OneIG-Bench: Omni-dimensional Nuanced Evaluation for Image Generation

Jingjing Chang[1,2] **Yixiao Fang**[2,†] **Peng Xing**[2] **Shuhan Wu**[2] **Wei Cheng**[2]
Rui Wang[2] **Xianfang Zeng**[2] **Gang Yu**[2,‡] **Hai-Bao Chen**[1,‡]

[1] SJTU [2] StepFun
[†] Project lead [‡] Corresponding author

Dataset Code Project Page

Abstract

Text-to-image (T2I) models have garnered significant attention for generating high-quality images aligned with text prompts. However, rapid T2I model advancements reveal limitations in early benchmarks, lacking comprehensive evaluations, for example, the evaluation on reasoning, text rendering and style. Notably, recent state-of-the-art models, with their rich knowledge modeling capabilities, show promising results on the image generation problems requiring strong reasoning ability, yet existing evaluation systems have not adequately addressed this frontier. To systematically address these gaps, we introduce **OneIG-Bench**, a meticulously designed comprehensive benchmark framework for fine-grained evaluation of T2I models across multiple dimensions, including prompt-image alignment, text rendering precision, reasoning-generated content, stylization, and diversity. By structuring the evaluation, this benchmark enables in-depth analysis of model performance, helping researchers and practitioners pinpoint strengths and bottlenecks in the full pipeline of image generation. Specifically, **OneIG-Bench** enables flexible evaluation by allowing users to focus on a particular evaluation subset. Instead of generating images for the entire set of prompts, users can generate images only for the prompts associated with the selected dimension and complete the corresponding evaluation accordingly. Our codebase and dataset are now publicly available to facilitate reproducible evaluation studies and cross-model comparisons within the T2I research community.

1 Introduction

Recent years have witnessed remarkable advancements in text-to-image (T2I) models across image quality, semantic alignment, text rendering precision, and knowledge-driven reasoning in image generation [55, 46, 19, 60, 38, 53, 45]. However, the development of evaluation systems has significantly lagged behind model progress: most existing benchmarks remain confined to single-dimensional assessments, lacking comprehensiveness. For instance, T2ICompBench [34], GenEval [25], and DSG-1k [16] focus on short-text semantic understanding, while DPG-Bench [31] introduces dense prompt evaluation but lackly coverage of limited dimensions like style and text. Although WorldGenBench [80] addresses world knowledge and reasoning, a more comprehensive multi-dimensional evaluation framework is urgently needed to provide scientific model assessments and guide technological development.

To drive the advancement of text-to-image models, we posit that the development of a holistic benchmark framework capable of evaluating models across multiple interconnected dimensions is

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

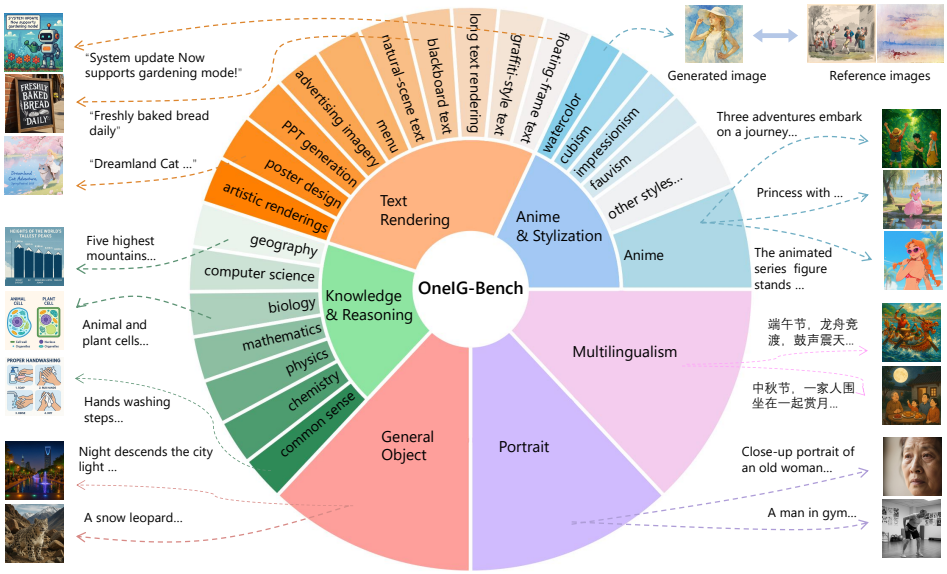


Figure 1: **Overview of OneIG-Bench**. OneIG-Bench comprises six core categories, each designed to evaluate targeted capabilities across distinct generative dimensions, with approximately 200 carefully curated prompts per category to ensure comprehensive coverage of diverse scenarios.

imperative for fostering rigorous, comprehensive assessment. We introduce **OneIG-Bench** shown in Figure 1, comprising over 1000 omni-dimensional prompts, primarily sourced from real-world user inputs, systematically designed to comprehensively evaluate text-to-image models across diverse generative capabilities. Based on distinct generative themes, we classify the evaluation dataset into six core assessment categories: *General Object*, *Portrait*, *Anime and Stylization*, *Text Rendering*, *Knowledge and Reasoning* and *Multilingualism*. By leveraging our framework for the taxonomic differentiation of generative themes, we enable a more nuanced evaluation of models' multidimensional capabilities. Consequently, this systematic assessment allows users to identify models with tailored capabilities that align precisely with their specific application requirements.

For different evaluation dimensions, we have carefully devised quantitative indicators, taking into account various factors to ensure the comprehensiveness and objectivity of the evaluation. These indicators are designed to precisely measure the model's performance in different aspects related to specific subject domains, enabling a more accurate and in-depth assessment. Specifically for the evaluation of *General object*, *Portrait*, *Anime and Stylization*, we have developed an evaluation system for the ability to comply with input prompts. Regarding *Anime and Stylization*, we also incorporate the stylistic similarity into our evaluation system, specifically designed to assess models' capabilities in reproducing diverse artistic styles. In the *Text Rendering* evaluation segment, we focus on three core metrics: the edit distance between generated text and ground truth, the text completion rate in one visual output, and the overall text generation accuracy. For the *Knowledge and Reasoning* part, our assessment centers on whether the model possesses the required domain knowledge and can accurately interpret user intent, thereby enabling the generation of semantically coherent and logically consistent images. For the *Multilingualism* part, we evaluate the alignment between the images generated from cultural element prompts and the corresponding cultural elements. Additionally, we also evaluate the diversity across all six dimensions.

We summarize our key contributions in the following three points:

- We present **OneIG-Bench**, which consists of six prompt sets, with the first five — 245 *Anime and Stylization*, 244 *Portrait*, 206 *General Object*, 200 *Text Rendering*, and 225 *Knowledge and Reasoning* prompts — each provided in both English and Chinese, and 200 *Multilingualism* prompts, designed for the comprehensive evaluation of current text-to-image models.
- A systematic quantitative evaluation is developed to facilitate objective capability ranking through standardized metrics, enabling direct comparability across models. Specifically, our evaluation framework allows T2I models to generate images only for prompts associated

with a particular evaluation dimension, and to assess performance accordingly within that dimension.

- State-of-the-art open-sourced methods as well as the proprietary model are evaluated based on our proposed benchmark to facilitate the development of text-to-image research.

2 Related Works

2.1 Text to Image Models

Text-to-image (T2I) generation aims to develop models that produce images semantically consistent with given text descriptions. Early explorations during the generative adversarial network (GAN) [26, 57] era laid foundational work, but these methods suffered significant limitations due to mode collapse, often failing to generate even simple subject matters accurately. In recent years, generative approaches based on the diffusion model paradigm have emerged as a dominant trend [30, 59]. Notable advancements include Unet-based architectures like Stable Diffusion XL [46], Diffusion Image Transformer (DiT)-based models [9, 10, 42], double-stream MMDiT frameworks [60, 19], and hybrid designs combining double-stream and single-stream networks [78, 73, 38]. These models have achieved remarkable progress in generating high-quality images that closely align with textual semantics. With technological advancements, the evaluation framework for T2I models must evolve from single-dimensional attribute assessments (e.g., color, shape) to multi-layered evaluations, encompassing semantic alignment, stylistic consistency, and text rendering accuracy. Meanwhile, autoregressive-based models [12, 51, 54, 61, 77, 63, 68, 74] have demonstrated unique strengths in knowledge modeling and complex text comprehension, necessitating the integration of systematic reasoning capability evaluations into the assessment framework. In summary, the rapid development of T2I generation underscores the urgent need for a comprehensive and rigorous evaluation system to accurately measure model performance, identify strengths and weaknesses, and foster sustainable progress in the field.

2.2 Text to Image Evaluation

In the early stages of text-to-image development, researchers typically employed some metrics to evaluate image generation quality, such as FID [28](Fréchet Inception Distance), SSIM [69](Structural Similarity Index Measure), PSNR(Peak Signal to Noise Ratio), etc. However, these methods do not provide a comprehensive understanding of the model's capabilities and fail to capture the model's ability to comprehend higher-order semantics.

In recent years, the technology of text-to-image models has been evolving rapidly, and the corresponding evaluation system urgently needs to be innovated and upgraded. Taking the evaluation of Stable Diffusion 1.5 [55] and Stable Diffusion XL [46] as examples, most of the existing benchmark evaluations (Attend [7], Hrs-bench [5], CC500 [20]) focus on judging the degree of restoration of the core elements in the prompts. In the face of the rapid evolution of model technology, it has become difficult to comprehensively and accurately measure the actual performance and innovative potential of the models. Subsequently, evaluation methods such as PartiPrompt [77], DrawBench [56], TIFA [32], Gecko [70], EvalAlign [62], T2ICompBench [34], T2ICompBench++ [33], GenEval [25], EvalMuse [27], DPG-Bench [31], and GenAI-Bench [39] introduced Visual Language Models (CLIP [49],BLIP [41], MLLMs [3, 67, 14]) as evaluators. These approaches aim to maximize the utilization of model capabilities for jointly assessing prompts and images. However, these evaluation methods primarily focus on the prompt following ability of text-to-image models, often neglecting other critical aspects. Some alternative methods(MJHQ-30K [40], HPSv2 [72], Pick-a-pic [37]) have attempted to assess the aesthetic quality of images generated by these models. Recently, to keep up with the advances in text-to-image models, evaluation frameworks such as WISE [43], WorldGenBench [80], Commonsense-T2I [22], and PhyBench [48] have been developed to assess the models' knowledge and reasoning capabilities.

3 Benchmark

3.1 Benchmark Overview

With the rapid advancement of text-to-image (T2I) models, the existing evaluation frameworks for T2I models urgently require improvement to measure model strengths and weaknesses comprehensively. Early studies have explored evaluation methods from diverse perspectives, such as compositional text-to-image generation tasks [7, 34, 22], including concept correlation, attribute binding (focusing on color attributes), and spatial relationship modeling. However, evaluations solely focusing on compositional content exhibit significant limitations [25, 31, 27], failing to adequately address broader natural language understanding and other quantitative image assessment dimensions.

In response, recent evaluation frameworks like GenEval [25], EvalMuse [27], and DPG-Bench [31] adopt an object-centric structured paradigm to quantify T2I model performance on specific tasks. Nevertheless, these methods predominantly rely on vision-language models (VLMs) [2, 4], object detection models [15, 11] or visual question answering (VQA) [1] models for element-level alignment assessment, suffering from notable deficiencies in evaluating style consistency and text rendering accuracy, with a lack of high-precision metrics. Additionally, human-based evaluation is prohibitively costly in terms of time and resources, making automated evaluation with limited prompts increasingly critical. Notably, with the rapid development of reasoning-oriented models [45], this study proactively introduces reasoning task evaluation to accurately measure models' knowledge representation and reasoning-driven image generation capabilities.

Table 1 systematically reviews the advantages and disadvantages of recent evaluation methods, presenting a comprehensive framework named OneIG-Bench from six dimensions: scene coverage, prompt distribution diversity, evaluation content, multilingualism support, automation level and leaderboard availability. OneIG-Bench covers core T2I scenarios (e.g., style image generation, text rendering, reasoning-based drawing), supports multi-format prompt inputs (including long/short texts and phrase/tag-based prompts), and employs customized automated metrics for different scenarios. Experimental results demonstrate that this framework effectively identifies performance bottlenecks and strengths of current models, providing a scientific basis for T2I model optimization. **Hereafter, unless otherwise specified, OneIG-Bench refers to OneIG-Bench-EN.**

Table 1: **Comparison between OneIG-Bench and other previous benchmarks**. In the column of prompt diversity, **L** denotes long prompt, **S** denotes the short prompt, **NP** denotes the natural language prompt, **T** denotes the tag-based prompt, and **P** denotes the phrase-based prompt.

