	45.64	52.78	59.83	69.58	51.33	66.33	70.57	76.07	57.57	66.32	38.43	44.75	41.57	45.57	48.96	59.97	41.72	54.84	53.33	60.00	0.00	0.83	45.32	46.44		
Playground v2.5	47.73	54.82	63.08	68.08	57.83	73.83	71.82	77.32	59.57	53.07	40.73	48.17	39.70	45.82	49.59	64.22	44.22	46.72	60.00	80.00	0.00	4.83	47.19	47.56		
PixArt-delta	41.01	48.24	53.83	59.25	46.33	52.83	62.07	71.32	53.07	53.57	34.60	42.77	32.44	37.44	53.59	56.59	36.96	49.46	46.67	73.33	0.00	0.00	38.23	40.10		
PixArt-alpha	44.37	50.50	55.50	61.00	52.33	56.33	63.82	74.07	50.32	52.57	38.71	44.90	37.82	41.32	58.84	52.46	42.62	47.09	50.00	76.67	0.00	0.83	45.70	53.16		
PixArt-sigma	62.00	58.12	70.66	75.25	69.33	78.83	75.07	77.32	67.57	69.57	57.65	49.50	65.20	56.57	66.96	61.72	66.59	54.59	<b>83.33</b>	70.00	1.83	1.83	62.11	52.41		
LUMINA-Next	50.93	52.46	64.58	66.08	56.83	59.33	67.57	71.82	69.32	67.07	44.75	45.63	51.44	43.20	51.09	59.72	44.72	54.46	70.00	66.67	0.00	0.83	47.56	49.05		
Hunyuan-DiT	51.38	53.28	69.33	69.00	65.83	69.83	78.07	73.82	64.07	63.32	42.62	45.45	50.20	41.57	59.22	61.84	47.84	51.09	56.67	73.33	0.00	0.83	40.10	44.20		
<b>AR based Open-Source Models</b>																										
Llamagen	41.67	38.22	53.00	50.00	48.33	42.33	59.57	60.32	51.07	47.32	35.89	32.61	38.82	31.57	40.84	47.22	49.59	46.22	46.67	33.33	0.00	0.00	39.73	35.62		
LightGen	53.22	43.41	66.58	47.91	55.83	47.33	74.82	45.82	69.07	50.57	46.74	41.53	62.44	40.82	61.71	50.47	50.34	45.34	53.33	53.33	0.00	6.83	50.92	50.55		
Show-o	59.72	58.86	73.08	75.83	74.83	79.83	78.82	78.32	65.57	69.32	53.67	50.38	60.95	56.82	68.59	68.96	66.46	56.22	63.33	66.67	3.83	2.83	55.02	50.92		
Infinity	62.07	62.32	73.08	75.41	74.33	76.83	72.82	77.57	72.07	71.82	56.64	54.98	60.44	55.57	<b>74.22</b>	64.71	60.22	59.71	<b>80.00</b>	<b>73.33</b>	10.83	23.83	54.28	56.89		
JanusPro	66.50	65.02	79.33	78.25	79.33	82.33	78.32	73.32	<b>80.32</b>	79.07	59.71	58.82	66.07	56.20	70.46	<b>70.84</b>	67.22	59.97	60.00	70.00	<b>28.83</b>	<b>33.83</b>	65.84	60.25		
T2I-R1	<b>68.59</b>	<b>67.19</b>	<b>82.90</b>	<b>81.63</b>	<b>86.50</b>	<b>83.00</b>	<b>83.47</b>	<b>79.43</b>	78.73	<b>82.46</b>	<b>69.05</b>	<b>68.00</b>	<b>71.64</b>	<b>69.47</b>	72.43	69.95	<b>69.40</b>	<b>70.40</b>	60.00	63.33	27.60	26.24	<b>67.54</b>	<b>60.45</b>		
<b>Closed-Source Models</b>																										
DALL-E 3	74.96	70.81	78.72	78.50	79.50	79.83	80.82	78.82	75.82	76.82	73.39	67.27	73.45	67.20	72.01	71.34	63.59	60.72	89.66	86.67	66.83	54.83	72.93	60.99		
MidJourney v6	70.78	67.70	76.00	69.08	77.83	69.33	81.32	73.07	68.82	64.82	68.54	67.62	57.82	61.95	69.84	63.96	57.46	60.34	83.33	73.33	75.83	73.83	65.10	68.46		
MidJourney v7	68.74	65.69	77.41	76.00	77.58	81.83	82.07	76.82	72.57	69.32	64.66	60.53	62.70	62.70	81.22	71.59	60.72	64.59	83.33	80.00	24.83	20.83	68.83	63.61		
FLUX.1 Pro	67.32	69.89	79.08	78.91	78.83	81.33	82.82	83.82	75.57	71.57	61.10	65.37	62.32	65.57	69.84	71.47	65.96	67.72	63.00	63.00	35.83	55.83	71.80	68.80		
GPT-4o	<b>89.15</b>	<b>88.29</b>	<b>90.75</b>	<b>89.66</b>	<b>91.33</b>	<b>87.08</b>	<b>84.57</b>	<b>84.57</b>	<b>96.32</b>	<b>97.32</b>	<b>88.55</b>	<b>88.35</b>	<b>87.07</b>	<b>89.44</b>	<b>87.22</b>	<b>83.96</b>	<b>85.59</b>	<b>83.21</b>	<b>90.00</b>	<b>93.33</b>	<b>89.83</b>	<b>86.83</b>	<b>89.73</b>	<b>93.46</b>		

For each of the three difficulty levels (basic following, advanced following, designer-level), we calculate the average score of its associated dimensions for both short and long versions of the prompts. We also calculate the overall average scores of all dimensions across all the three difficulty levels. From Tab. 3, we can have a set of key observations, which are detailed below.

#### 4.1.1 Diffusion-based Open-Source Models

**(i) Overall performance.** From the upper panel of Tab. 3, we see that SD-3.5 attains the highest scores on short prompts, whereas FLUX-1 Dev delivers the strongest instruction-following performance on long prompts among diffusion-based open-source models—likely due to its MMDiT architecture and large parameter count. Trailing closely are SD-3.5, SD-3, SANA-1.5, and PixArt-Sigma. A notable outlier is the Playground series. Despite its relatively modest average image quality, it shows consistently improved performance under long-prompt conditions, which can be attributed to its distinctive mechanism for deeply embedding textual semantics within the generation pipeline.

**(ii) Text Rendering.** Among diffusion-based open-source models, only FLUX-1 Dev, SANA 1.5 (and SANA Sprint), and the Stable Diffusion family (SD-XL, SD-3, SD-3.5) support in-image text rendering. While SD-3 and SD-3.5 perform well with short prompts, their effectiveness degrades with longer prompts, probably due to difficulties in isolating textual elements from the broader descriptive context. In contrast, FLUX-1 Dev maintains consistent performance across both short and extended prompts, making it a more reliable choice for text rendering tasks.

