UniGenBench++: A Unified Semantic Evaluation Benchmark for Text-to-Image Generation

Yibin Wang[1,2,4*], **Zhimin Li**[4*], **Yuhang Zang**[3*], **Jiazi Bu**[3,5], **Yujie Zhou**[3,5], **Yi Xin**[2], **Junjun He**[2,3], **Chunyu Wang**[4], **Qinglin Lu**[4], **Cheng Jin**[1,2†], **Jiaqi Wang**[2†]
[1]Fudan University, [2]Shanghai Innovation Institute [3]Shanghai AI Lab,
[4]Hunyuan, Tencent, [5]Shanghai Jiaotong University
Project Page: codegoat24.github.io/UniGenBench/

Fig. 1: **Benchmark Overview.** (1) Our **UniGenBench++** covers diverse prompt themes, subjects, and comprehensive evaluation criteria. (2) Each prompt includes multiple test points and is assessed through a streamlined MLLM-based pipeline for reliable and efficient evaluation. (3) We conduct comprehensive evaluations of both open- and closed-source models using both English and Chinese prompts in short and long forms, systematically revealing their strengths and weaknesses across various aspects.

***Abstract*—Recent progress in text-to-image (T2I) generation underscores the importance of reliable benchmarks in evaluating how accurately generated images reflect the semantics of their textual prompt. However, (1) existing benchmarks lack the diversity of prompt scenarios and multilingual support, both essential for real-world applicability; (2) they offer only coarse evaluations across primary dimensions, covering a narrow range of sub-dimensions, and fall short in fine-grained sub-dimension assessment. To address these limitations, we introduce UniGenBench++, a unified semantic assessment benchmark for T2I generation. Specifically, it comprises 600 prompts organized hierarchically to ensure both coverage and efficiency: (1) spans across diverse real-world scenarios, *i.e.,* 5 main prompt themes and 20 subthemes; (2) comprehensively probes T2I models' semantic consistency over 10 primary and 27 sub evaluation criteria, with each prompt assessing multiple testpoints. To rigorously assess model robustness to variations in prompt length and language, we provide both English and Chinese versions of each prompt in short and long forms. Leveraging the general world knowledge and fine-grained image understanding capabilities of a closed-source Multi-modal Large Language Model (MLLM), *i.e.,* Gemini-2.5-Pro, an effective pipeline is developed for reliable benchmark construction and streamlined model assessment. Moreover, to further facilitate community use, we propose a robust evaluation model that enables offline assessment of T2I model outputs. Through comprehensive benchmarking of both open- and closed-sourced T2I models, we systematically reveal their strengths and weaknesses across various aspects.**

***Index Terms*—Text-to-image generation, semantic evaluation, and benchmark.**

TABLE I
SEMANTIC EVALUATION BENCHMARK COMPARISON. "-" INDICATES THAT THE ASPECT IS NOT DISCUSSED IN ITS ORIGINAL PAPER.

Benchmark	Primary Dimension	Sub Dimension	Prompt Theme	Prompt Length	Prompt Num.	Multi-Testpoint per Prompt	Multilingual Support	Dedicated Offline Eval Model
GenEval	6	-	-	short	553	✗	✗	✓
T2I-CompBench++	8	-	-	short	2,400	✗	✗	✓
DPG-Bench	5	-	-	long	1,065	✗	✗	✗
WISE	6	-	-	short	1,000	✗	✗	✗
TIIF-Bench	9	-	-	short/long	5,000	1∼2	✗	✗
UniGenBench++ (**Ours**)	10	27	20	short/long	600	1∼10	✓	✓

I. INTRODUCTION

RECENT progress in text-to-image (T2I) generation [1]–[19] has highlighted the ability to generate high-quality images directly from natural language descriptions. Technically, current T2I models can be broadly divided into two paradigms. (1) Diffusion-based methods, including Stable Diffusion [2], [5], Playground [16], and FLUX [9], [19], iteratively refine Gaussian noise using U-Net or Transformer backbones to generate images. (2) Autoregressive (AR) approaches, such as Infinity [20], Janus series [21]–[23], and BLIP3-o [24], treat images as token sequences and synthesize them via next-token prediction or progressive scaling. Recent methods incorporate reinforcement learning [25]–[28] to improve T2I models' instruction following capability [29], [30] and the visual quality of generated images [31]–[33]. With these rapid advancements, assessing T2I models, particularly their semantic generation capability, *i.e., how accurately generated images reflect the semantics of their textual prompt*, has emerged as a critical challenge. Traditional benchmarks [34], [35] typically evaluate T2I models by probing various compositional generation and employ CLIP-based metrics for quantitative assessment. However, CLIP-based scorers remain limited in capturing the fine-grained semantic information and complex world knowledge or logical reasoning. Therefore, several studies [36], [37] evaluate the implicit semantic understanding and world knowledge integration capabilities of T2I models using powerful visual-language models (VLMs) [38] as the evaluator. Recent efforts broaden T2I evaluation by incorporating long-prompt semantics generation [39], [40] and additional evaluation dimensions [40] such as style and text generation.