Benchmark	Scenes				Prompt Diversity			Evalution				Multilingualism	Auto Eval	Leaderboard
	General	Style	Text	Reason	Length	Type	Count	Alignment	Text	Style	Diversity			
PartiPrompt [77]	✓	✗	✗	✗	L, S	NP	1,600	✗	✗	✗	✗	✗	✗	✗
DrawBench [56]	✓	✗	✓	✗	S	NP	200	✗	✗	✗	✗	✗	✗	✗
TIFA [32]	✓	✗	✗	✗	S	NP	4,000	✓	✗	✗	✗	✗	✓	✗
T2ICompBench [34]	✓	✗	✗	✗	S	NP	6,000	✓	✗	✗	✗	✗	✓	✗
GenEval [39]	✓	✗	✗	✗	S	NP	553	✓	✗	✗	✗	✗	✓	✗
EVALALIGN [62]	✓	✓	✓	✓	S	NP	100	✓	✗	✗	✗	✗	✓	✗
WISE [43]	✓	✗	✗	✓	S	NP	1,000	✓	✗	✗	✗	✗	✓	✗
EvalMuse [27]	✓	✗	✗	✗	S	NP	199	✓	✗	✗	✗	✗	✓	✓
DPG-Bench [31]	✓	✗	✗	✗	L	NP	1,065	✓	✗	✗	✗	✗	✓	✗
OneIG-Bench	✓	✓	✓	✓	L, S	NP,T,P	2,440*	✓	✓	✓	✓	✓	✓	✓

∗:**OneIG-Bench** consists of two subsets: **OneIG-Bench-EN** and **OneIG-Bench-ZH**. **OneIG-Bench-EN** includes 245 *Anime and Stylization* prompts, 244 *Portrait* prompts, 206 *General Objectz* prompts, 200 *Text Rendering* prompts, and 225 *Knowledge and Reasoning* prompts. **OneIG-Bench-ZH** comprises manually translated and verified Chinese versions of the prompts in **OneIG-Bench-EN**, along with additional prompts from the Multilingualism category, totaling $1,320$ prompts. In total, **OneIG-Bench** contains $2,440$ prompts.

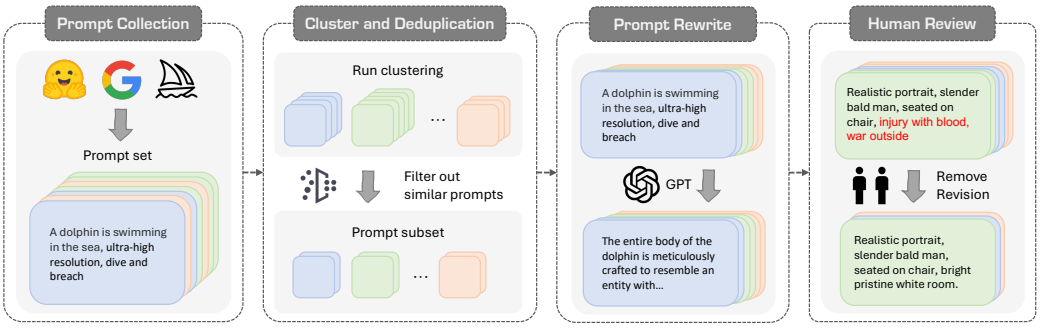


Figure 2: **The construction pipeline of OneIG-Bench**. The construction pipeline comprises four methodical steps to derive our final assessment prompts, ensuring the diversity and comprehensiveness of the benchmark.

3.2 Benchmark Construction

In constructing the evaluation prompt set, as illustrated in Figure 2, we have established four core steps to ensure the diversity and comprehensiveness of the prompt set, aligning with the design principles of our benchmark framework.

In the first step, we curated prompts and generation scenes by filtering publicly accessible internet data, user inputs, and some established datasets, thereby ensuring that the benchmark focuses on content aligned with real-world user needs rather than rare or specialized contexts.

In the second step, we apply a clustering approach to balance the distribution of prompts across different scenes and semantic dimensions, ensuring that no single category dominates. Within each cluster, we identify some prompts with high semantic overlap, which can compromise evaluation diversity; thus, we implement a deduplication pipeline that filters out redundant prompts based on cosine similarity to the cluster center embeddings. Besides, we endeavor to maintain a relatively consistent proportion of prompts across the five defined dimensions during prompt selection.

Following deduplication and subset curation, we employ a large language model (LLM) to rewrite the original prompts. Concurrently, constraints are applied to the word-level length distribution of prompts, enabling a structured analysis of model performance across varying text complexities. The prompt corpus is intentionally structured into three length categories: concise texts (fewer than 30 words), mid-complexity scenarios (30–60 words), and elaborate texts (exceeding 60 words). The corresponding ratio of three categories is around 1:2:1.

Finally, we performed manual reviews to filter out prompts containing sensitive content or conflicting semantics, ensuring the rationality of all benchmark prompts. This critical step not only enhances the dataset's quality and reliability but also guarantees its suitability for fair and unbiased model evaluation across diverse generative scenarios.

Through a rigorously designed construction pipeline, this structured evaluation framework facilitates a granular assessment of model performance across multiple dimensions.

4 Evaluation

4.1 Metrics

To illustrate the metrics used in our benchmark, we provide the following general definitions. Given the text prompt set $\mathbb{T} = \{T_1, T_2, ..., T_n\}$, for each $T_k \in \mathbb{T}$, the generated images are defined as $\mathbb{G}^k = \{G_1^k, G_2^k, ..., G_m^k\}$, where m denotes the number of generated images, and a total of $m \times n$ images are generated as evaluation images.. To evaluate the style score, for each style-specific prompt $T_k \in \mathbb{T}$, we define $\mathbb{R} = \{R_1^k, R_2^k, R_3^k\}$ as the set of corresponding style reference images, where l denotes the number of reference images. For text rendering, the original target text string in T_k is defined as s_k, and the generated string in the corresponding image is defined as $\hat{s}_k$.

Semantic alignment. We follow the method introduced in DSG [16] on *General Object*, *Portrait*, *Anime and Stylization* to assess the semantic matching degree of each text-image sample. For each prompt, we initially leverage GPT-4o [44] to generate a question dependency graph. In the process

of constructing the graph, our focus lies in formulating questions related to the overall information, spatial relationships, and the attributes of diverse objects. During the evaluation, Qwen2.5-VL-7B [4] is utilized to answer questions derived from the corresponding prompt and the generated image. A score of 1 is assigned for each correctly answered question. However, when calculating the aggregate score for a prompt, leaf node scores are conditionally validated: they contribute to the total score only if the root node question is answered correctly; otherwise, leaf node scores are reset to 0. The final score for each prompt is computed as the sum of validated scores divided by the total number of questions.

Text Rendering. To accurately evaluate the text-generation capability of text-to-image models, we designed specialized text evaluation metrics. To extract the generated string $\hat{s}_k$, we first use a state-of-the-art Vision-Language Model (VLM, e.g., Qwen2.5-VL-7B [4]) to parse the text string and then clean it by removing symbols and consecutive spaces The metrics are as follows:

(1) **Edit Distance (ED)**: It is defined as the average edit distance between the generated text of evaluation images and the ground-truth text to be generated. We define the edit distance score of the i-th evaluation image as $\text{ED}_i = \mathcal{L}(\hat{s}_i, s_i)$, where $\mathcal{L}(\cdot)$ denotes the Levenshtein distance function. Therefore, the overall edit distance score of the model is: $\text{ED} = \frac{1}{n \times m} \sum_{i=1}^{n \times m} \text{ED}_i$.

(2) **Completion Rate (CR)**: It is defined as the proportion of the number of evaluation images with completely correct generated text to the total number of evaluation images. We define that the score for the i-th generated image in this criterion is $\text{CR}_i = 1$ if and only if the edit distance score of the i-th image is 0, *i.e.*, $\text{ED}_i = 0$. Therefore, the overall CR score of the model is defined as $\text{CR} = \sum_{i=1}^{m \times n} \text{CR}_i / (m \times n)$.

(3) **Word Accuracy (WAC)**: It is defined as a metric representing the ratio of all correctly generated words to the total number of words in the original target text strings among all prompts.

Based on our analysis on the evaluation results, we define the edit distance upper bound as ϕ, and edit distance exceeding ϕ indicate deficiencies in the model's text-rendering capability. To facilitate metric ranking and readability, we integrated three metrics into a composite metric and defined the text score(S_{text}) as follows:

$$\text{S}_{\text{text}} = 1 - \min(\phi, \text{ED}) * (1 - \text{CR}) * (1 - \text{WAC}) / \phi \tag{1}$$

where $\phi = 100$ in OneIG-Bench. Considering that Chinese characters typically occupy twice as many bytes as English letters, we use $\phi = 50$ in Equation 1 when computing the text score for OneIG-Bench-ZH, in order to maintain a comparable normalization scale.

Knowledge and Reasoning. We perform the evaluations using GPT-4o [44] and LLM2CLIP [35]. Specifically, GPT-4o is responsible for generating the textual reasoning answers, which serve as the core reference for evaluation. LLM2CLIP then measures the alignment between text and image by calculating the cosine similarity between the GPT-4o-generated answer and the corresponding generated image.

Style. For stylization evaluation, we curated multiple reference images per style and employed a dual-style extraction framework to mitigate bias and enhance robustness. Specifically, the CSD [58] model and OneIG style image encoder(fine-tuned from CLIP [49], using images generated by CSGO [76]) are leveraged to encode the images and generate the corresponding embeddings. Subsequently, for each encoder, we quantitatively assess the model's style capacity by computing the cosine similarity between the style embeddings of generated images and those of reference images. For each generated image, we choose the maximum similarity as the score of the image. The style similarity($\text{S}_{[\text{csd,oneig}]}$) of one style image encoder is defined as:

$$\text{S}_{[\text{csd,oneig}]} = \frac{1}{n} \sum_{k=1}^{n} \left[\frac{1}{m} \sum_{i=1}^{m} \left(\max_{j} \cos(\mathcal{F}(G_i^k), \mathcal{F}(R_j^k)) \right) \right] \tag{2}$$

where $\mathcal{F}(\cdot)$ denotes the corresponding style image encoder. The final style score S_{style} is defined as $\text{S}_{\text{style}} = (\text{S}_{\text{csd}} + \text{S}_{\text{oneig}})/2$.

Diversity. In addition, we apply a form of similarity calculation to evaluate the diversity of the generation of the model introduced in [23]. The calculation of diversity is defined as follows: for a given model, we first compute the pairwise cosine similarity between every pair of images generated

from the same prompt within a set of multiple generated outputs. These similarities are averaged per prompt to yield an intra-prompt similarity score. We then aggregate these intra-prompt averages across all prompts in the evaluation dataset using a global mean, resulting in an overall diversity metric. Following [23], we also applied DreamSim [21] to compute the cosine similarity and the formula is as follows:

$$\text{SIM}_{ij}^{k} = \cos(\mathcal{F}(G_i^k), \mathcal{F}(G_j^k)) \tag{3}$$

where SIM_{ij}^{k} denotes the cosine similarity between images generated by one text prompt, $\mathcal{F}(\cdot)$ represents DreamSim [21] model. Thus, the diversity score($\text{S}_{\text{diversity}}$) can be defined:

$$\text{S}_{\text{diversity}} = \frac{1}{n}\sum_{k=1}^{n}\left[\frac{1}{C_m^2}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\left(1-\text{SIM}_{ij}^{k}\right)\right]. \tag{4}$$

4.2 Results and Analysis

We evaluate a range of well-known image generation models on our benchmark, including unified multimodal models (Janus-Pro [13], BLIP3-o [8], BAGEL [18], Show-o2 [75], OmniGen2 [71]), open-source models (Stable Diffusion 1.5 [55], Stable Diffusion XL [46], Stable Diffusion 3.5 [60], Flux.1-dev [38], CogView4 [78], SANA [73], Lumina-Image 2.0 [47], and HiDream-I1-Full [29]) with A800 GPUs, as well as closed-source models (Imagen3 [36], Recraft V3 [66], Kolors 2.0 [64], Seedream 3.0 [24], Imagen4 [17] and GPT-4o [45]). To present a comprehensive comparison, we aggregate evaluation metrics across multiple dimensions, as summarized in Table 2. **We define the sets of images generated based on the OneIG-Bench prompt categories *General Object* $\mathcal{O}$, *Portrait* $\mathcal{P}$, *Anime and Stylization* $\mathcal{A}$ (prompts without stylization), $\mathcal{S}$ (prompts with stylization), *Text Rendering* $\mathcal{T}$, *Knowledge and Reasoning* $\mathcal{KR}$ and *Multilingualism* $\mathcal{L}$.** A more detailed, fine-grained analysis of individual dimensions on OneIG-Bench and the quantitive results on OneIG-Bench-ZH follows in the subsequent sections and the appendix.