**(iii) Style Control.** Prompt length exerts opposite effects on different models. In systems such as the Playground series and PixArt-Alpha, the internal "style" vocabulary is limited, *e.g.*, terms like *Ghibli*

or *Cyberpunk* are poorly grounded. Consequently, short style-control prompts that rely solely on such keywords often fail to yield coherent visual results. Longer prompts that provide additional visual context can compensate for this limitation, substantially improving the output quality. In contrast, models such as SD 3.5 and PixArt-Sigma are trained on datasets where stylistic instructions typically appear as brief, label-like tokens. When these cues are embedded within longer descriptive prompts, their salience is diminished, leading to reduced stylistic fidelity and degraded generation quality.

**(iv) Designer-Level Prompts.** The designer-level prompts encompass the densest and most diverse set of requirements, offering the most comprehensive test of a model’s instruction-following capability. As a result, the ranking of models on this dimension closely aligns with that of the overall performance.

**(v) Robustness to Prompt Length.** Top-performing models such as FLUX.1 Dev, SD 3.5, PixArt-Sigma, and SANA 1.5 demonstrate strong robustness to variations in prompt length, producing consistent results across short and long versions of semantically equivalent prompts. In contrast, weaker models such as SDXL, SD 3, PixArt-Alpha, and Playground series exhibit large performance discrepancies. This provides preliminary evidence that the instruction comprehension capability of a T2I model is positively correlated with its image generation quality.

#### 4.1.2 AR-based Open-Source Models

**(i) Overall Performance.** From the middle panel of Tab. 3, we see that Janus-Pro [17] achieves the best overall instruction-following performance, which can be attributed to its large-scale pretraining and unified training across generation and understanding.

**(ii) Text Rendering.** Autoregressive architectures are inherently limited in their ability to render text within images. Among them, only Show-o [41] and Janus-Pro support basic text generation, but both of them perform substantially worse than state-of-the-art diffusion-based models.

**(iii) Style Control.** Even compared to the most advanced open-source diffusion-based models, Janus-Pro and Infinity exhibit strong performance in style control. This is because autoregressive architecture enables more accurate interpretation of descriptive semantics, leading to improved image fidelity under complex stylistic instructions.

**(iv) Designer-Level Prompts.** Due to the high visual complexity embedded in designer-level prompts, current AR-based models are constrained by their image quality (referring to visual fidelity rather than instruction adherence), resulting in only moderate overall scores. Nevertheless, rankings along this dimension remain consistent with the models’ overall performance.

**(v) Robustness to Prompt Length.** As diffusion models, those AR models with stronger overall performance (*e.g.*, Janus-Pro, Infinity) remain stable across prompt lengths, while weaker models show large gaps. This further supports the link between instruction understanding and generation.

#### 4.1.3 Diffusion-based vs. AR-based Open-Source Models

Although AR-based models tend to produce images with lower visual fidelity, their autoregressive architecture, jointly trained on generation and understanding tasks, grants them strong instruction-following capabilities. For instance, Janus-Pro outperforms diffusion-based model PixArt-Sigma on TIIF Bench. We present qualitative examples in Fig. 7 to illustrate this finding.

#### 4.1.4 Closed-Source Models

**(i) Overall Performance.** From the bottom panel of Tab. 3, we see that GPT-4o demonstrates the strongest instruction-following ability among all models, largely due to its autoregressive architecture and superior language understanding. It excels not only in capturing object attributes and spatial relations, but also in handling complex logical reasoning. Its success rate of text rendering is exceptionally high, with no competitor close to its level.

**(ii) Text Rendering.** Thanks to its robust comprehension capabilities, GPT-4o leads by a wide margin in text rendering. Notably, MidJourney V7 shows a significant degradation compared to V6.

**(iii) Style Control.** With a vast knowledge base and strong semantic understanding, GPT-4o dominates the style control dimension, consistently producing stylistically aligned outputs.

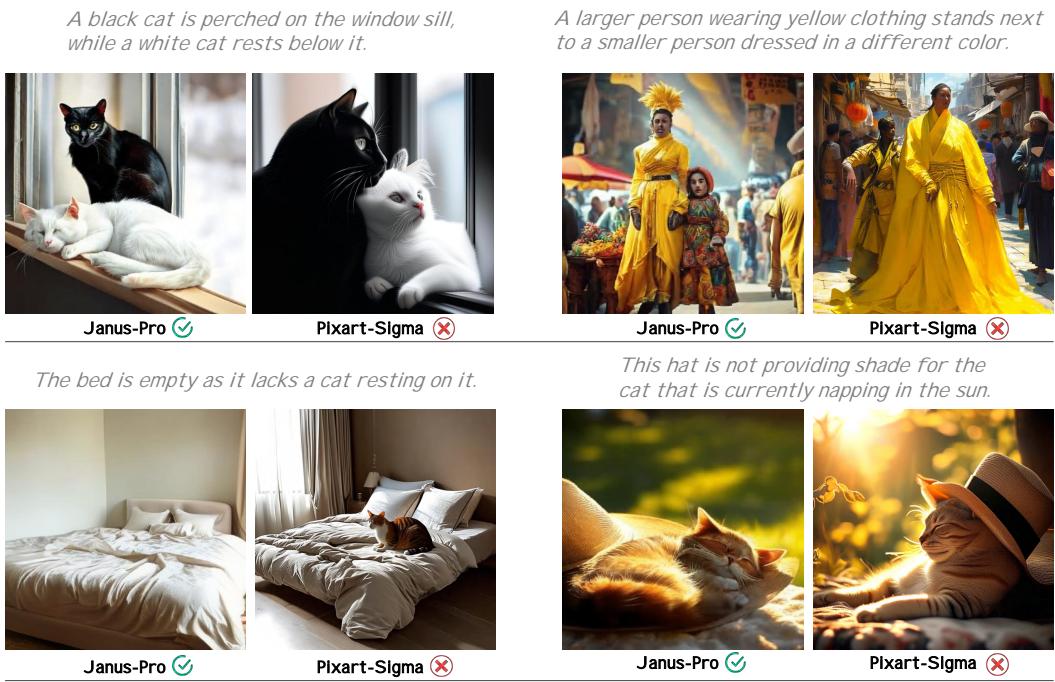


Figure 7: Although AR-based models typically produce images with lower visual fidelity, their autoregressive architecture—jointly trained for both generation and understanding tasks—endows them with strong instruction-following capabilities. We observe that Janus-Pro outperforms the diffusion-based model PixArt-Sigma on TIIF BENCH, on prompts requiring reasoning skills such as **differentiation**, **comparison**, and **negation**.

(iv) **Designer-Level Prompts.** Closed-source models generally perform well on designer-level prompts, likely due to high-quality data exposure during training. GPT-4o, with its advanced understanding of rich, detailed instructions, outperforms all others in this dimension.

(v) **Robustness to Prompt Length.** Closed-source models, trained on large-scale high-quality datasets with well-engineered architectures, exhibit strong robustness to prompt length variations. This further supports the correlation between instruction comprehension and generation quality.

#### 4.1.5 Closed- vs. Open-Source Models

While closed-source models generally outperform open-source models in generated image quality, their instruction-following capabilities are not always superior. For instance, Flux.1 Pro actually achieves overall lower scores than Flux.1 Dev, particularly under long-prompt scenarios. GPT-4o, benefiting from its autoregressive architecture and strong comprehension abilities, currently maintains a significant lead over all other models, no matter open-source and closed-source.

