OneIG-Bench: Omni-dimensional Nuanced Evaluation for Image Generation
– Supplemental Material –

Jingjing Chang[1,2] **Yixiao Fang**[1†] **Peng Xing**[1] **Shuhan Wu**[1] **Wei Cheng**[1]
Rui Wang[1] **Xianfang Zeng**[1] **Gang YU**[1‡] **Hai-Bao Chen**[2‡]
[1]StepFun [2]SJTU
[†]Project lead [‡] Corresponding author

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 1 and Figure 1, 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 [9], SigLIP [16], support a maximum of 77 tokens, a prompt of up to 60 words can generally be processed directly by these encoders.

Table 1: **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	50.1	0.275	0.475	0.250
Knowledge & Reasoning	20.5	0.960	0.018	0.022
Total Distribution	45.0	0.389	0.377	0.234
Total Distribution w/o K & R	51.1	0.246	0.467	0.287

The Portrait category, however, shows a slight deviation from this 1:2:1 distribution in Figure 2 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.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

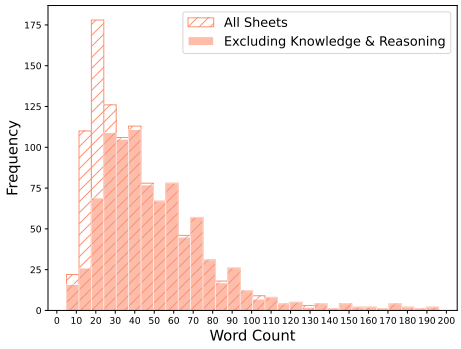


Figure 1: **Word Count of the Overall Prompts of OneIG-Bench**. The word count distribution of OneIG-Bench's prompts ranges from 0 to 200.

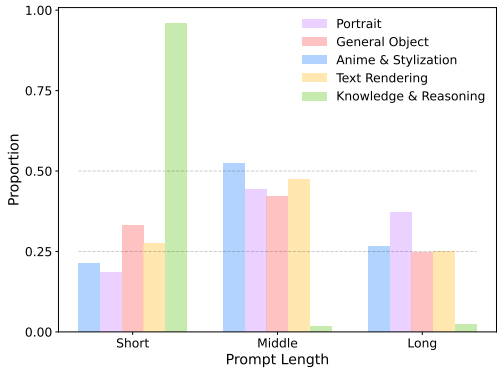


Figure 2: **The distribution of prompt word counts across Short, Middle, Long categories**.

A.2 Implementation

Our experiments on image generation with open-source methods are configured according to Table 2. 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 [11] from 40 to 50. The number of inference steps for Flux.1-dev [5] is set to be consistent with that used in the official API [1]. With the exception of Stable Diffusion 1.5 [10], 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 2: **Configurations used by open-source methods when generating images**. **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
Stable Diffusion 1.5 [10]	0.9B	7.5	512×512	50
Stable Diffusion XL [7]	2.6B	5.0	1024×1024	50
Stable Diffusion 3.5 Large [11]	8.1B	4.5	1024×1024	50
Flux.1-dev [5]	12B	3.5	1024×1024	28
CogView4 [14]	6B	3.5	1024×1024	50
SANA-1.5 1.6B (PAG) [15]	1.6B	5.0	1024×1024	20
SANA-1.5 4.8B (PAG) [15]	4.8B	5.0	1024×1024	20
Lumina-Image 2.0 [8]	2.6B	4.0	1024×1024	50
HiDream-I1-Full [4]	17B	5.0	1024×1024	50

For closed-source methods, we present the corresponding release or update dates of the methods in Table 3 to facilitate alignment with subsequent experimental results.

Table 3: **Release/Update date of closed-source methods**. **Release/Update Date** represents the version of the corresponding method when generating images.

Method	Imagen3 [2]	Recraft v3 [13]	Kolors 2.0 [12]	Seedream 3.0 [3]	GPT-4o [6]
Release/Update Date	2025-01-23	2024-10-30	2025-04-15	2025-04-15	2025-04-29

A.3 The Details on Prompts Rewriting

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 shown as Table 4.