Table 2: **Overall quantitative comparison of different methods on OneIG-Bench**. The table showcases the results of five core metrics for various methods. indicate the first, second, third, fourth, and fifth performance, respectively.

Method	Alignment ↑	Text ↑	Reasoning ↑	Style ↑	Diversity ↑
Assessment Sets	$\mathcal{O},\mathcal{P},\mathcal{A},\mathcal{S}$	$\mathcal{T}$	$\mathcal{KR}$	$\mathcal{S}$	$\mathcal{O},\mathcal{P},\mathcal{A},\mathcal{S},\mathcal{T},\mathcal{KR}$
Janus-Pro [13]	0.553	0.001	0.139	0.276	0.365
BLIP3-o [8]	0.711	0.013	0.223	0.361	0.229
BAGEL [18]	0.769	0.244	0.173	0.367	0.251
BAGEL+CoT [18]	0.793	0.020	0.206	0.390	0.209
Show-o2-1.5B [75]	0.798	0.002	0.219	0.317	0.186
Show-o2-7B [75]	0.817	0.002	0.226	0.317	0.177
OmniGen2 [71]	0.804	0.680	0.271	0.377	0.242
Stable Diffusion 1.5 [55]	0.565	0.010	0.207	0.383	0.429
Stable Diffusion XL [46]	0.688	0.029	0.237	0.332	0.296
Stable Diffusion 3.5 Large [60]	0.809	0.629	0.294	0.353	0.225
Flux.1-dev [38]	0.786	0.523	0.253	0.368	0.238
CogView4 [78]	0.786	0.641	0.246	0.353	0.205
SANA-1.5 1.6B (PAG) [73]	0.762	0.054	0.209	0.387	0.222
SANA-1.5 4.8B (PAG) [73]	0.765	0.069	0.217	0.401	0.216
Lumina-Image 2.0 [47]	0.819	0.106	0.270	0.354	0.216
HiDream-I1-Full [29]	0.829	0.707	0.317	0.347	0.186
Imagen3 [36]	0.843	0.343	0.313	0.359	0.188
Recraft V3 [66]	0.810	0.795	0.323	0.378	0.205
Kolors 2.0 [64]	0.820	0.427	0.262	0.360	0.300
Seedream 3.0 [24]	0.818	0.865	0.275	0.413	0.277
Imagen4 [17]	0.857	0.805	0.338	0.377	0.199
GPT-4o [45]	0.851	0.857	0.345	0.462	0.151

4.2.1 Semantic Alignment and Diversity

In the alignment dimension shown in Table 2, Imagen4 [17], GPT-4o [45] and Imagen3 [36] consistently outperform other models, with Imagen4 and GPT-4o showing superior alignment performance. Furthermore, as indicated in Table 3, most models achieve significantly higher alignment accuracy

when responding to natural language prompts than to tag-based or phrase-based prompts. A possible explanation is that tag and phrase prompts may introduce ambiguity, such as attribute confusion between entities or logical inconsistencies, which complicates semantic alignment. Longer prompts tend to involve greater semantic complexity and structural variation, often resulting in lower alignment scores. Notably, models incorporating T5 [52] or other large language models appear more robust in handling long prompts, exhibiting less semantic degradation.

Table 3: **The alignment and diversity evaluation results**. **NP** denotes the natural language prompt. **T&P** denotes the tag-based and phrase-based prompt. **Short**, **Medium** and **Long** represent the length of the prompts, where **Short** denote the number of words is less than 30, **Medium** denotes the number between 30 and 60, and **Long** denotes the number exceeding 60. indicate the first, second, third, fourth, and fifth performance respectively.

Method	Alignment↑					Diversity↑				
	NP	T&P	Short	Medium	Long	NP	T&P	Short	Medium	Long
Janus-Pro [13]	0.557	0.533	0.609	0.548	0.515	0.372	0.304	0.407	0.332	0.347
BLIP3-o [8]	0.719	0.671	0.754	0.712	0.674	0.237	0.161	0.283	0.192	0.198
BAGEL [18]	0.776	0.734	0.782	0.769	0.759	0.257	0.197	0.344	0.190	0.194
BAGEL+CoT [18]	0.798	0.767	0.824	0.793	0.767	0.214	0.164	0.244	0.184	0.189
Show-o2-1.5B [75]	0.805	0.760	0.800	0.799	0.793	0.191	0.142	0.226	0.162	0.157
Show-o2-7B [75]	0.825	0.778	0.825	0.819	0.807	0.182	0.130	0.216	0.152	0.151
OmniGen2 [71]	0.812	0.768	0.809	0.805	0.799	0.250	0.174	0.329	0.185	0.190
Stable Diffusion 1.5 [55]	0.570	0.541	0.616	0.558	0.537	0.434	0.381	0.481	0.389	0.403
Stable Diffusion XL [46]	0.688	0.685	0.732	0.685	0.657	0.303	0.239	0.337	0.263	0.278
Stable Diffusion 3.5 Large [60]	0.818	0.762	0.826	0.808	0.795	0.229	0.194	0.267	0.194	0.206
Flux.1-dev [38]	0.791	0.759	0.794	0.785	0.780	0.243	0.190	0.302	0.194	0.199
CogView4 [78]	0.796	0.737	0.792	0.788	0.777	0.211	0.153	0.277	0.158	0.159
SANA-1.5 1.6B(PAG) [73]	0.769	0.726	0.770	0.764	0.752	0.231	0.153	0.284	0.177	0.192
SANA-1.5 4.8B(PAG) [73]	0.773	0.729	0.782	0.767	0.749	0.223	0.154	0.264	0.181	0.191
Lumina-Image 2.0 [47]	0.825	0.788	0.829	0.819	0.812	0.224	0.149	0.282	0.171	0.180
HiDream-I1-Full [29]	0.834	0.806	0.849	0.834	0.806	0.192	0.142	0.260	0.139	0.140
Imagen3 [36]	0.849	0.809	0.859	0.841	0.832	0.189	0.173	0.246	0.146	0.153
Recraft V3 [66]	0.816	0.781	0.838	0.809	0.788	0.209	0.178	0.246	0.178	0.180
Kolors 2.0 [64]	0.824	0.798	0.847	0.814	0.807	0.308	0.230	0.359	0.261	0.259
Seedream 3.0 [24]	0.818	0.815	0.838	0.825	0.789	0.280	0.246	0.342	0.235	0.227
Imagen4 [17]	0.860	0.843	0.875	0.854	0.847	0.199	0.197	0.276	0.147	0.149
GPT-4o [45]	0.857	0.820	0.869	0.851	0.838	0.154	0.124	0.177	0.134	0.134

Diversity is informative when assessed among models with comparable levels of alignment. In real-world applications, generative models are generally expected to produce varied outputs while maintaining close adherence to the input prompts. Although Stable Diffusion 1.5 [55] and Janus-Pro [13] achieve notably high diversity scores, these results are less indicative of true generative quality, as they largely stem from the models' inconsistent preservation of semantic alignment in the generated images. And Kolors 2.0 [64] exhibits outstanding diversity without compromising its alignment performance, and is regarded as a model with excellent diversity performance.

While stylization can be viewed as a specific facet of semantic alignment, performance in this dimension does not entirely coincide with overall alignment outcomes. GPT-4o [45] retains a clear lead in stylization, while both Seedream 3.0 [24] and the SANA series methods also exhibit strong performance Notably, Stable Diffusion 1.5 [55] demonstrates impressive stylization capabilities despite its relatively poor performance in semantic alignment. This may be attributed to its data cleaning process, which likely preserved a broad range of stylistic patterns and enabled the model to generate images with distinct stylistic characteristics.

4.2.2 Text Rendering

We evaluate the performance of text-to-image models with a particular focus on text rendering, which assesses how accurately these models reproduce textual content within generated images. For clearer comparison across varying textual complexities, the images are categorized by prompt length.

As shown in Table 4, Seedream 3.0 [24] achieves the best performance across nearly all sub-dimensions of completion ratio and word accuracy count, as well as in edit distance for short and medium-length prompts. Closer inspection of the generated images reveals that the relatively higher

Table 4: **The text rendering evaluation results.** In the table, **Short**, **Medium** and **Long** represent the length of the prompts, where **Short** denote the number of words is less than 30, **Medium** denotes the length between 30 and 60, and **Long** denotes the length exceeding 60. indicate the first, second, third, fourth, and fifth performance respectively.

Method	Edit Distance (ED)↓			Completion Rate (CR)↑			Word Accuracy (WAC)↑		
	Short	Medium	Long	Short	Medium	Long	Short	Medium	Long
Janus-Pro[13]	33.041	59.695	295.020	0.000	0.000	0.000	0.001	0.001	0.000
BLIP3-o[8]	31.627	59.584	258.510	0.005	0.000	0.000	0.014	0.022	0.012
BAGEL[18]	25.650	44.453	242.415	0.005	0.079	0.000	0.128	0.288	0.148
BAGEL+CoT[18]	30.550	57.397	255.220	0.000	0.005	0.000	0.024	0.033	0.011
Show-o2-1.5B [75]	32.205	59.232	296.580	0.000	0.000	0.000	0.002	0.004	0.001
Show-o2-7B [75]	31.709	59.179	294.720	0.000	0.000	0.000	0.004	0.006	0.001
OmniGen2 [71]	17.000	36.145	181.525	0.255	0.145	0.000	0.543	0.614	0.534
Stable Diffusion 1.5 [55]	36.227	60.480	290.245	0.000	0.011	0.000	0.004	0.020	0.004
Stable Diffusion XL [46]	35.045	61.824	290.470	0.000	0.008	0.000	0.020	0.056	0.029
Stable Diffusion 3.5 Large [60]	19.459	38.711	248.280	0.432	0.255	0.005	0.749	0.740	0.512
Flux.1-dev [38]	26.855	44.189	227.565	0.223	0.161	0.000	0.387	0.577	0.430
CogView4 [78]	17.173	29.437	193.420	0.200	0.150	0.010	0.437	0.593	0.517
SANA-1.5 1.6B (PAG) [73]	31.573	61.566	288.225	0.059	0.003	0.000	0.143	0.090	0.031
SANA-1.5 4.8B (PAG) [73]	25.027	55.634	268.025	0.086	0.000	0.000	0.228	0.079	0.030
Lumina-Image 2.0 [47]	26.259	54.547	270.530	0.059	0.013	0.000	0.199	0.199	0.083
HiDream-I1-Full [29]	14.464	31.026	177.765	0.195	0.166	0.005	0.435	0.602	0.576
Imagen3 [36]	34.005	83.171	239.565	0.068	0.071	0.000	0.279	0.447	0.371
Recraft V3 [66]	17.050	26.945	74.181	0.050	0.061	0.006	0.267	0.371	0.430
Kolors 2.0 [64]	18.236	37.930	235.953	0.125	0.118	0.000	0.374	0.483	0.257
Seedream 3.0 [24]	7.204	20.596	169.307	0.699	0.451	0.109	0.893	0.822	0.688
Imagen4 [17]	18.273	42.918	121.625	0.181	0.184	0.060	0.364	0.575	0.745
GPT-4o [45]	18.255	37.850	85.430	0.150	0.171	0.070	0.323	0.495	0.673

edit distances observed for Seedream 3.0 are mainly associated with the difficulty of rendering long text inputs, particularly in structured formats such as articles and PowerPoint slides. Images generated with long text inputs may contain more frequent textual deviations. While Seedream 3.0 performs well on certain long prompts, the overall complexity of long-text rendering contributes to an increase in edit distance.