We observe that most models exhibit strong instruction-following capabilities in object attribute dimensions such as color and material. However, their performance drops significantly when dealing with spatial relations or logical reasoning. It is worth mentioning that GPT-4o is the only model that maintains relatively stable performance across these more complex dimensions.

## 4.2 TIIF-Bench vs. Existing Benchmarks

To further assess the effectiveness of our TIIF-Bench evaluation methodology, we select four strong open-source models (SD 3.5, SANA 1.5, PixArt-Sigma, and Janus-Pro) and four representative closed-source models (GPT-4o, DALL·E 3, Flux.1 Pro, and MidJourney v6), and evaluate them on three benchmarks: COMPBENCH++, GENAI BENCH, and our proposed TIIF-BENCH. For each model, we generate images for all prompts. The resulting images are then scored using the evaluation method of each benchmark. Full scoring results are provided in Tab. 4 and Tab. 5.

Observing model performance on GENAI BENCH (Tab. 4), we find that the CLIP-based VQAScore fails to capture fine-grained text-image alignment. As a result, multiple models, including SD 3.5, SANA 1.5, PixArt-Sigma, Flux.1 Pro, MidJourney v6, receive identical scores, hindering precise evaluation. Similarly, examining COMPBENCH++ results (Tab. 5), we notice that several models achieve identical scores and there exist notable discrepancies between expert model-based and GPT-based scores. For example, FLUX.1 Pro is rated high by GPT-4o but is rated significantly lower by expert models. In contrast, TIIF-Bench yields a clear and consistent ordering of the evaluated models.

To further assess how well each benchmark aligns with human preferences, we conduct a user study as described in Sec. 4.3. Specifically, ten volunteers are invited and they are asked to rank the outputs of the eight models for each prompt, without knowing the model identities. We then compute the *Spearman ρ* correlation between the benchmark rankings and human-annotated rankings. As shown in Tab. 6, COMPBENCH++ shows only weak alignment with human judgment. In contrast, TIIF-BENCH exhibits near-perfect correlation with human preferences (see Tab. 7), highlighting the effectiveness of our fine-grained evaluation framework.

Table 4: Performance of eight T2I models on GENAI BENCH, evaluated using VQAScore. Results are reported for both basic and advanced prompt categories, as well as the overall average.

Model	Basic	Advanced	Overall
SD 3.5	0.88	0.65	0.75
SANA 1.5	0.86	0.66	0.75
PixArt-Sigma	0.86	0.65	0.75
Flux.1 Pro	0.81	0.68	0.75
Janus-Pro	0.80	0.64	0.71
DALLE-3	0.85	0.76	0.81
MidJourney v6	0.85	0.68	0.75
GPT-4o	0.86	0.83	0.85

Table 5: Evaluation results on COMPBENCH++. COMPBENCH++ provides both a GPT-based evaluation method and an expert model-based evaluation method. We report the results from both.

Model	AVG		Color		Shape		Texture		Numeracy		2D Spatial		3D Spatial		Non-Spatial		Complex	
	Expert	GPT	Expert	GPT	Expert	GPT	Expert	GPT	Expert	GPT	Expert	GPT	Expert	GPT	Expert	GPT	Expert	GPT
SD 3.5	0.507	68.89	0.767	83.40	0.596	85.15	0.706	87.44	0.621	45.03	0.277	44.76	0.403	45.06	0.315	85.03	0.376	75.26
SANA 1.5	0.507	69.72	0.755	88.83	0.536	79.59	0.686	78.92	0.611	53.10	0.362	51.03	0.412	44.26	0.312	84.46	0.382	77.56
PixArt-Sigma	0.436	56.92	0.587	82.60	0.476	62.45	0.569	71.08	0.549	30.10	0.247	46.83	0.366	27.00	0.308	67.40	0.383	67.90
FLUX.1 Pro	0.459	70.48	0.704	85.06	0.529	70.08	0.596	75.83	0.616	72.28	0.276	59.50	0.344	50.73	0.290	77.29	0.315	73.10
Janus-Pro	0.285	47.34	0.393	57.45	0.264	42.34	0.349	48.49	0.335	36.90	0.080	17.86	0.209	29.83	0.299	71.36	0.355	74.56
DALLE-3	0.511	76.94	0.803	84.03	0.637	85.15	0.774	86.88	0.617	62.90	0.253	63.20	0.353	68.06	0.297	85.73	0.353	79.58
MidJourney v6.1	0.507	70.93	0.750	84.33	0.525	72.68	0.787	85.22	0.670	55.22	0.278	52.27	0.384	51.06	0.310	89.80	0.349	76.88
ChatGPT 4o	0.572	86.59	0.795	95.34	0.593	84.30	0.831	91.04	0.800	79.86	0.450	85.83	0.406	70.46	0.311	95.50	0.389	90.40

Table 6: Alignment of Expert- and GPT-based evaluations with human preferences on CompBench++.

Comp++	Color	Shape	Texture	Numeracy	2D Spatial	3D Spatial	Non-Spatial	Complex
<i>Spearman ρ</i>								
Experts	0.58	0.78	0.61	0.24	0.00	0.14	0.07	0.36
GPT	0.36	0.49	0.43	0.60	0.60	0.79	0.60	0.52

Table 7: Alignment of TIIF-Bench results with human preferences. GNED is used as the metric for the text rendering dimension, while all other dimensions are evaluated using VLM-based scoring.

Model	Basic Following						Advanced Following						Real-World					
	Attribute			Relation		Reasoning	Attribute + Relation			Attribute + Reasoning		Relation + Reasoning	Style	Text	Complex			
	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short			
<i>Spearman ρ</i>																		
VLM Eval/PNED	0.81	0.88	0.88	0.91	0.93	1.00	0.93	0.95	0.93	0.98	0.95	1.00	0.81	0.81	0.98	1.00	0.85	0.81

### 4.3 User Study

To further assess how well the TIIF Bench’s evaluation scores align with human preferences, we conduct a user study involving ten independent, uninformed participants. These participants are asked to rank the quality of the images generated by the eight T2I models. For each dimension in the benchmark, we sample 3 evaluation sets. Each set contains one prompt and 8 corresponding images, one from each model. The participants are asked to rank the images based on their visual quality.

During ranking, participants are required to consider two criteria: (1) how well the image follows the prompt; and (2) the overall visual quality, including clarity and aesthetic appeal. For each set, we compute a ranking based on the aggregated user input. These rankings are then averaged across the 3 sets along each dimension to produce a final ranking of the 8 models for that dimension. Fig. 8 illustrates the full user study process for one such set.

To quantify the alignment between benchmark evaluation scores and human preferences, we compute the Spearman rank correlation coefficient  $\rho$  between the two corresponding ranking sequences. As shown in Tab. 7, the evaluation results of TIIF Bench exhibit a high degree of alignment with human preferences across all dimensions.