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 $\mathbb{P}_{\text{rewritten}}$

Table 4: **Prompt Template for Rewriting Task**. **[Initial Prompt]** and **[Target Word Count]** correspond to the input arguments of the GPT-4o API function defined in Algorithm 1.

Prompt Template: Generating Rewritten Prompts Confronting the Specific Ratio
You are a precise rewriting assistant. **Task:** - 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 5: **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 5: **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 6 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 [6] demonstrates exceptional style-following ability across all categories, significantly outperforming other methods.

Table 6: **The style scores on different categories of styles**. indicate the first, second, third, fourth, and fifth performance, respectively.

Method	Traditional	Media	Anime
Stable Diffusion 1.5 [10]	0.483	0.298	0.349
Stable Diffusion XL [7]	0.316	0.307	0.339
Stable Diffusion 3.5 Large [11]	0.356	0.315	0.335
Flux.1-dev [5]	0.367	0.298	0.391
CogView4 [14]	0.376	0.294	0.369
SANA-1.5 1.6B(PAG) [15]	0.438	0.331	0.370
SANA-1.5 4.8B(PAG) [15]	0.443	0.340	0.379
Lumina-Image 2.0 [8]	0.351	0.325	0.360
HiDream-I1-Full [4]	0.331	0.295	0.368
Imagen3 [2]	0.378	0.309	0.371
Recraft v3 [13]	0.418	0.347	0.332
Kolors 2.0 [12]	0.370	0.336	0.360
Seedream 3.0 [3]	0.424	0.358	0.368
GPT-4o [6]	0.532	0.404	0.411

A.5 Visualization Results

We selected the top five 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 ordered from left to right according to their performance ranking, from highest to lowest. 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 3 shows that GPT-4o [6] demonstrates 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 4, Seedream 3.0 [3] demonstrates high accuracy and aesthetic quality. Some methods, such as GPT-4o [6], 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. Recraft v3 [13], 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 [4] 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. In contrast, Stable Diffusion 3.5 [11] exhibits a significant performance gap compared to the above four methods.

From a reasoning perspective, only GPT-4o [6] demonstrates both logical coherence and textual accuracy in Figure 5. Although Recraft v3 [13] falls short of GPT-4o in terms of clarity and correctness, it still produces text that is generally readable. HiDream-I1-Full [4] provides limited textual and visual content but manages to convey a certain degree of knowledge and reasoning, albeit with insufficient accuracy. In contrast, Imagen3 [2] tends to generate overly redundant outputs, with excessive textual and graphical elements that obscure the intended message, and often includes incorrect information. While Stable Diffusion 3.5 Large [11] outperforms several other methods,