As shown in Figure 6, although GPT-4o [45] demonstrates strong visual accuracy, it shows no particular advantage in quantitative evaluation. This is mainly due to our strict evaluation criterion: any mismatch in capitalization (e.g., uppercase vs. lowercase) is counted as one unit of edit distance. This rule has a direct impact on GPT-4o's overall rendering scores.

Compared to Imagen3 [36], Imagen4 [17] shows a clear improvement in text rendering, with nearly double the performance on metrics such as ED and CR. Nonetheless, its overall visual appeal and the clarity of rendered text for long prompts still leave room for enhancement.

Notably, Recraft V3 [66] consistently achieves excellent edit distance performance across all prompt lengths, with its values for long prompts showing a large gap from the second-best result. However, its performance in completion ratio and word accuracy count is relatively less impressive. This discrepancy may be attributed to its layout-first strategy [65], in which a layout is generated prior to text insertion. This approach effectively reduces severe errors and prevents chaotic outputs by decomposing the original text rendering task into several relatively simpler subtasks, thereby significantly enhancing the reliability of text rendering in the final images.

4.2.3 Knowledge and Reasoning

In the knowledge and reasoning dimension, shown in Table 5 and in Figure 6, GPT-4o [45] demonstrates substantially stronger capabilities than other models. It consistently outperforms its counterparts in both knowledge retention and reasoning ability across nearly all subject categories evaluated. In general, closed-source models outperform open-source models in knowledge and reasoning capabilities. Notably, no single model has shown a particularly outstanding performance in specific

subjects, indicating that the reasoning abilities of current models are largely derived from a balanced and conventional training dataset.

Table 5: **The Knowledge and Reasoning evaluation results.** indicate the first, second, third, fourth, and fifth performance respectively.

Method	Geography	Computer Science	Biology	Mathmatics	Physics	Chemistry	Common Sense
Janus-Pro [13]	0.153	0.134	0.123	0.130	0.144	0.131	0.189
BLIP3-o [8]	0.210	0.225	0.219	0.223	0.219	0.232	0.275
BAGEL [18]	0.191	0.152	0.183	0.149	0.168	0.169	0.248
BAGEL+CoT [18]	0.206	0.184	0.206	0.200	0.212	0.208	0.273
Show-o2-1.5B [75]	0.218	0.197	0.235	0.201	0.222	0.206	0.295
Show-o2-7B [75]	0.222	0.206	0.244	0.217	0.227	0.212	0.297
OmniGen2 [71]	0.266	0.270	0.271	0.284	0.257	0.278	0.314
Stable Diffusion 1.5 [55]	0.217	0.203	0.211	0.212	0.207	0.185	0.251
Stable Diffusion XL [46]	0.246	0.216	0.244	0.235	0.234	0.239	0.289
Stable Diffusion 3.5 Large [60]	0.291	0.299	0.283	0.306	0.292	0.292	0.319
Flux.1-dev [38]	0.239	0.257	0.247	0.265	0.247	0.253	0.298
CogView4 [78]	0.223	0.251	0.237	0.279	0.239	0.252	0.296
SANA-1.5 1.6B(PAG) [73]	0.207	0.203	0.222	0.218	0.211	0.193	0.280
SANA-1.5 4.8B(PAG) [73]	0.214	0.206	0.224	0.224	0.214	0.203	0.282
Lumina-Image 2.0 [47]	0.208	0.206	0.225	0.222	0.215	0.197	0.278
HiDream-I1-Full [29]	0.324	0.324	0.305	0.312	0.318	0.305	0.347
Imagen3 [36]	0.304	0.319	0.298	0.303	0.320	0.315	0.338
Recraft V3 [66]	0.323	0.337	0.303	0.320	0.319	0.328	0.344
Kolors 2.0 [64]	0.255	0.252	0.256	0.263	0.258	0.277	0.314
Seedream 3.0 [24]	0.246	0.295	0.313	0.253	0.297	0.270	0.277
Imagen4 [17]	0.334	0.346	0.314	0.343	0.342	0.346	0.351
GPT-4o [45]	0.351	0.348	0.323	0.334	0.350	0.355	0.364

The models' reasoning abilities can be categorized into five tiers as follows: the first tier includes GPT-4o [45], Imagen4 [17], the second tier consists of Recraft V3 [66], HiDream-I1-Full [29], and Imagen3 [36], the third tier includes only Stable Diffusion 3.5 Large [60], and the fourth tier includes Seedream 3.0 [24], Lumina-Image 2.0 [47], and Kolors 2 [64], other models form the last tier. The following figure visualizes the reasoning scores, which correspond closely with the aforementioned ranking.

5 Conclusion

We introduce a comprehensive text-to-image benchmark, namely **OneIG-Bench**, which establishes a systematic framework for omni-dimensional nuanced evaluation through categorization of generation themes. Specifically, we have meticulously designed general scenarios including human figures and conventional objects, text rendering scenarios, and anime/style scenarios, and have crafted evaluation metrics for each scenario to comprehensively measure text-to-image performance. By decomposing evaluation into these discrete dimensions, the benchmark facilitates in-depth comparative analysis of models' strengths and limitations. This approach not only provides researchers with a rigorous evaluation framework but also serves as a guiding tool for identifying technical bottlenecks and prioritizing methodological innovations in the field.

Limitation: While this study presents a novel and systematic benchmark, several limitations should be acknowledged. (1) Knowledge and reasoning represents a relatively novel task in the image generation domain, and most existing models currently lack robust reasoning capabilities. While we have confirmed that our metric rankings align closely with human evaluations, there may exist more rational and effective evaluation approaches yet to be explored. (2) Moreover, aesthetic models tend to exhibit unexpected biases, while body quality assessment models often lack sufficient discriminative power and generalizability. We will further investigate both dimensions to develop a more robust and precise evaluation method.

Acknowledgements

We are particularly grateful to Bizhu Huang, Shuli Gao, Kang An, and Wen Sun for their invaluable support throughout the course of this research.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[5] Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20041–20053, 2023.

[6] black-forest labs. The official api of flux-1.dev. `https://api.us1.bfl.ai/scalar#tag/tasks/POST/v1/flux-dev`, 2024.

[7] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.

[8] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025.

[9] 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. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[10] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[11] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[12] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020.

[13] 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. *arXiv preprint arXiv:2501.17811*, 2025.

[14] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[15] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021.

[16] Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *ICLR*, 2024.

[17] Google deepmind Imagen4 team. Imagen4. `https://storage.googleapis.com/deepmind-media/Model-Cards/Imagen-4-Model-Card.pdf`, 2025.

[18] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

[19] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[20] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

[21] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data, 2023.

[22] Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. Commonsense-t2i challenge: Can text-to-image generation models understand commonsense? *arXiv preprint arXiv:2406.07546*, 2024.

[23] Rohit Gandikota and David Bau. Distilling diversity and control in diffusion models. *arXiv preprint arXiv:2503.10637*, 2025.

[24] Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025.

[25] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[27] Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, and Chongyi Li. Evalmuse-40k: A reliable and fine-grained benchmark with comprehensive human annotations for text-to-image generation model evaluation, 2024.

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

[29] HiDream-ai. Hidream-i1. `https://github.com/HiDream-ai/HiDream-I1`, 2025.

[30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[31] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

[32] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023.

[33] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[34] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

[35] Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Xiyang Dai, Dongdong Chen, Chong Luo, et al. Llm2clip: Powerful language model unlock richer visual representation. *arXiv preprint arXiv:2411.04997*, 2024.

[36] Imagen-Team-Google. Imagen 3, 2024.

[37] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.

[38] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

[39] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.

[40] 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, 2024.

[41] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[42] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pages 23–40. Springer, 2024.

[43] Yuwei Niu, Munan Ning, Mengren Zheng, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

[44] OpenAI. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024.

[45] OpenAI. Introducing 4o image generation. https://openai.com/index/introducing-4o-image-generation/, 2025.

[46] 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. *arXiv preprint arXiv:2307.01952*, 2023.

[47] Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Xinyue Li, Dongyang Liu, Xiangyang Zhu, Will Beddow, Erwann Millon, Wenhai Wang Victor Perez, Yu Qiao, Bo Zhang, Xiaohong Liu, Hongsheng Li, Chang Xu, and Peng Gao. Lumina-image 2.0: A unified and efficient image generative framework, 2025.

[48] Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, et al. Phybench: Holistic evaluation of physical perception and reasoning in large language models. *arXiv preprint arXiv:2504.16074*, 2025.

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[51] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[52] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[53] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[54] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[55] 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 (CVPR)*, pages 10684–10695, June 2022.

[56] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[57] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In *International conference on machine learning*, pages 30105–30118. PMLR, 2023.

[58] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024.

[59] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[60] Stability-AI. stable-diffusion-3.5-large. `https://github.com/Stability-AI/sd3.5`, 2024.

[61] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

[62] Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, Mengping Yang, Cheng Zhang, and Hao Li. Evalalign: Evaluating text-to-image models through precision alignment of multimodal large models with supervised fine-tuning to human annotations. *arXiv e-prints*, pages arXiv–2406, 2024.

[63] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[64] Kuaishou Kolors team. Kolors2.0. `https://app.klingai.com/cn/`, 2025.

[65] Recraft team. How to create sota image generation with text recrafts ml team insights. `https://www.recraft.ai/blog/how-to-create-sota-image-generation-with-text-recrafts-ml-team-insights`, 2024.

[66] Recraft team. Recraft v3. `https://www.recraft.ai/blog/recraft-introduces-a-revolutionary-ai-model-that-thinks-in-design-language?utm_source=ai-bot.cn`, 2024.

[67] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[68] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[69] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[70] Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Pinelopi Papalampidi, Ira Ktena, Chris Knutsen, et al. Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings. *arXiv preprint arXiv:2404.16820*, 2024.

[71] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

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

[73] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer, 2025.

[74] 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. *arXiv preprint arXiv:2408.12528*, 2024.

[75] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

[76] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*, 2024.

[77] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

[78] THUKEG Z.ai. Cogview4. `https://github.com/THUDM/CogView4`, 2025.

[79] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[80] Daoan Zhang, Che Jiang, Ruoshi Xu, Biaoxiang Chen, Zijian Jin, Yutian Lu, Jianguo Zhang, Liang Yong, Jiebo Luo, and Shengda Luo. Worldgenbench: A world-knowledge-integrated benchmark for reasoning-driven text-to-image generation, 2025.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim our contributions in the abstract and introduction 1. We have three main contributions: presenting **OneIG-Bench** with different generation dimensions, introducing a systematic quantitative evaluation, and state-of-the-art methods are evaluated based on our proposed benchmark to facilitate the development of the text-to-image research.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss the current limitation 5 in our paper.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We state our assumptions and define the equations in section 4.1.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all the information.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code, datasets, and models used in this paper are all publicly accessible.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specify all evaluation settings and data.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report the average score of evaluation results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide our computer resources in Results and Analysis section 4.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential impacts in section 5 5

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The assets are properly credited and license and terms of use is explicitly mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets are documented.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We have described the usage of the LLMs in the paper.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Technical Appendices and Supplementary Material

A.1 Word Count Distribution Statistics

The word count distribution of our OneIG-Bench prompts, as shown in Table 6 and Figure 3, follows a **Short** : **Middle** : **Long** ratio of approximately 1:2:1. The choice of 30 and 60 words as the thresholds for distinguishing short, middle, and long prompts is based on the following reasoning: According to common rules for understanding token lengths, 1 word is approximately equal to 4/3 tokens, and 1-2 sentences are roughly equivalent to 30 tokens. This means that a simple 1-2 sentence prompt has a length of about 20-25 words. To ensure diversity in sentence structure and accuracy in stylization or portraiture in the image, we set the boundaries for short and middle prompts at 30 words. Furthermore, since some text encoders, such as CLIP [50], SigLIP [79], support a maximum of 77 tokens, a prompt of up to 60 words can generally be processed directly by these encoders.

Table 6: **Word count distribution of OneIG-Bench prompts in different categories**. In the table, **Avg** represents the average word count of prompts in different categories (including total and total w/o Knowledge & Reasoning). **Short**, **Medium** and **Long** represent the length of the prompts, where **Short** denote the number of words is less than 30, **Medium** denotes the number between 30 and 60, and **Long** denotes the number exceeding 60. "K & R" is the abbreviation for "Knowledge & Reasoning".