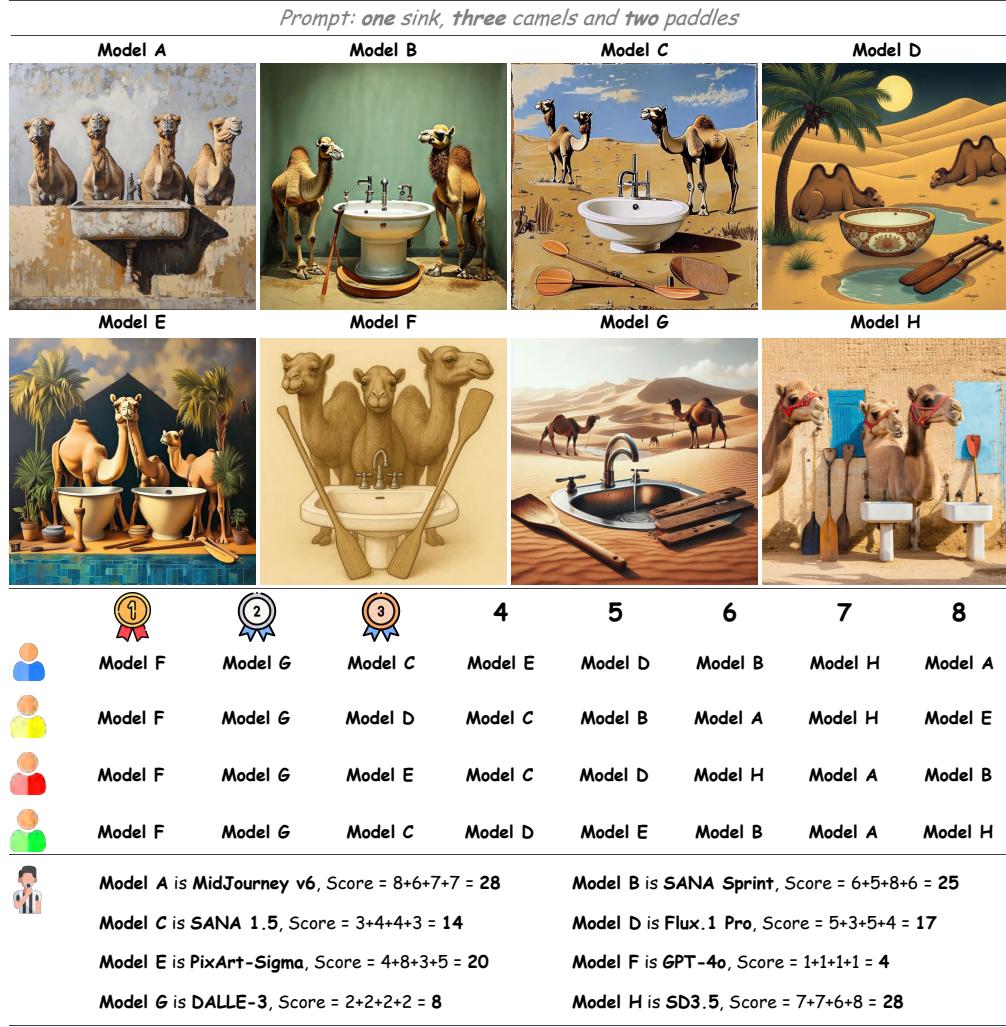


Figure 8: Illustration of the user study procedure. The figure shows one evaluation set from a specific dimension in COMPBENCH++. In practice, each dimension involves three such sets, and users perform a separate ranking for each.

## 5 Conclusion

We introduced TIIF-BENCH, a comprehensive and hierarchical benchmark designed to assess the instruction-following capabilities of T2I models. TIIF-Bench encompassed diverse combinations of concepts, as well as comprehensive evaluation dimensions, especially along text rendering, style control, and designer-level prompts. Additionally, we propose a fine-grained evaluation methodology leveraging the world knowledge embedded in VLMs, along with GNED, a new metric specifically developed to assess text rendering quality. Extensive experiments were conducted to analyze instruction-following behaviors of current T2I models during image generation. Our findings provide valuable insights for future development of T2I systems.

**Limitations.** Despite TIIF-Bench’s advantages over existing benchmarks, there remain some limitations. First, the prompts contain mainly common objects. In future work, we will include obscure vocabulary or rare objects to further challenge the instruction following capability of T2I models. Second, currently, all prompts are in English. Incorporating linguistic diversity is an important direction for future work. Last but not least, how could stylistic variations (*e.g.*, formal vs. conversational) impact model performance is not considered yet. An investigation on this problem may offer valuable insight in practical applications.

## References

- [1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
- [2] 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, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *Forty-First International Conference on Machine Learning*, June 2024.
- [3] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt- $\Sigma$ : Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-Image Generation, March 2024.
- [4] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. PIXART- $\delta$ : Fast and Controllable Image Generation with Latent Consistency Models, January 2024.
- [5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt- $\$ \alpha \$$ : Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis, December 2023.
- [6] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three Insights towards Enhancing Aesthetic Quality in Text-to-Image Generation, February 2024.
- [7] Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shirrao, Chase Lambert, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving Text-to-Image Alignment with Deep-Fusion Large Language Models, October 2024.
- [8] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- [9] Junsong Chen, Shuchen Xue, Yuyang Zhao, Jincheng Yu, Sayak Paul, Junyu Chen, Han Cai, Enze Xie, and Song Han. SANA-Sprint: One-Step Diffusion with Continuous-Time Consistency Distillation, March 2025.
- [10] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, and Song Han. SANA 1.5: Efficient Scaling of Training-Time and Inference-Time Compute in Linear Diffusion Transformer, March 2025.
- [11] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient High-Resolution Image Synthesis with Linear Diffusion Transformers, October 2024.
- [12] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving Image Generation with Better Captions.
- [13] Imagen-Team-Google, Jason Baldridge, Jakob Bauer, Mukul Bhutani, and Others. Imagen 3, December 2024.

- [14] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, Xu Luo, Zehan Wang, Kaipeng Zhang, Xiangyang Zhu, Si Liu, Xiangyu Yue, Dingning Liu, Wanli Ouyang, Ziwei Liu, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-Next: Making Lumina-T2X Stronger and Faster with Next-DiT, June 2024.
- [15] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyang Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-DiT: A Powerful Multi-Resolution Diffusion Transformer with Fine-Grained Chinese Understanding, May 2024.
- [16] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation, June 2024.
- [17] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling.
- [18] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation, October 2024.
- [19] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis, 2024.
- [20] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction, 2024.
- [21] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yi Xin, Xinyue Li, Qi Qin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining, 2025.
- [22] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025.
- [23] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025.
- [24] Chengzhuo Tong, Ziyu Guo, Renrui Zhang, Wenyu Shan, Xinyu Wei, Zhenghao Xing, Hongsheng Li, and Pheng-Ann Heng. Delving into rl for image generation with cot: A study on dpo vs. grpo, 2025.
- [25] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aoju Zhou, Bin Wei, Shanghang Zhang, Peng Gao, Chunyuan Li, and Hongsheng Li. Mavis: Mathematical visual instruction tuning with an automatic data engine, 2024.
- [26] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [27] Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluis Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024.
- [28] Midjourney Team. Midjourney. <https://www.midjourney.com/>, 2025.
- [29] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation, 2024.
- [30] 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, 2023.
- [31] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Ming Ding, Xiaohan Zhang, Xiaotao Gu, Shiyu Huang, Minlie Huang, Jie Tang, and Yuxiao Dong. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation, 2025.
- [32] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation, 2025.
- [33] Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment, 2023.
- [34] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation, 2024.