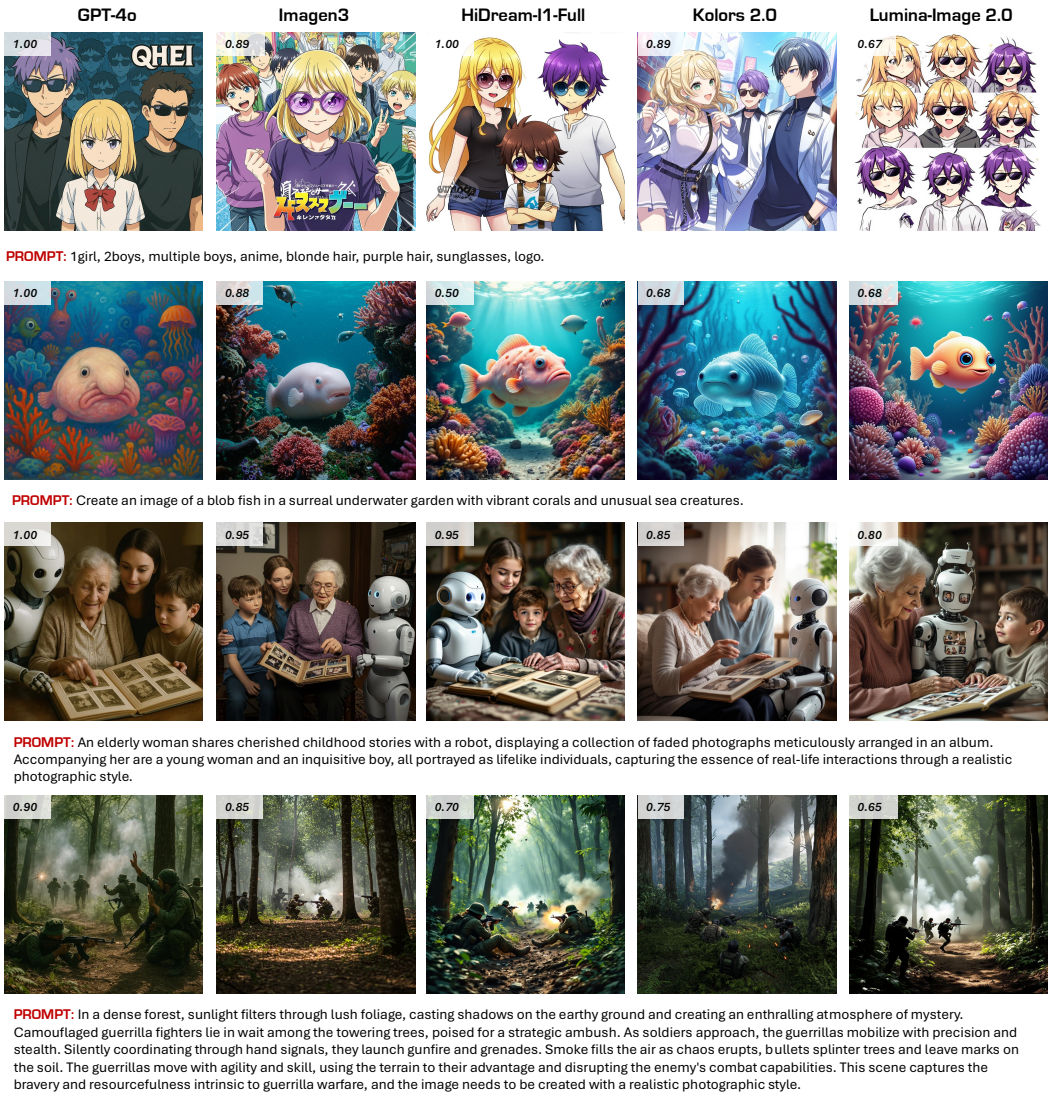


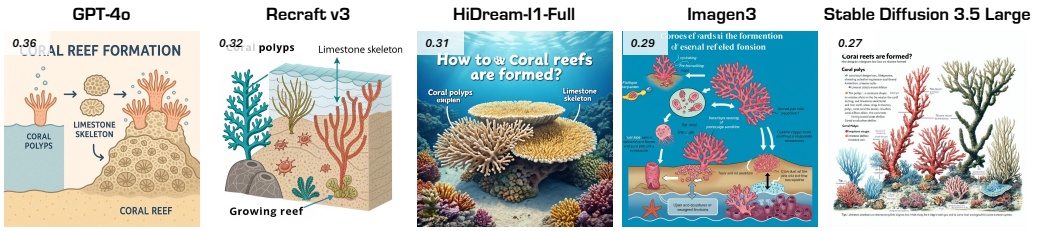
Figure 3: Visualization results of methods GPT-4o [6], Imagen3 [2], HiDream-I1-Full [4], Kolors 2.0 [12] and Lumina-Image 2.0 [8] on **alignment**. **Row 1** corresponds to **tag/phrase prompt**: The variation in the scores of the visual samples are mainly influenced by: "2 boys" and "logo". **Row 2** 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: "gunfire", "sunlight", "grenades". The mentioned keywords may correspond to more than one question–answer pair.

the performance gap between it and the leading methods remains substantial in Figure 5. Therefore, Knowledge and Reasoning remains a critical area that warrants further investigation and refinement for generative methods.