Category	Avg	Short	Middle	Long
Portrait	56.4	0.184	0.443	0.373
General Object	46.5	0.330	0.422	0.248
Anime & Stylization	50.6	0.212	0.522	0.265
Text Rendering	51.2	0.275	0.475	0.250
Knowledge & Reasoning	20.5	0.960	0.018	0.022
Total Distribution	45.2	0.389	0.377	0.234
Total Distribution w/o K & R	51.3	0.246	0.467	0.287

The Portrait category, however, shows a slight deviation from this 1:2:1 distribution in Figure 4 due to the explicit requirement for portraits in the prompts, ensuring that the generated characters do not include stylized figures like those found in anime. As a result, the average word count for prompts in this category is higher than in other categories. On the other hand, the Knowledge & Reasoning category, which focuses on reasoning tasks, does not revise the prompts to conform to the word count ratio, leading to a noticeably lower average word count compared to other categories. In general, excluding the Knowledge & Reasoning category, OneIG-Bench prompts' word count results align closely with the 1:2:1 ratio.

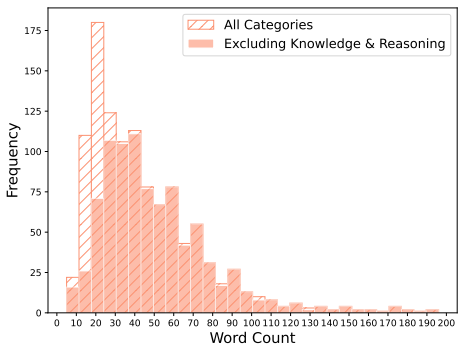


Figure 3: **Word Count of the Overall Prompts of OneIG-Bench**. The word count distribution of OneIG-Bench's prompts ranges from 0 to 200.

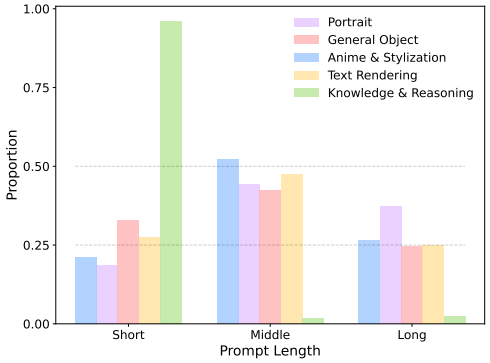


Figure 4: **The distribution of prompt word counts across Short, Middle, Long categories**.

A.2 Implementation

Our experiments on image generation with unified multimodal methods and open-source methods are configured according to Table 7. For all methods, the CFG and step parameters follow the methods' default settings. To ensure consistency and ensure the quality of image generation, we increased the default steps for Stable Diffusion 3.5 Large [60] from 40 to 50. The number of inference steps for Flux.1-dev [38] is set to be consistent with that used in the official API [6]. With the exception of Stable Diffusion 1.5 [55], which cannot generate images with a resolution of 1024×1024, all other methods are configured to generate images at a resolution of 1024×1024.

Table 7: **Configurations for unified multimodal and open-source methods**. **Size** represents the parameter size of the corresponding method. **CFG** represents the **guidance scale** of the corresponding method. **Resolution** represents the resolution of the image generated by the corresponding method. **Step** represents the number of inference steps during the image generation process.

Method	Size	CFG	Resolution	Step
Janus-Pro [13]	7B	5.0	384×384	-
BLIP3-o [8]	8B	3.0	1024×1024	-
BAGEL [18]	7B	4.0	1024×1024	50
BAGEL+CoT [18]	7B	4.0	1024×1024	50
Show-o2-1.5B [75]	1.5B	5.0	432×432	50
Show-o2-7B [75]	7B	5.0	432×432	50
OmniGen2 [71]	3B+4B	4.0	1024×1024	50
Stable Diffusion 1.5 [55]	0.9B	7.5	512×512	50
Stable Diffusion XL [46]	2.6B	5.0	1024×1024	50
Stable Diffusion 3.5 Large [60]	8.1B	4.5	1024×1024	50
Flux.1-dev [38]	12B	3.5	1024×1024	28
CogView4 [78]	6B	3.5	1024×1024	50
SANA-1.5 1.6B (PAG) [73]	1.6B	5.0	1024×1024	20
SANA-1.5 4.8B (PAG) [73]	4.8B	5.0	1024×1024	20
Lumina-Image 2.0 [47]	2.6B	4.0	1024×1024	50
HiDream-I1-Full [29]	17B	5.0	1024×1024	50

For closed-source methods, we present the corresponding release or update dates of the methods in Table 8 to facilitate alignment with subsequent experimental results.

Table 8: **Release/Update date of closed-source methods**. Release/Update date represents the version of the corresponding method when generating images.

Method	Imagen3 [36]	Recraft V3 [66]	Kolors 2.0 [64]	Seedream 3.0 [24]	Imagen4 [17]	GPT-4o [45]
Release/Update Date	2025-01-23	2024-10-30	2025-04-15	2025-04-15	2025-05-20	2025-04-29

A.3 The Details on Prompts Rewriting

Algorithm 1: Initial Prompts Rewritten by GPT-4o

Input: n: the length of the initial prompts list,
$\mathbb{P}_{\text{init}}$: the initial prompts $[p_1, p_2, \ldots, p_n]$.
Output: $\mathbb{P}_{\text{rewritten}}$: the rewritten prompts $[p'_1, p'_2, \ldots, p'_n]$
$\mathbb{P}_{\text{sorted}} \leftarrow$ sorted_by_word_count $(\mathbb{P}_{\text{init}})$
$R \leftarrow 100 * \text{sorted}(\text{beta.rvs}(2.37, 2.86, n))$
for $i \leftarrow 1$ ***to*** n **do**
  initial_prompt $\leftarrow \mathbb{P}_{\text{sorted}}[i]$
  target_word_count $\leftarrow R[i]$
  rewritten_prompt $\leftarrow$ GPT-4o_API(Prompt Template, initial_prompt, target_word_count)
  $\mathbb{P}_{\text{rewritten}}[i] \leftarrow$ rewritten_prompt
return $\mathbb{P}_{\text{rewritten}}$

As shown in Algorithm 1, the initial prompts are first sorted based on their word count. Then, using a Beta distribution with parameters (2.37, 2.86), which roughly follows the ratio 0-0.3:0.3-0.6:0.6-1

$\approx$ 1:2:1, a list of desired prompt lengths is generated and subsequently sorted. The sorted prompts are then matched with the corresponding desired lengths, and the GPT-4o API is called to rewrite each prompt according to its specified length, resulting in the rewritten prompts. Without loss of generality, lengths in the range of 60-100 can be mapped to a range of greater, thereby generating longer prompts. In this process, the corresponding prompt for rewriting is as follows:

Prompt of Prompts Rewriting

You are a precise rewriting assistant.
Task Description:

- Rewrite the <initial_prompt > according to the <target_word_count> of the prompt.
- For <initial_prompt > longer than the <target_word_count>
  - Shorten the prompt by carefully removing specific but non-essential details.
  - Do not simply delete words or generalize the description.
- For <initial_prompt > shorter than the <target_word_count>
  - Expand the prompt by adding specific, meaningful, and vivid details that enhance the scene.
  - Do not introduce abstract or generalized commentary.
- Ensure the rewritten prompt
  - The prompt should be coherent, natural, fluent, logically structured.
  - Please maintain the initial tone and intent as much as possible.

Important:

- Only output the final rewritten prompt without any additional words.

A.4 The Details and Analysis on Stylization

Table 9: **The styles in OneIG-Bench corresponding to specific categories**.

Category	Style
Traditional	abstract expressionism, art nouveau, Baroque, Chinese ink painting, cubism, fauvism, impressionism, line art, minimalism, pointillism, pop art, Rococo, Ukiyo-e
Media	clay, crayon, graffiti, LEGO, pencil sketch, stone sculpture, watercolor
Anime	Celluloid, Chibi, comic, Cyberpunk, Ghibli, Impasto, Pixar, pixel art, 3d rendering,

The Anime & Stylization category encompasses a variety of styles, which are systematically grouped into three subcategories in Table 9: **Traditional**, **Media**, and **Anime**. The **Traditional** category primarily includes styles rooted in classical and historical art movements from around the world. The **Media** category includes styles defined by specific artistic media and material-based techniques. The **Anime** category represents a collection of stylized and detailed visual aesthetics commonly associated with animation and pop culture. And Table 10 presents the scores of different methods across various style categories. The calculation process is as follows: for each method, the average style score is first calculated based on the images generated for each prompt within each style. Then, the score for each style category is obtained by averaging the scores of the individual styles within that category. It is clear that GPT-4o [45] demonstrates exceptional style-following ability across most categories, significantly outperforming other methods.

A.5 Overall quantitive results of OneIG-Bench-ZH

To facilitate comparison, we present the quantitative results of OneIG-Bench-ZH in Table 11 and showcase qualitative examples on OneIG-Bench-ZH and OneIG-Bench from several top-performing methods in Figure 5.

Table 10: **The style scores on different categories of styles**. indicate the first, second, third, fourth, and fifth performance, respectively.

Method	Traditional	Media	Anime
Janus-Pro [13]	0.224	0.212	0.412
BLIP3-o [8]	0.381	0.287	0.390
BAGEL [18]	0.363	0.297	0.440
BAGEL+CoT [18]	0.372	0.329	0.499
Show-o2-1.5B [75]	0.269	0.288	0.410
Show-o2-7B [75]	0.283	0.295	0.380
OmniGen2 [71]	0.360	0.314	0.458
Stable Diffusion 1.5 [55]	0.483	0.298	0.349
Stable Diffusion XL [46]	0.316	0.307	0.339
Stable Diffusion 3.5 Large [60]	0.356	0.315	0.335
Flux.1-dev [38]	0.367	0.298	0.391
CogView4 [78]	0.376	0.294	0.369
SANA-1.5 1.6B(PAG) [73]	0.438	0.331	0.370
SANA-1.5 4.8B(PAG) [73]	0.443	0.340	0.379
Lumina-Image 2.0 [47]	0.351	0.325	0.360
HiDream-I1-Full [29]	0.331	0.295	0.368
Imagen3 [36]	0.378	0.309	0.371
Recraft V3 [66]	0.418	0.347	0.332
Kolors 2.0 [64]	0.370	0.336	0.360
Seedream 3.0 [24]	0.383	0.365	0.524
Imagen4 [17]	0.336	0.365	0.452
GPT-4o [45]	0.532	0.404	0.411

We evaluate a relatively smaller set of well-known image generation models on OneIG-Bench-ZH, including unified multimodal models (Janus-Pro [13], BLIP3-o [8], BAGEL [18]), open-source models (CogView4 [78], Lumina-Image 2.0 [47], and HiDream-I1-Full [29]) using A800 GPUs, as well as closed-source models (Kolors 2.0 [64], Seedream 3.0 [24], and GPT-4o [45]). In OneIG-Bench-ZH, *Multilingualism* part consists of 100 culture-related prompts and 100 portrait-related prompts. To provide a clear overall comparison, the evaluation results across multiple dimensions are summarized in Table 11. On OneIG-Bench-ZH, GPT-4o [45] demonstrates outstanding performance, ranking first across most evaluation dimensions in Table 11 and Figure 5. In contrast, Seedream 3.0 [24] excels particularly in Chinese text rendering, significantly outperforming GPT-4o. However, most models show limited capability in generating Chinese text, with many nearly incapable of producing legible Chinese characters. It is also worth noting that for most models, performance on alignment, reasoning, and style dimensions is slightly weaker on OneIG-Bench-ZH than on OneIG-Bench, indirectly reflecting that their ability to understand and generate Chinese semantics still requires further improvement.

Table 11: **Overall quantitative comparison of different methods on OneIG-Bench-ZH**. The table showcases the results of five core metrics for various methods. indicate the first, second, third, fourth, and fifth performance, respectively.