- [35] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023.
- [36] Xianfeng Wu, Yajing Bai, Haoze Zheng, Harold Haodong Chen, Yexin Liu, Zihao Wang, Xuran Ma, Wen-Jie Shu, Xianzu Wu, Harry Yang, and Ser-Nam Lim. LightGen: Efficient Image Generation through Knowledge Distillation and Direct Preference Optimization, March 2025.
- [37] Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling Autoregressive Text-to-image Generative Models with Continuous Tokens, October 2024.
- [38] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning, March 2022.
- [39] Jiayi Lei, Renrui Zhang, Xiangfei Hu, Weifeng Lin, Zhen Li, Wenjian Sun, Ruoyi Du, Le Zhuo, Zhongyu Li, Xinyue Li, Shitian Zhao, Ziyu Guo, Yiting Lu, Peng Gao, and Hongsheng Li. Imagine-e: Image generation intelligence evaluation of state-of-the-art text-to-image models, 2025.
- [40] Shitian Zhao, Qilong Wu, Xinyue Li, Bo Zhang, Ming Li, Qi Qin, Dongyang Liu, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Peng Gao, Bin Fu, and Zhen Li. Lex-art: Rethinking text generation via scalable high-quality data synthesis, 2025.
- [41] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One Single Transformer to Unify Multimodal Understanding and Generation, October 2024.

## Appendix

### A1 Meta Prompts

#### A1.1 Meta Prompt for Length Augmentation

For each generated prompt, we use GPT-4o to create a longer version of it by paraphrasing and elaborating in natural language, while preserving the original meaning. The prompt is as follows:

```
###RAW CAP###
You are a professional writer. Observe the sentence above, which describes a visual scene.
Please expand the sentence by increasing its linguistic richness. You may elaborate with more complex structure, rhetorical flourishes, or stylistic details-however, **DO NOT** introduce any new objects, entities, or events. The visual content must remain unchanged.
Return only the expanded sentence, without any extra explanation or commentary.
```

#### A1.2 Meta Prompt for Evaluation

For each prompt and its corresponding generated image, we insert the associated list of yes/no questions into a predefined meta prompt. The resulting prompt, along with the image, is then passed to VLMs (GPT-4o or Qwen-VL2.5-72B) for evaluation. The meta prompt is as follows:

```
You are tasked with carefully examining the provided image and answering the following yes or no questions:
Questions:
##YNQuestions##

Instructions:
1. Answer each question on a separate line, starting with "yes" or "no",
followed by a brief reason.
2. Maintain the exact order of the questions in your answers.
3. Provide only one answer per question.
4. Return only the answers-no additional commentary.
5. Each answer must be on its own line.
6. Ensure the number of answers matches the number of questions.
```

## A2 Additional Visualizations

### A2.1 Failure Cases of Strong Closed-source Models in the Relation Dimension

Through extensive experiments on TIIF BENCH, we observe that most models exhibit strong instruction-following capabilities in object attribute dimensions such as color and material. However, their performance degrades when handling spatial relations. Fig. 9 presents several failure cases from strong closed-source models when generating images for prompts involving spatial relations.

### A2.2 Additional Examples of Style Control

Style control, as a key dimension for evaluating T2I models' ability to manage global image quality, content understanding, and aesthetic coherence, has been largely overlooked in previous T2I benchmarks. TIIF BENCH introduces this dimension explicitly. Fig. 10 shows some illustrative examples.

### A2.3 Additional Examples of text rendering

TIIF BENCH introduces text rendering as a novel evaluation dimension to assess a model's ability to generate complex, non-natural textures such as embedded text. We adopt two metrics for this task:



**Model:** DALLE 3

**Prompt:** In the scene before us, a radiant, fluffy **white cloud**, its billowy form resembling soft tufts of cotton, is gracefully positioned to the right of a serene blue cup, which stands with a quiet dignity, its smooth cerulean surface reflecting a tranquil hue, serving as a vivid contrast to the ethereal cloud beside it.



**Model:** MidJourney v6

**Prompt:** In a serene and tranquil setting, the woman, with an almost contemplative air about her, is distinctly not perusing the enigmatic and rustic charm of **wooden books** that lie upon a similarly crafted **wooden surface**, whose rich textures intertwine seamlessly, inviting the observer to ponder the timeless connection between nature and artifice.



**Model:** MidJourney v7

**Prompt:** Upon observing the setting before me, I noticed that the metallic fork, with its polished, gleaming tines catching the light ever so slightly, was resting quietly **on the bottom of the fabric shirt**, which lay crumpled and soft, its material seeming to embrace the cool, hard presence of the utensil with a kind of resigned familiarity



**Model:** GPT4o

**Prompt:** In the scene unfolding before the viewer's eyes, the tall, **elegantly conical chef hat**, with its pristine white fabric crisp and unmarred, stands hidden just behind the large, perfectly round, and glistening spherical snowball, rendering it invisible from this particular perspective.

Figure 9: Failure cases of closed-source models on prompts involving spatial relations.

OCR Recall and the proposed GNED. Visualizations of both metrics are shown in Fig.6. Additionally, Fig.11 presents qualitative examples of how different T2I models perform on this dimension.

#### A2.4 Additional Examples of Real-world Prompts

The designer-level prompts feature the most dense and diverse set of requirements, providing the most comprehensive evaluation of a model’s instruction-following capabilities. Fig. 12 showcases examples of how current T2I models perform on this dimension.

### A3 More Experiments

#### A3.1 QwenVL2.5-72B as Evaluation Model on TIIF-Bench Testmini Subset

In addition to GPT-4o, we also employed QwenVL2.5-72B as an evaluation model to provide fine-grained scoring of the generated outputs from common T2I models. The results are summarized in Tab. 8. From this table, we observe that although the specific scores vary with those evaluated by GPT-4o due to the different world knowledge embedded in the two VLMs, the ranking order of the evaluated models remains largely consistent. We calculated the Spearman rank correlation coefficient  $\rho$  between these two sets of scores, and the results are presented in Tab. 9. In summary, the conclusions derived from Tab. 3 of the main paper remain unchanged by using QwenVL2.5-72B as the evaluation model.

#### A3.2 Results on the Entire TIIF-Bench

We evaluated all open-source models on the complete set of prompts in TIIF-Bench, scoring the generated images using GPT-4o and QwenVL2.5-72B. The results are presented in Tab. 10 and Tab. 11. Note that we did not evaluate the closed-source models on the full set of prompts due to their slower generation speed and higher cost. From Tab. 10 and Tab. 11, we see that although absolute

scores differ between the **ENTIRE TIIF-Bench** and the **testmini subset**, the ranking of models remains largely unchanged and the conclusions are consistent. Among diffusion-based models, SD-3.5 and FLUX-1 Dev lead on short and long prompts, respectively, while Janus-Pro tops the AR-based category. High-performing models exhibit strong robustness to prompt length, whereas weaker models are highly sensitive to it, confirming that a T2I model’s instruction-comprehension ability is positively correlated with its image-generation quality.