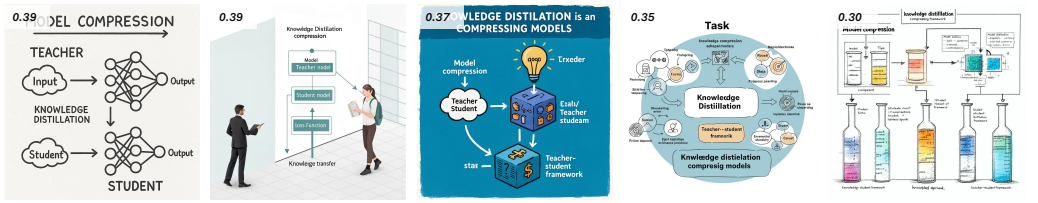
The visualization of diversity results is presented in Figure 6, where the ranking of diversity scores aligns well with visual inspection, supporting the validity of our proposed diversity metrics.

Figures 7, 8, and 9 show that GPT-4o [6] performs well across most styles, though it struggles with specific ones such as 3D rendering. Stable Diffusion 1.5 [10], despite its lower visual quality, effectively captures traditional style features. Overall, SANA-1.5 4.8B (PAG)[15] and Seedream 3.0[3] also demonstrate strong stylistic consistency, ranking just behind GPT-4o.

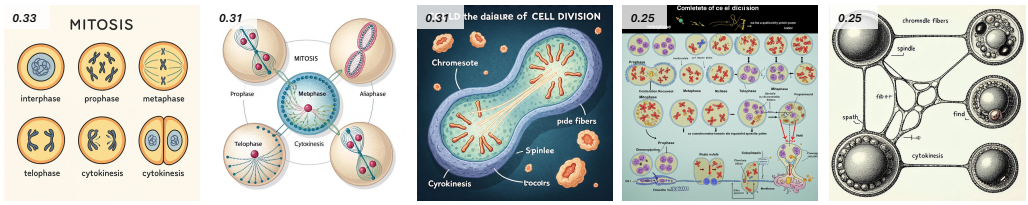
Figure 4: Visualization results of methods Seedream 3.0 [3], Recraft v3 [13], GPT-4o [6], HiDream-I1-Full [4] and Stable Diffusion 3.5 Large [11] on **text**. **Row 1, 2** correspond to **short prompts**. **Row 3, 4** correspond to **middle prompts**. **Row 5, 6** correspond to **long prompts**.



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



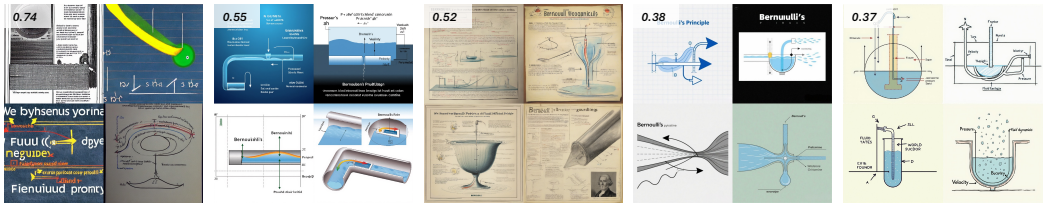
PROMPT: The task is to sketch a diagram illustrating how knowledge distillation is applied in compressing models. Let's delve into this approach and visualize its key components. Tips: Model compression, knowledge distillation, teacher-student framework.



PROMPT: Could you draw the complete sequence of cell division, focusing on the stages of mitosis? Tips: chromosomes, spindle fibers, cytokinesis.

Figure 5: Visualization results of methods GPT-4o [6], Recraft v3 [13], HiDream-I1-Full [4], Imagen3 [2] and Stable Diffusion 3.5 Large [11] on **reasoning**. **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 working mechanism of knowledge distillation in model compression, which must include both the teacher model and the student model. **Row 3** focuses on the stages of mitosis, including prophase, metaphase, anaphase, and telophase.