Method	Alignment ↑	Text ↑	Reasoning ↑	Style ↑	Diversity ↑
Assessment Sets	$\mathcal{O}_{zh}, \mathcal{P}_{zh}, \mathcal{A}_{zh}, \mathcal{S}_{zh}, \mathcal{L}_{zh}$	$\mathcal{T}_{zh}$	$\mathcal{KR}_{zh}$	$\mathcal{S}_{zh}$	$\mathcal{O}_{zh}, \mathcal{P}_{zh}, \mathcal{A}_{zh}, \mathcal{S}_{zh}, \mathcal{T}_{zh}, \mathcal{KR}_{zh}, \mathcal{L}_{zh}$
Janus-Pro [13]	0.324	0.148	0.104	0.264	0.358
BLIP3-o [8]	0.608	0.092	0.213	0.369	0.233
BAGEL [18]	0.672	0.365	0.186	0.357	0.268
BAGEL+CoT [18]	0.719	0.127	0.219	0.385	0.197
Cogview4 [78]	0.700	0.193	0.236	0.348	0.214
Lumina-Image 2.0 [47]	0.731	0.136	0.221	0.343	0.240
HiDream-I1-Full [29]	0.620	0.205	0.256	0.304	0.300
Kolors 2.0 [64]	0.738	0.502	0.226	0.331	0.333
Seedream 3.0 [24]	0.793	0.928	0.281	0.397	0.243
GPT-4o [45]	0.812	0.650	0.300	0.449	0.159

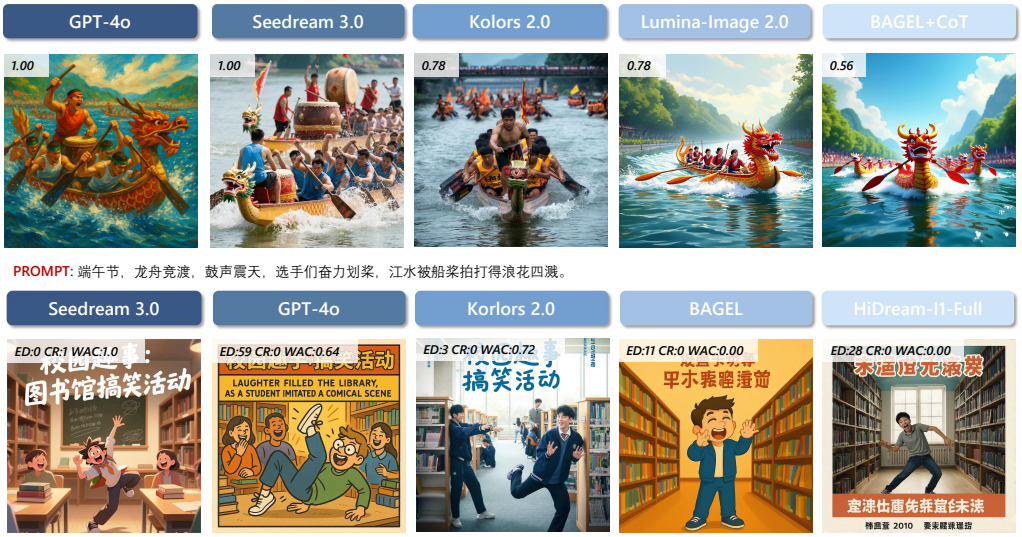


Figure 5: **An illustration of the generation results and the corresponding scores on OneIG-Bench-ZH.** The first row shows the alignment results and the second row shows the the text rendering results. And the evaluation scores are displayed in the upper left corner of the image.

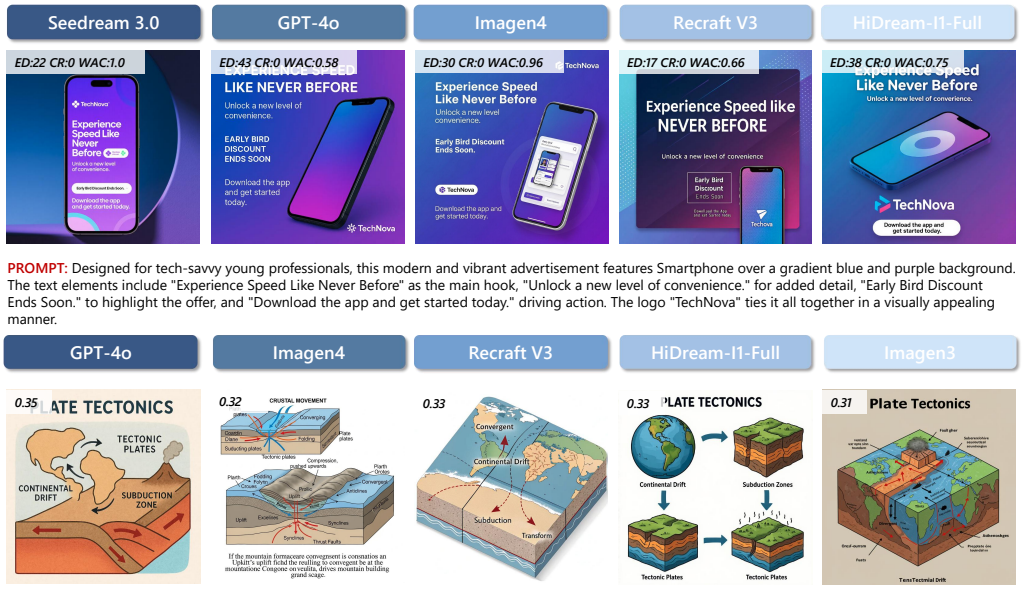


Figure 6: **An illustration of the generation results and the corresponding scores.** The first row shows the text rendering results and the second row shows the the knowledge and reasoning results. And the evaluation scores are displayed in the upper left corner of the image.

A.6 Benchmark Reliability Analysis

A.6.1 User Study

User studies have validated the high alignment between OneIG-Bench's evaluation model results and human annotations. To measure human agreement with the evaluation results of OneIG-Bench, we randomly select 430 prompts from OneIG-Bench, distributed as follows:

- Alignment: 150 prompts (50 from Anime & Stylization, 50 from Portrait, 50 from General Object)

- Text: 60 prompts
- Reasoning: 60 prompts
- Style: 51 prompts
- Diversity: 100 prompts (20 from each category)

For each prompt, generation results are randomly sampled from 4 distinct models. Subsequently, over 30 human annotators are arranged to rank these outputs. The resulting human rankings are then compared with the rankings generated by the evaluation model.

Due to possible ties (e.g., all 4 models performing the same in certain prompts), traditional correlation metrics like Spearman or Kendall are not applicable. To address this, we adopt a custom metric called Ranking Consistency (RC), which measures the proportion of pairwise comparisons where the model ranking agrees with the human ranking. If two items are ordered the same way in both human and model rankings, they are considered a consistent pair. RC is the ratio of consistent pairs to all possible pairs.

Table 12 presents the average RC values across different evaluation dimensions, reflecting the accuracy and reliability of our ranking mechanism.

Table 12: **Ranking consistency between OneIG-Bench evaluations and human annotations**.

	Alignment	**Text**	**Reasoning**	**Style**	**Diversity**
Avg. of RC	0.8133	0.9139	0.9167	0.7931	0.8857

Despite the strong reliability observed in user studies, the evaluation remains susceptible to hallucinations and related artifacts arising from the use of LLMs and VLMs.

For the semantic alignment evaluation, to reduce the likelihood of hallucinated responses, we design the prompts with clear structure and straightforward wording, while imposing strict constraints on the model's responses—ensuring they can directly address "Yes" or "No" questions. Since the format of questions in this section is relatively simple, we find that VLMs perform significantly better compared to their performance on more complex image understanding benchmarks.

Also, we observe some hallucinations in tasks during text extraction. Specifically, in our evaluation of Qwen2.5-VL-7B's extracted text sequences, we find a hallucination rate of approximately 6.49%. We further analyze these cases and identified some common patterns in hallucinated outputs. As a result, we've implemented a filtering and cleaning process within our codebase to remove typical hallucinated terms. This process effectively reduces hallucinated text to approximately 1%, ensuring fairness and reliability in our subsequent evaluations. Here we provide representative text-based cases in Table 13.

Table 13: **VLM failure cases**.

Case	**Human results**	**VLM results**
Case 1	VOICE OF JUSTICE THE LEGEND OF A FEMALE LAWYER	VOICE OF JUSTICE THE LEGEND OF A FEMALE LAWYER **No text recognized**
Case 2	NOW SHOWING CLASSIC FILM FESTIVAL ADULTS $5 – POPCORN INCLUDED THEATER 1EUBRR PLT 20TM	NOW SHOWING **addCriterion** CLASSIC FILM FESTIVAL ADULTS $5 – POPCORN INCLUDED THEATER 1EUBRR PLT 20TM

A.6.2 Comparison with Existing Benchmarks

OneIG-Bench shows good overall alignment with existing commonly used benchmarks in the alignment dimension. However, due to differences in prompt content, there are also some variations in relative rankings. We select the two most widely adopted alignment benchmarks for T2I models, as other dimensions currently lack sufficient model coverage for meaningful comparison. As shown in the Table 14, OneIG-Bench results are largely consistent with both GenEval [25] and DPG Bench[31], with a notably higher agreement with DPG Bench. This may be attributed to differences in prompt

length distribution: GenEval primarily focuses on short prompts, while DPG Bench emphasizes medium to long prompts, which aligns more closely with the distribution in OneIG-Bench.

For alignment metrics, our benchmark includes a balanced mix of short, medium, and long prompts, reflecting a more realistic task setting. Furthermore, OneIG-Bench prompts are more diverse in form—ranging from tags and phrases to natural language instructions—enhancing generalizability. In addition, we are the first to introduce the Anime & Stylization category, which has been largely overlooked in existing benchmarks.

Overall, the distribution of prompt lengths, the diversity in prompt forms, and the inclusion of different content domains collectively influence the generation quality of T2I models and, consequently, their alignment rankings. As generative tasks grow increasingly complex, demanding performance across diverse areas like alignment, text rendering, and reasoning, AIGC researchers need a benchmark that offers broad coverage and more closely mirrors real-world application scenarios. Our benchmark perfectly meets this demand.

Table 14: **Model performance across benchmarks and inter-benchmark correlations**.

Model	OneIG	#Rank	GenEval	#Rank	DPG Bench	#Rank
Stable Diffusion 1.5	0.565	8	0.430	8	63.18	8
Stable Diffusion XL	0.688	7	0.550	7	74.65	7
Flux. 1-dev	0.786	3	0.660	6	83.79	6
CogView4	0.786	3	0.730	4	85.13	3
SANA-1.5 1.6B (PAG)	0.762	6	0.82	2	84.50	5
SANA-1.5 4.8B (PAG)	0.765	5	0.81	3	84.70	4
Lumina-Image 2.0	0.819	2	0.730	4	87.20	1
HiDream-I1-Full	0.829	1	0.830	1	85.89	2
PLCC (OneIG & X)	1	-	0.6037	-	0.8503	-
SPCC (OneIG & X)	1	-	0.5964	-	0.8743	-

A.7 Visualization Results

We first give the best normalized polar visualization of SOTA methods on OneIG-Bench in Table 15. It can be observed that the closed-source methods outperform the other two categories of methods overall. And we selected some methods based on their overall performance across non-style metrics and showcased representative examples for each. For the style dimension, we further selected images from three finer-grained subcategories to illustrate the results. **In all visualizations, the methods are broadly selected based on overall performance. In the following figures, each image tile is labeled in the top-left corner with the score achieved by the corresponding method under the current evaluation metric.**

Figure 7 shows that Imagen4 [17] and GPT-4o [45] demonstrate strong capabilities in semantic alignment. Notably, in multi-person generation tasks, many methods struggle to accurately fulfill requirements at the individual level and often exhibit confusion in assigning attributes to the correct subjects. In addition, some methods tend to overlook fine-grained details while focusing on the primary generation task, which significantly hinders their ability to achieve high alignment scores.

Text rendering is an important task for current generative methods. As shown in Figure 8, Seedream 3.0 [24] demonstrates high accuracy and aesthetic quality. Some methods, such as GPT-4o [45], demonstrate limitations in adhering to case sensitivity (distinguishing between uppercase and lowercase letters), which compromises the textual accuracy of the rendered content. While the generated images may appear visually impressive, these subtle errors can lead to a noticeable divergence between subjective visual quality and objective evaluations based on metrics such as ED and WAC. Imagen4 [17] demonstrates good accuracy in text generation, but the overall visual quality of the images is relatively poor. Recraft V3 [66], although rarely producing major errors in text generation, tends to make mistakes at the word level and suffers from inconsistencies in typography and layout coherence. Overall, HiDream-I1-Full [29] performs reasonably well in text rendering—it may not outperform Seedream 3.0, Recraft V3, or GPT-4o, but it is able to fulfill the basic prompt requirements.