Table 8: Performance of closed-source models and state-of-the-art open-source models on TIIF-Bench **testmini** subset evaluated by QwenVL2.5-72B. Evaluated systems are grouped into (i) **diffusion-based** open-source models, (ii) **autoregressive** open-source models, and (iii) **closed-source** models. Within each group, the highest score is shown in bold and color-coded for that group.

Model	Overall		Basic Following						Advanced Following										Designer							
			Avg			Attribute		Relation		Reasoning		Avg			Attribute + Relation		Attribute + Reasoning		Relation + Reasoning		Style		Text		Real World	
	short	long	short	long	short	long	short	long	short	long	short	short	long	short	long	short	long	short	long	short	long	short	long	short	long	
<b>Diffusion based Open-Source Models</b>																										
FLUX-1 dev	66.24	<b>66.72</b>	<b>74.41</b>	76.67	72.50	75.50	78.20	<b>79.78</b>	<b>72.52</b>	74.73	60.72	60.95	66.76	65.50	61.76	60.74	56.60	57.49	63.33	60.00	44.49	<b>54.75</b>	<b>74.63</b>	<b>72.01</b>		
SD XL	51.13	43.92	61.62	50.52	56.50	44.00	71.02	58.50	57.35	49.06	46.97	41.98	56.13	50.83	40.87	37.78	45.16	39.32	66.67	66.67	17.19	1.36	49.25	47.76		
SD 3	64.79	64.79	73.97	<b>76.92</b>	<b>78.50</b>	<b>78.50</b>	77.06	77.16	66.36	<b>75.08</b>	58.48	62.58	62.90	67.67	53.47	61.28	59.50	62.85	56.67	70.00	57.01	22.62	71.64	67.91		
SD 3.5	<b>68.69</b>	64.92	73.72	72.10	77.50	66.50	74.79	77.16	68.87	72.64	<b>65.59</b>	<b>63.41</b>	<b>70.85</b>	<b>68.22</b>	<b>65.03</b>	<b>62.93</b>	<b>61.03</b>	<b>61.66</b>	56.67	60.00	<b>73.30</b>	46.15	70.15	69.03		
SANA Sprint	61.55	58.49	72.25	68.04	70.00	66.50	77.27	68.59	69.48	69.04	58.32	58.64	64.54	67.25	52.88	54.04	60.46	60.76	73.33	53.33	19.91	14.93	66.04	72.01		
SANA 1.5	62.57	63.48	73.92	72.31	71.50	73.00	<b>82.21</b>	78.39	68.04	65.52	60.36	60.36	65.65	67.33	56.41	56.13	62.20	60.18	66.67	<b>76.67</b>	28.51	23.53	61.94	70.52		
Playground v2	45.38	53.26	54.29	67.14	47.50	65.00	65.22	74.60	50.15	61.83	44.45	52.22	51.12	60.01	41.16	47.57	45.06	53.72	56.67	63.33	0.45	4.07	51.12	49.25		
Playground v2.5	46.18	50.99	59.50	63.21	53.00	61.00	69.10	77.91	56.40	50.71	43.34	48.62	55.33	58.21	40.20	40.74	38.12	49.89	56.67	66.67	0.90	6.79	45.90	47.76		
PixArt-delta	41.65	46.87	49.86	55.37	39.50	45.50	58.44	70.43	51.62	50.19	40.38	44.32	47.56	49.03	39.42	35.52	37.12	49.98	56.67	73.33	0.45	2.71	44.03	45.15		
PixArt-alpha	43.42	50.71	53.32	58.95	49.50	54.50	63.43	72.17	47.02	50.18	43.14	46.74	55.23	48.04	36.58	42.55	42.79	50.90	50.00	76.67	1.36	3.17	51.49	54.85		
PixArt-sigma	57.46	57.04	67.74	68.19	65.50	69.50	74.33	72.11	63.40	62.96	56.71	54.52	62.47	59.67	57.51	55.08	54.84	52.64	<b>76.67</b>	73.33	2.71	4.98	63.06	63.06		
LUMINA-Next	46.83	51.81	59.62	62.48	49.50	61.50	63.30	65.51	66.04	60.44	43.72	47.20	47.52	51.35	42.65	42.06	44.90	50.87	53.33	66.67	2.71	6.33	51.49	61.57		
Hunyuan-DiT	49.14	52.67	65.39	67.79	59.00	63.00	79.89	76.82	57.27	63.56	51.61	52.25	62.49	59.93	49.14	45.71	49.38	54.74	53.33	73.33	0.45	2.26	31.34	34.70		
<b>AR based Open-Source Models</b>																										
Llamagen	41.21	40.35	49.92	49.58	44.50	47.00	56.28	56.17	48.98	45.58	42.17	40.44	42.52	42.70	43.70	38.53	45.36	44.10	40.00	43.33	3.62	5.43	45.90	40.30		
LightGen	52.84	46.42	68.70	53.99	61.00	52.00	73.69	54.52	71.40	50.52	54.10	45.76	66.82	48.37	52.22	42.93	51.07	50.64	43.33	43.33	2.26	10.86	53.73	59.70		
Show-o	57.34	59.24	69.99	<b>71.30</b>	66.50	<b>78.24</b>	<b>76.47</b>	69.88	67.00	69.24	58.25	<b>59.89</b>	67.21	64.33	54.26	56.75	61.08	56.19	46.67	66.67	4.98	11.31	71.64	68.66		
Infinity	60.65	59.66	70.90	71.63	73.00	73.00	73.75	<b>74.44</b>	65.96	67.44	59.80	57.81	<b>68.92</b>	63.78	60.53	<b>56.87</b>	55.04	<b>56.81</b>	<b>56.67</b>	<b>73.33</b>	22.17	26.70	69.78	61.19		
JanusPro	<b>65.38</b>	<b>61.10</b>	<b>74.99</b>	73.19	<b>74.50</b>	78.00	73.69	70.51	<b>76.77</b>	<b>71.04</b>	<b>61.77</b>	56.03	65.71	<b>66.48</b>	<b>62.01</b>	55.62	<b>61.16</b>	49.34	43.33	70.00	<b>38.46</b>	<b>42.08</b>	<b>79.48</b>	<b>73.51</b>		
<b>Closed-Source Models</b>																										
DALL-E 3	74.47	72.94	77.35	78.40	77.62	75.00	80.22	79.67	74.22	80.54	70.11	68.45	76.65	75.05	68.39	68.07	63.64	59.92	79.31	80.00	74.07	75.51	76.12	62.69		
MidJourney v6	77.51	69.83	73.92	73.87	76.83	76.50	81.80	81.61	63.13	63.49	67.81	64.81	74.75	73.96	61.07	59.70	60.11	60.81	<b>93.33</b>	33.33	95.48	95.93	84.42	83.12		
MidJourney v7	65.92	62.43	73.96	74.63	75.00	82.00	78.74	78.51	68.12	68.55	63.44	62.59	70.60	74.03	64.43	59.58	58.84	61.34	66.67	33.33	31.67	34.39	79.22	75.32		
FLUX-1 Pro	63.75	63.53	71.39	73.57	70.00	68.50	68.51	79.97	75.66	72.23	64.63	61.42	70.69	72.99	62.34	57.27	64.65	57.11	63.00	63.00	34.39	36.65	69.94	66.78		
GPT-4o	<b>84.19</b>	<b>84.61</b>	<b>85.30</b>	<b>86.55</b>	<b>81.00</b>	<b>82.12</b>	<b>86.16</b>	<b>84.12</b>	<b>88.74</b>	<b>94.50</b>	<b>81.24</b>	<b>79.75</b>	<b>81.95</b>	<b>81.55</b>	<b>80.03</b>	<b>79.85</b>	<b>80.88</b>	<b>75.68</b>	76.67	<b>86.67</b>	<b>92.76</b>	<b>90.05</b>	<b>89.55</b>	<b>88.06</b>		