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 6: Visualization results of methods Stable Diffusion 1.5 [10], Kolors 2.0 [12], Stable Diffusion XL [7], Seedream 3.0 [3] and Flux-1.dev [5] on **diversity**.

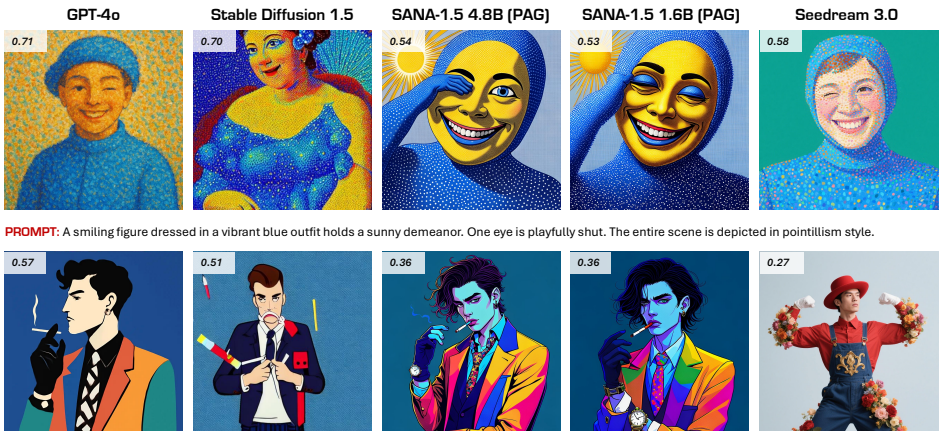


Figure 7: Visualization results of methods GPT-4o [6], Stable Diffusion 1.5 [10], SANA-1.5 4.8B (PAG) and 1.6B (PAG) [15], and Seedream 3.0 [3] on **traditional** styles. The styles are **pointillism** and **minimalism**.

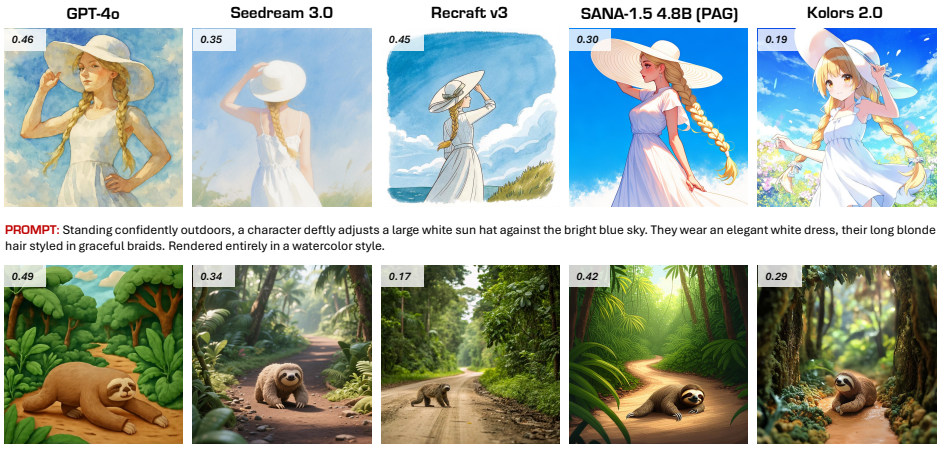


Figure 8: Visualization results of methods GPT-4o [6], Seedream 3.0 [3], Recraft v3[13], SANA-1.5 4.8B (PAG) [15] and Kolors 2.0 [12] on **media** styles. The styles are **watercolor** and **clay**.

Figure 9: Visualization results of methods GPT-4o [6], Flux-1.dev [5], SANA-1.5 4.8B (PAG) [15], Imagen3 [2] and SANA-1.5 1.6B(PAG) [15] on **anime** styles.The styles are **comic** and **3d rendering**.

References

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

[2] Google deepmind Imagen3 team. Imagen3. https://deepmind.google/technologies/imagen-3/, 2024.

[3] 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.

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

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

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

[13] 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.

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

[15] 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.

[16] 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.