From a reasoning perspective, only GPT-4o [45] demonstrates both logical coherence and textual accuracy in Figure 9. Although Imagen4 [17] and Recraft V3 [66] fall short of GPT-4o in terms of clarity and correctness, they produce text that is generally readable. HiDream-I1-Full [29] provides limited textual and visual content but manages to convey a certain degree of knowledge and reasoning, albeit with insufficient accuracy. In contrast, Imagen3 [36] tends to generate overly redundant outputs, with excessive textual and graphical elements that obscure the intended message, and often includes incorrect information. Therefore, Knowledge and Reasoning remains a critical area that warrants further investigation and refinement for generative methods.

The visualization of diversity results is presented in Figure 10, where the ranking of diversity scores aligns well with visual inspection, supporting the validity of our proposed diversity metrics, although the diversity observed in some methods may partly stem from insufficient alignment capabilities.

Figures 11, 12, and 13 show that GPT-4o [45] performs well across most styles, though it struggles with specific ones especially some anime styles. Stable Diffusion 1.5 [55], despite its lower visual quality, effectively captures traditional style features. Although Imagen4 [17] does not achieve a high overall score in style, it performs notably well in media and anime styles. It is worth noting that unified multimodal methods, such as BAGEL+CoT [18], OmniGen2 [71], BAGEL [18] and Janus-Pro [13], exhibit particularly strong performance in the anime style. Overall, Seedream 3.0[24] and SANA-1.5 4.8B (PAG)[73] also demonstrate strong stylistic consistency, ranking just behind GPT-4o.

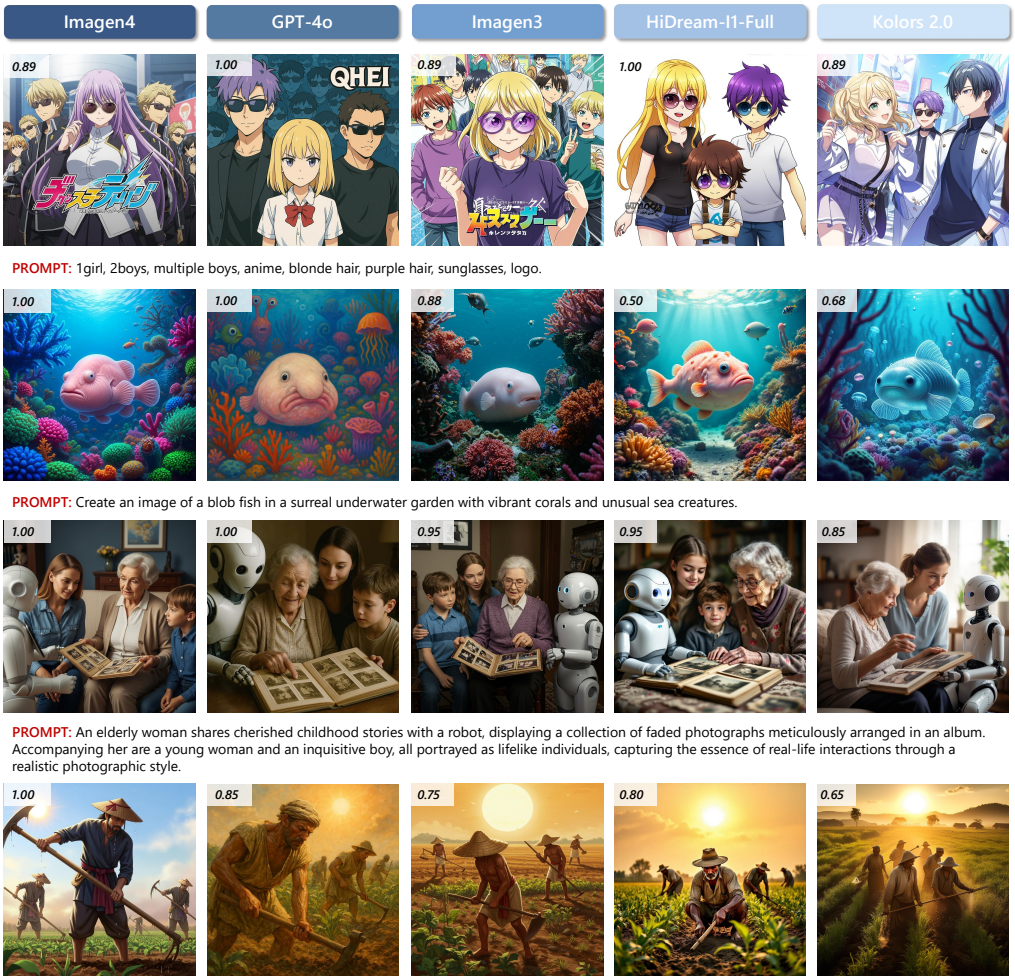


Figure 7: **Visualization results of SOTA methods. Alignment** of Imagen4 [17], GPT-4o [45], Imagen3 [36], HiDream-I1-Full [29] and Kolors 2.0 [64] are evaluated respectively. **Row 1** corresponds to **tag/phrase prompt**: The variation in the scores of the visual samples are mainly influenced by: "2 boys" and "logo". **Row 2** corresponds to **short prompt**: The variation in the scores of the visual samples are mainly influenced by: "blob fish", "surreal underwater" and "unusual sea creatures". **Row 3** corresponds to **middle prompt**: The variation in the scores of the visual samples are influenced by: "inquisitive boy", "realistic photography style" and whether each individual mentioned in the prompt is accurately and uniquely generated in the image. **Row 4** corresponds to **long prompt**: The variation in the scores of the visual samples are influenced by: "ancient farmers", "sweat", "facial expressions", "relentless dedication". The mentioned keywords may correspond to more than one question–answer pair.

Table 15: **Best normalized polar visualization of SOTA methods on OneIG-Bench.** Anti-Clockwisely, **Alignment**: NP, T&P, Short, Middle, Long; **Text**: ED-Short, ED-Middle, ED-Long, CR-Short, CR-Middle, CR-Long, WAC-Short, WAC-Middle, WAC-Long; **Reasoning**: Geography, Computer Science, Biology, Mathematics, Physics, Chemistry, Common Sense; **Style**: Traditional, Media, Anime; **Diversity**: NP, T&P, Short, Middle, Long. The absolute average values of each evaluation metric are list on legends.

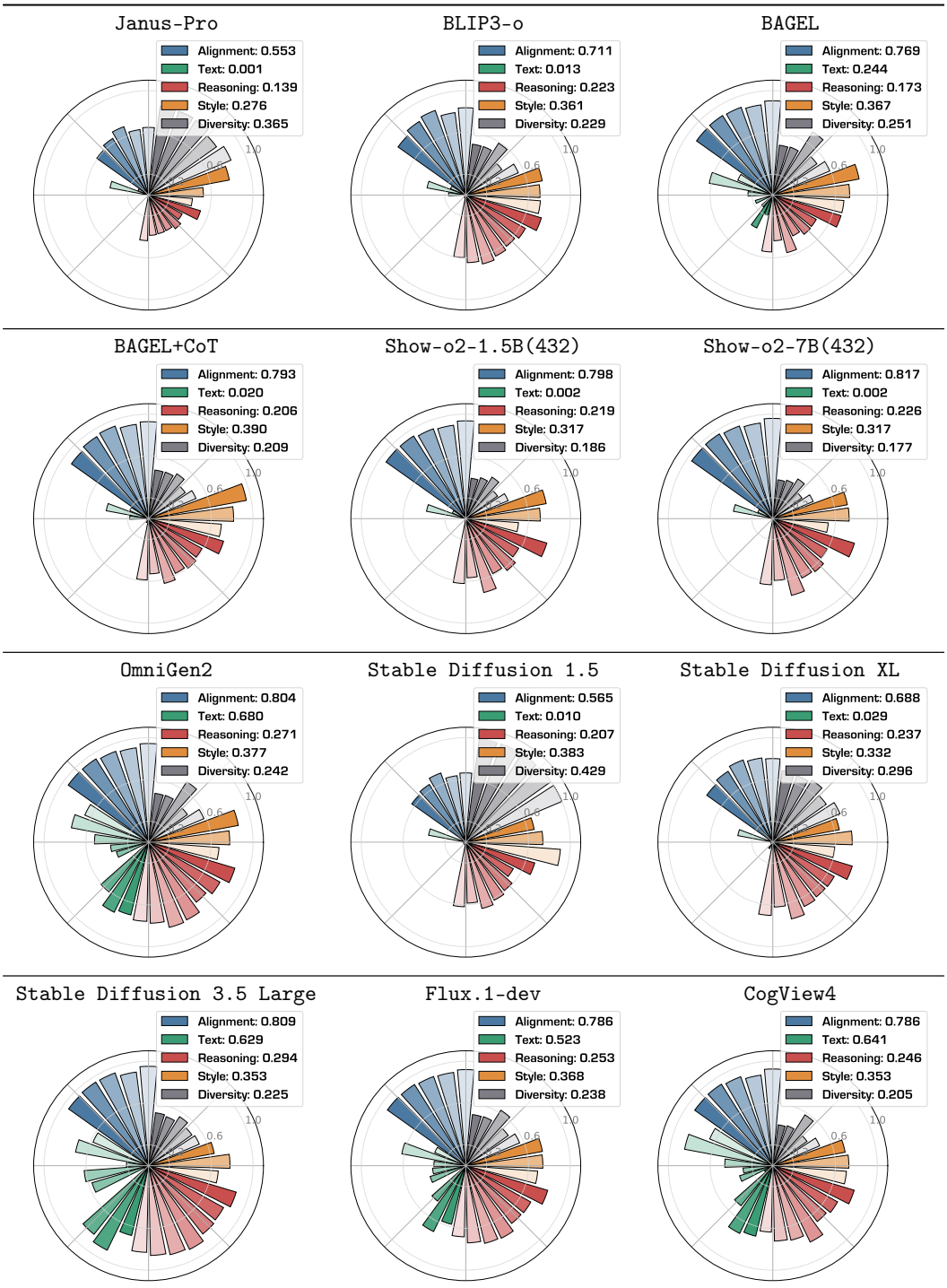


Table 15: **[Continued] Best normalized polar visualization of SOTA methods on OneIG-Bench.** Anti-Clockwisely, **Alignment**: NP, T&P, Short, Middle, Long; **Text**: ED-Short, ED-Middle, ED-Long, CR-Short, CR-Middle, CR-Long, WAC-Short, WAC-Middle, WAC-Long; **Reasoning**: Geography, Computer Science, Biology, Mathematics, Physics, Chemistry, Common Sense; **Style**: Traditional, Media, Anime; **Diversity**: NP, T&P, Short, Middle, Long. The absolute average values of each evaluation metric are list on legends.