Table 9: Similarity of evaluation results between GPT-4o and QwenVL2.5-72B, measured by *Spearman*  $\rho$ .

T2I Model Groups	Overall Performance on TIIF-BENCH																	
	Short Prompts						Long Prompts											
Diffusion-based Open-Source Models	0.973																	
AR-based Open-Source Models	1.000																	
Closed-Source Models	1.000																	
All Models	0.981																	

*Prompt: Pirate ship battle with exploding cannonballs and plank-walking in LEGO style*



Figure 10: **Style Control** evaluates a T2I model’s ability to manage global image quality, content comprehension, and overall aesthetic coherence. We provide illustrative examples for eight representative T2I models on this dimension.

*Prompt: A picture of a charming illustration showcasing a teacup and a slice of cake in warm, inviting tones, with the text on it: "drink", "tea", "eat", "cake", "relax", "enjoy", "delight".*



Figure 11: **Text Rendering** is introduced as a novel evaluation dimension for assessing a model’s ability to generate complex, non-natural textures—embedded human language text. We present illustrative examples for eight representative T2I models on this dimension.

*Prompt: a golden-haired, eyeglass-free elderly American man presenting a gold medal engraved with "USA's Best BOY" to a middle-aged man with short dark hair, wearing a yellow t-shirt with the text "Tarrif is the new trick" written in bold*



Figure 12: **Designer-level** prompts comprise the densest and most diverse set of requirements, offering the most comprehensive test of a model’s instruction-following capabilities. We provide illustrative examples showing how eight representative T2I models perform on this dimension.

Table 10: Performance of state-of-the-art open-source models on TIIF-Bench **whole set** evaluated by GPT4-o. Evaluated systems are grouped into (i) **diffusion-based** open-source models and **autoregressive** open-source models. Within each group, the highest score is shown in bold and color-coded for that group.

Model	Overall		Basic Following				Advanced Following								Designer									
			Avg	Attribute	Relation	Reasoning	Avg	Attribute + Relation	Attribute + Reasoning	Relation + Reasoning	Style	Text												
	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long								
<b>Diffusion based Open-Source Models</b>																								
FLUX.I dev	63.47	<b>67.32</b>	70.03	<b>71.84</b>	77.75	<b>78.80</b>	78.02	<b>78.67</b>	54.31	<b>58.04</b>	62.69	65.11	66.95	69.74	60.99	60.75	62.47	<b>64.66</b>	63.33	72.67	43.25	<b>58.91</b>	64.14	63.65
SD 3	66.83	63.69	70.23	69.73	80.05	78.15	76.14	74.83	54.50	56.22	64.96	<b>64.62</b>	72.88	<b>72.48</b>	59.44	61.39	62.10	63.15	75.33	77.00	58.24	26.87	<b>62.78</b>	63.15
SD 3.5	<b>69.59</b>	64.96	<b>71.74</b>	70.07	<b>80.25</b>	78.35	78.61	75.61	<b>56.38</b>	56.24	66.90	62.69	<b>73.41</b>	68.33	60.73	57.30	<b>64.51</b>	62.95	78.33	71.67	<b>71.79</b>	49.76	62.28	64.39
SANA 1.5	65.17	62.17	70.18	69.03	76.85	75.25	<b>78.73</b>	77.60	54.97	54.23	65.75	62.45	72.42	68.66	<b>63.84</b>	<b>61.76</b>	63.69	60.66	<b>93.00</b>	81.00	16.91	14.13	66.13	<b>66.25</b>
Playground v2	46.30	54.63	53.10	63.31	55.75	67.00	63.44	73.87	40.11	49.06	44.97	54.82	50.90	63.34	39.48	49.99	46.95	54.37	69.33	81.67	1.10	2.16	49.63	50.25
Playground v2.5	46.34	54.04	54.33	63.04	57.75	69.25	63.32	74.07	41.90	45.79	45.41	54.31	53.14	63.90	39.69	49.52	46.40	52.45	65.33	80.00	1.48	5.08	48.01	46.28
PixArt-delta	43.32	48.92	50.62	55.52	50.55	54.90	62.86	68.07	38.44	43.60	42.05	47.37	47.45	51.38	37.40	43.00	43.50	48.74	66.33	85.67	0.19	1.15	43.18	43.80
PixArt-alpha	45.83	51.12	52.27	57.87	54.95	58.75	63.02	70.34	38.83	44.53	45.17	50.04	52.79	56.62	39.76	45.80	45.91	49.27	66.00	86.00	0.72	1.39	50.50	47.39
PixArt-sigma	58.17	58.80	64.88	67.21	70.30	75.00	71.59	74.88	52.75	51.75	58.80	58.09	65.13	65.74	56.08	54.86	58.28	56.80	89.33	<b>87.33</b>	3.45	3.88	56.58	58.93
Lumina-Next	50.37	49.83	57.32	57.49	61.20	60.85	64.93	66.54	45.83	45.08	50.27	48.23	55.25	54.13	45.84	43.07	52.55	49.70	77.00	77.67	1.01	0.96	49.75	50.50
Hunyuan-DiT	50.22	53.63	61.92	64.57	66.85	71.35	70.92	73.45	48.00	48.92	52.86	52.94	61.37	59.52	49.99	48.51	53.41	53.71	55.67	81.67	0.62	1.01	45.16	44.54
<b>AR based Open-Source Models</b>																								
LlamaGen	40.40	39.85	48.42	50.75	45.70	54.15	60.96	58.56	38.58	39.55	39.75	38.77	42.34	38.69	37.48	37.57	42.58	43.87	53.00	45.00	1.10	1.96	41.81	39.33
LightGen	52.77	46.31	61.49	51.68	65.90	58.80	69.26	54.94	49.30	41.31	54.01	46.14	59.16	48.93	52.38	45.41	55.29	47.55	67.00	58.00	2.63	6.47	53.97	55.33
Show-o	56.94	58.17	66.28	<b>69.56</b>	72.20	<b>78.30</b>	74.08	<b>76.52</b>	52.56	<b>53.85</b>	59.46	58.76	66.92	66.38	56.03	55.72	61.29	59.31	68.67	72.00	3.16	4.65	57.57	56.82
Infinity	63.11	60.83	69.55	67.02	75.70	76.15	<b>78.18</b>	74.27	54.77	50.63	64.58	60.95	70.98	68.94	<b>62.25</b>	58.42	63.87	58.33	<b>81.00</b>	<b>79.67</b>	21.17	19.54	60.05	61.54
Janus-Pro	<b>64.41</b>	<b>62.19</b>	<b>70.97</b>	68.15	<b>77.10</b>	76.55	76.84	76.45	<b>58.98</b>	51.45	<b>64.93</b>	<b>62.87</b>	<b>71.76</b>	<b>71.59</b>	61.88	<b>59.57</b>	<b>64.48</b>	<b>60.46</b>	71.00	68.33	<b>32.14</b>	<b>33.19</b>	<b>65.51</b>	<b>62.16</b>

Table 11: Performance of state-of-the-art open-source models on TIIF-Bench **whole set** evaluated by QwenVL2.5-72B. Evaluated systems are grouped into (i) **diffusion-based** open-source models and **autoregressive** open-source models. Within each group, the highest score is shown in bold and color-coded for that group.

Model	Overall		Basic Following						Advanced Following										Designer						
			Avg		Attribute		Relation		Reasoning		Avg		Attribute +Relation		Attribute +Reasoning		Relation +Reasoning		Style		Text		Real World		
	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	
<b>Diffusion based Open-Source Models</b>																									
FLUX.1 dev	65.96	<b>65.74</b>	<b>71.93</b>	<b>68.43</b>	<b>79.52</b>	74.25	<b>79.46</b>	<b>75.90</b>	<b>56.82</b>	55.15	60.46	<b>63.05</b>	<b>70.84</b>	67.07	<b>59.67</b>	58.55	<b>64.06</b>	63.06	60.00	70.00	44.49	<b>59.82</b>	67.51	67.87	
SD XL	53.40	44.97	58.92	50.77	62.30	52.35	69.87	60.55	44.58	39.41	51.02	43.25	57.00	49.28	43.80	38.66	52.70	44.41	80.67	63.00	17.96	2.87	51.74	54.22	
SD 3	65.84	63.30	67.89	67.81	76.35	73.95	74.06	73.50	53.24	<b>55.98</b>	62.68	62.88	67.98	<b>68.61</b>	57.96	<b>59.04</b>	61.08	<b>63.16</b>	74.00	76.67	59.39	31.80	<b>68.49</b>	67.00	
SD 3.5	<b>67.21</b>	63.98	68.64	66.98	76.80	<b>74.40</b>	75.88	73.25	53.24	53.30	<b>63.51</b>	61.96	69.76	65.87	57.73	57.78	61.03	62.62	69.33	67.33	<b>73.71</b>	53.40	67.37	67.87	
SANA Sprint	59.45	56.42	65.32	64.08	70.00	68.10	74.61	72.30	51.36	51.85	59.06	55.85	65.88	62.06	53.64	51.40	60.51	58.09	76.67	64.33	18.77	15.42	63.65	64.27	
SANA 1.5	62.96	60.88	68.09	66.00	73.50	70.70	77.05	74.49	53.71	52.81	62.60	59.75	67.78	64.89	58.64	55.10	63.78	61.75	79.33	75.00	26.63	24.47	66.25	<b>68.73</b>	
Playground v2	45.51	52.73	51.89	60.88	53.65	63.75	62.20	72.26	39.83	46.64	44.11	52.28	48.32	58.74	39.81	46.10	47.24	55.06	61.33	74.00	2.54	6.03	54.71	51.99	
Playground v2.5	45.13	52.36	52.96	61.06	55.90	65.90	62.78	72.24	40.20	45.04	44.36	52.02	51.00	60.62	39.96	45.94	45.72	52.85	58.00	69.67	1.77	7.66	50.87	51.36	
PixArt-delta	42.33	47.18	48.53	53.66	47.65	52.50	60.40	66.78	37.53	41.69	41.67	45.56	45.60	50.44	38.55	40.71	43.98	47.09	58.00	77.00	0.53	1.53	48.76	46.90	
PixArt-alpha	45.55	51.11	51.42	56.16	51.95	56.60	62.80	68.62	39.49	43.27	44.89	48.73	51.02	53.62	39.74	43.98	47.00	49.65	63.33	<b>84.33</b>	1.77	4.60	52.85	55.33	
PixArt-sigma	56.37	56.87	62.34	63.58	66.75	69.35	71.14	71.97	49.14	49.42	55.60	55.70	60.95	61.95	52.47	53.13	56.14	54.79	<b>83.67</b>	83.00	5.56	6.08	61.54	62.16	
LUMINA-Next	48.77	48.78	55.09	55.76	55.25	58.15	64.11	65.24	45.90	43.89	47.20	46.04	50.63	51.03	43.68	41.33	49.55	48.25	71.00	68.00	5.17	3.98	53.60	59.18	
Hunyuan-DiT	48.49	53.29	61.38	62.96	65.90	69.50	71.70	72.44	46.54	46.94	51.78	53.90	60.66	61.47	47.96	48.87	53.97	54.86	44.00	77.67	1.53	2.16	44.17	45.66	
<b>AR based Open-Source Models</b>																									
Llamagen	42.03	40.85	49.25	49.59	48.20	50.90	60.02	58.44	39.53	39.45	40.95	39.77	42.85	39.72	39.54	38.97	43.82	44.34	51.67	43.00	3.40	6.85	49.26	46.03	
LightGen	52.67	47.45	61.60	53.78	63.10	59.90	71.44	58.43	50.26	43.01	53.39	47.56	59.48	51.27	50.67	46.07	55.56	50.14	59.00	50.00	3.45	6.61	61.04	61.66	
Show-o	56.94	58.33	66.11	<b>67.47</b>	71.60	<b>75.35</b>	74.34	<b>73.90</b>	52.39	53.17	57.07	58.22	64.52	65.69	53.98	55.34	58.13	58.57	62.67	<b>64.67</b>	8.09	12.31	66.75	66.00	
Infinity	61.31	58.49	66.40	65.00	73.40	73.15	74.02	72.75	51.79	49.10	61.24	57.89	66.96	65.66	58.78	54.15	60.85	57.20	<b>69.67</b>	64.00	29.89	25.05	66.38	65.38	
JanusPro	<b>64.06</b>	<b>61.84</b>	<b>69.44</b>	66.82	<b>75.35</b>	75.10	<b>74.50</b>	72.10	<b>58.46</b>	<b>53.26</b>	<b>62.95</b>	<b>60.96</b>	<b>68.86</b>	<b>69.03</b>	<b>60.02</b>	<b>57.36</b>	<b>62.46</b>	<b>59.37</b>	62.00	54.00	<b>43.82</b>	<b>45.02</b>	<b>71.09</b>	<b>71.34</b>	