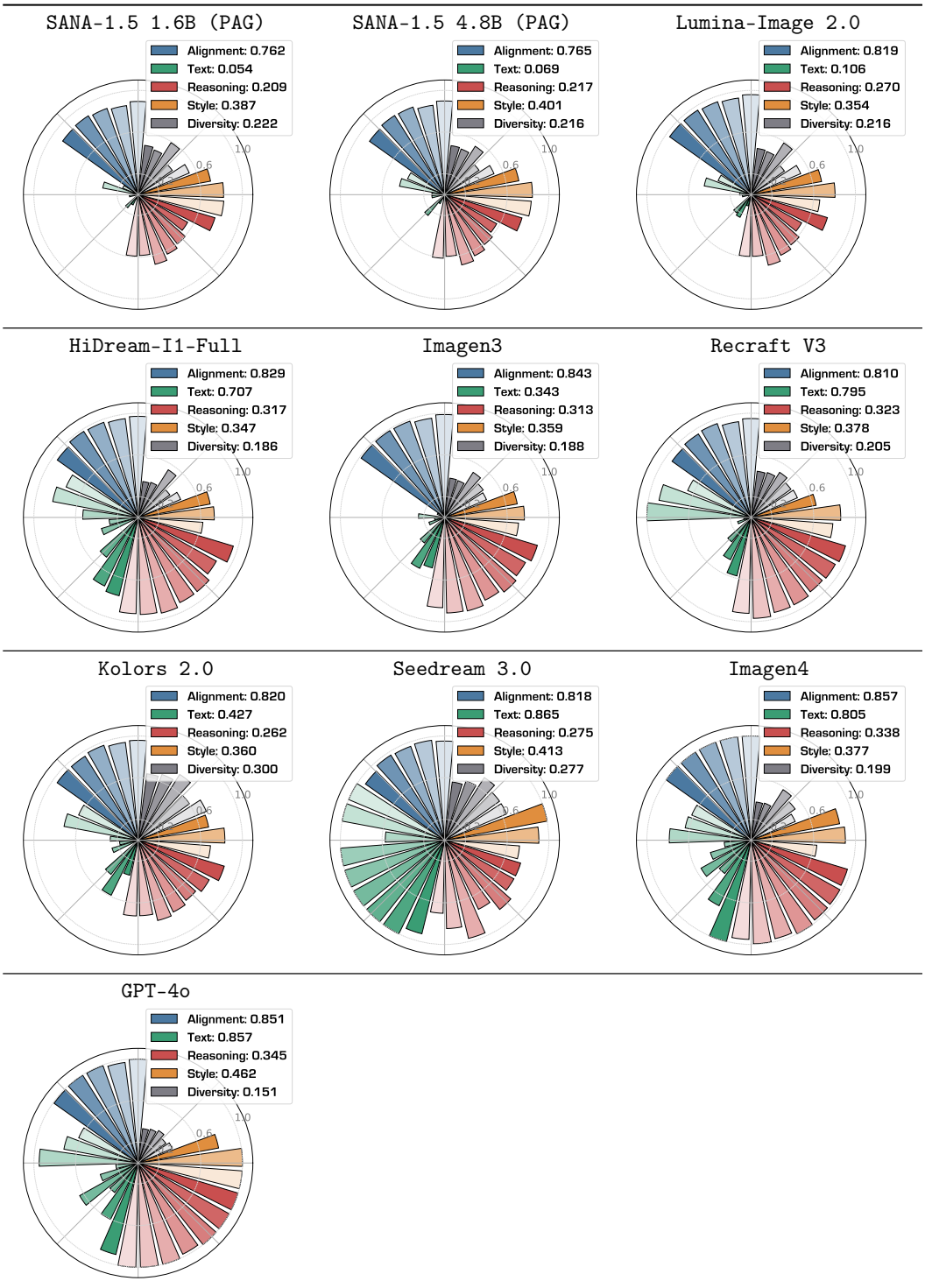
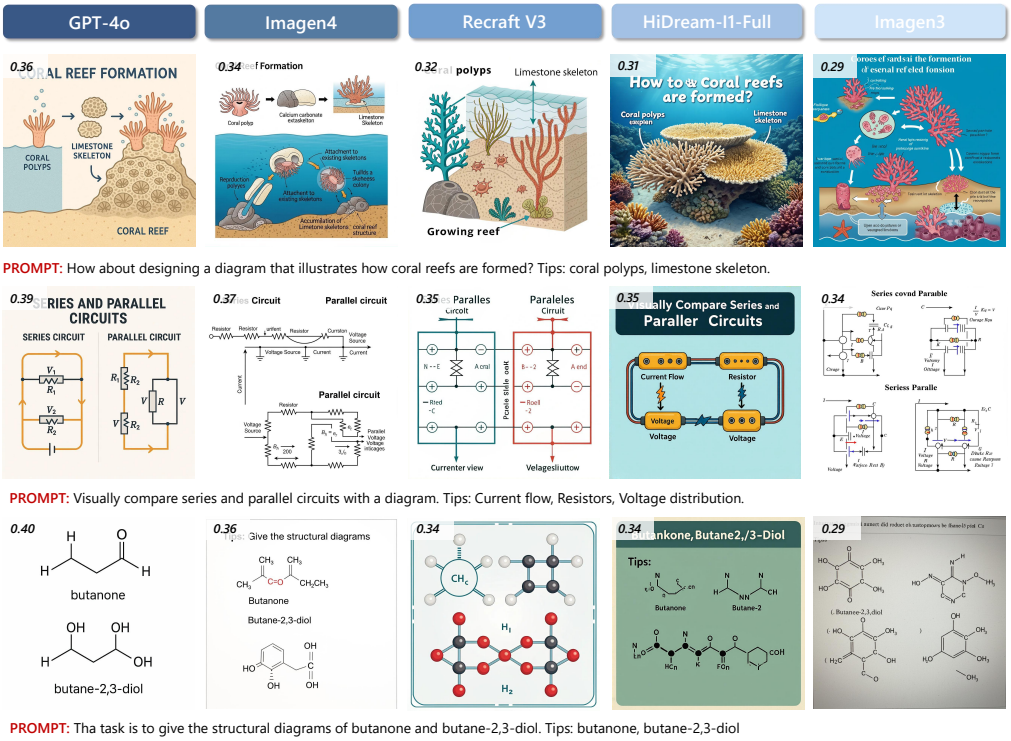


Figure 8: **Visualization results of SOTA methods. Text** of Seedream 3.0 [24], GPT-4o [45], Imagen4 [17], Recraft V3 [66], HiDream-I1-Full [29] are evaluated respectively. **Row 1, 2** correspond to **short prompts**. **Row 3, 4** correspond to **middle prompts**. **Row 5, 6** correspond to **long prompts**.

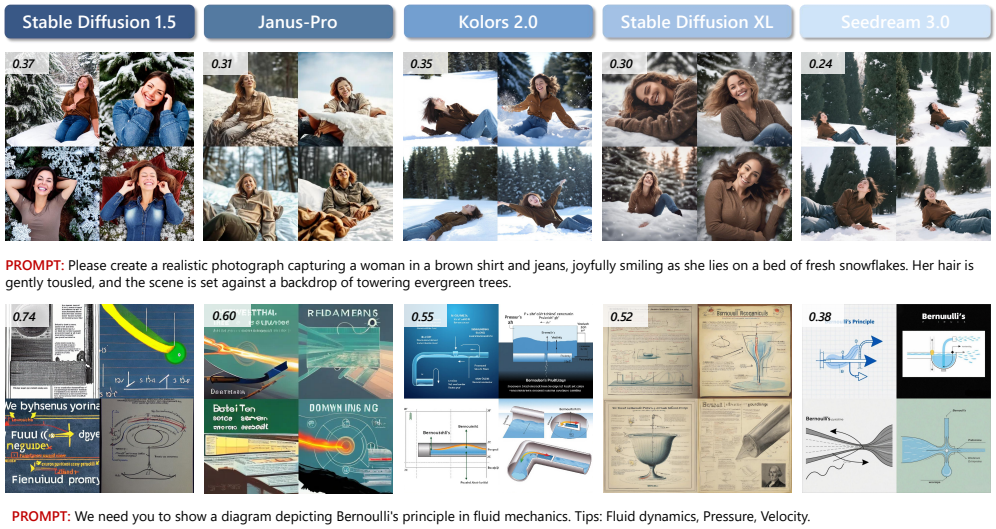


PROMPT: How about designing a diagram that illustrates how coral reefs are formed? Tips: coral polyps, limestone skeleton.

PROMPT: Visually compare series and parallel circuits with a diagram. Tips: Current flow, Resistors, Voltage distribution.

PROMPT: Tha task is to give the structural diagrams of butanone and butane-2,3-diol. Tips: butanone, butane-2,3-diol

Figure 9: **Visualization results of SOTA methods. Reasoning** of GPT-4o [45], Imagen4 [17], Recraft V3 [66], HiDream-I1-Full [29] and Imagen3 [36] are evaluated respectively. **Row 1** aims to illustrate how coral reefs are formed, highlighting key steps such as the growth of coral polyps and the gradual accumulation of calcium carbonate structures. **Row 2** demonstrates the circuit diagrams of series and parallel circuits. **Row 3** focuses on the structural diagrams of butanone and butane-2,3-diol.



PROMPT: Please create a realistic photograph capturing a woman in a brown shirt and jeans, joyfully smiling as she lies on a bed of fresh snowflakes. Her hair is gently tousled, and the scene is set against a backdrop of towering evergreen trees.

PROMPT: We need you to show a diagram depicting Bernoulli's principle in fluid mechanics. Tips: Fluid dynamics, Pressure, Velocity.

Figure 10: **Visualization results of SOTA methods. Diversity** of Stable Diffusion 1.5 [55], Janus-Pro [13], Kolors 2.0 [64], Stable Diffusion XL [46] and Seedream 3.0 [24] are evaluated respectively.

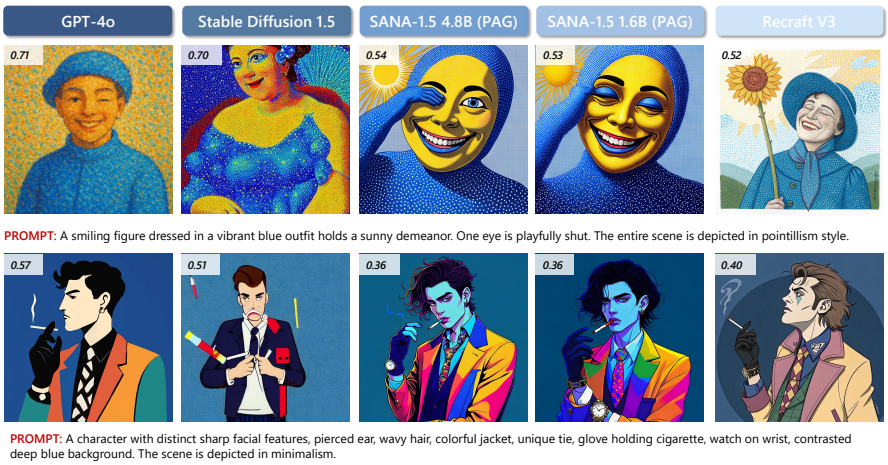


Figure 11: **Visualization results of SOTA methods. Traditional** styles of GPT-4o [45], Stable Diffusion 1.5 [55], SANA-1.5 4.8B (PAG) and 1.6B (PAG) [73], and Recraft V3 [66] are evaluated respectively. The styles are **pointillism** and **minimalism**.

Figure 12: **Visualization results of SOTA methods. Media** styles of GPT-4o [45], Imagen4 [17], Seedream 3.0 [24], Recraft V3[66] and SANA-1.5 4.8B (PAG) [73] are evaluated respectively. The styles are **pencil sketch** and **stone sculpture**.

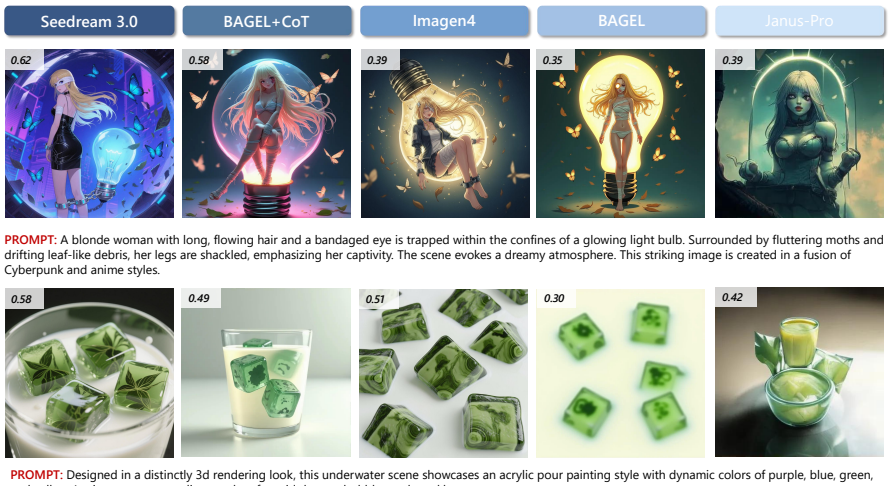


Figure 13: **Visualization results of SOTA methods. Anime** styles of Seedream 3.0 [24], BAGEL+CoT [18], Imagen4 [17], BAGEL [18] and Janus-Pro [13] are evaluated respectively The styles are **cyberpunk** and **3d rendering**.