Despite effectiveness, as shown in Tab. I, these benchmarks encounter two key limitations: (1) **Coarse evaluation on limited dimensions:** cover limited general dimensions (*e.g.,* lacking *grammar*, *action*), within which the sub-dimension coverage is also limited (*e.g.,* lacking *relation-similarity*, *inclusion*), and incapable of fine-grained assessment for each sub-dimension; (2) **Lacking diversity of prompt scenarios and multilingual evaluation:** only focus on evaluation dimension design but neglect the diversity of prompt scenarios and multilingual evaluation support, hindering comprehensive assessment in real-world applicability.

In light of these challenges, this work posits that **(1)** existing T2I models have already shown strong performance on several primary dimensions (*e.g.,* attributes) in current benchmarks [34], [39], [40]. This highlights the necessity of further decomposing these dimensions into explicit, comprehensive sub-dimension-level test points (*e.g.,* attribute-expression) to enable a more comprehensive and diagnostic evaluation of model capabilities, thereby uncovering fine-grained weaknesses that coarse metrics often overlook. **(2)** Real-world T2I generation involves diverse scenarios (*e.g.,* UI design, graphic art) and naturally spans multiple languages. The absence of such diversity in current benchmarks limits evaluation robustness, causing models that excel in constrained settings to falter in real-world applications.

To this end, we introduce **UNIGENBENCH++**, a unified semantic-generation benchmark tailored for fine-grained and comprehensive evaluation of T2I models. As illustrated in Fig. 1 (1), this benchmark **comprises 600 prompts organized within a hierarchical structure that ensures both coverage and efficiency**: (i) It provides a comprehensive assessment of semantic consistency across 10 primary and 27 sub-dimensions, each prompt targeting multiple specific test points. This design strikes a balance between fine-grained evaluation and efficiency, ensuring the benchmark captures diverse aspects of model semantic generation capability. (ii) It spans 5 major real-world primary generation scenarios and 20 sub-scenarios with diverse subject categories, encompassing practical domains that reflect authentic user requirements, thereby enabling evaluation under conditions that closely mirror real-world usage. Besides, to enable systematic evaluation of models' sensitivity to language and prompt length, each prompt is provided in both English and Chinese, and in short and long forms. For effective and efficient evaluation, in contrast to widely adopted paradigms, such as multi-turn conversational assessments with VLMs for each image evaluation [34], [35], [40], our benchmark **introduces a streamlined, point-wise evaluation pipeline**, as illustrated in Fig. 1 (2): given a prompt, its corresponding image, and a set of explicit designed test points (each accompanied by its in-context description within the prompt), the evaluation model, *i.e.,* Gemini-2.5-pro [41], sequentially analyses whether each semantic requirement is faithfully represented in the image and assigns an appropriate score. This lightweight and structured design reduces evaluation complexity while ensuring consistent, fine-grained, and interpretable judgments for every test point, thereby enabling more efficient and diagnostic assessment of T2I models. Moreover, to further facilitate community use, we provide a robust evaluation model that supports offline assessment of T2I model outputs.

We conduct a comprehensive bilingual (English/Chinese) and length-varied (short/long prompt) benchmarking across both closed-source models, such as GPT-4o [14], Nano Banana [13], Seedream-4.0 [11], and FLUX-kontext-max [9], as well as leading open-source counterparts, including Qwen-Image [15], Hidream [12], Lumina-DiMOO [42] and Bagel [43].

Fig. 2. **Benchmark Statistics.** (a) Word clouds for English and Chinese prompts in both short and long forms; (b) overall prompt length distribution; and (c) distribution of testpoint counts per prompt for short versus long versions.

As shown in Fig. 1 (3), both leading open- and closed-source models exhibit strong performance on prompts involving style and world knowledge, yet consistently struggle with logical reasoning that requires causal, contrastive, or other complex relational understanding. Furthermore, open-source models show larger performance fluctuations across dimensions, particularly underperforming in the *grammar* and *action* dimensions. This highlights the models' difficulty in handling grammar-conditioned instructions and depicting dynamic or behavior-centric content accurately.

The contributions of this paper are summarized as follows:

- We propose **UniGenBench++**, a unified benchmark for text-to-image (T2I) semantic generation evaluation, covering comprehensive evaluation dimensions, diverse prompt themes, and rich subject categories. Each prompt is provided in both English and Chinese, and in short and long forms, assessing multiple testpoints, ensuring both coverage and efficiency.
- We design a **streamlined, point-wise evaluation pipeline** that minimizes evaluation complexity while ensuring consistent, fine-grained, and interpretable judgments at the testpoint level.
- We provide a **dedicated offline evaluation model** that enables robust assessment of T2I model outputs to further facilitate community use.
- We conduct **extensive bilingual and length-varied benchmarking** across both closed- and open-source models, systematically revealing their strengths and weaknesses across diverse semantic aspects.

We hope that our benchmark could advance the development and evaluation of T2I models, driving further improvements in semantic consistency across diverse fine-grained tasks and fostering deeper insights into model performance across real-world scenarios.

II. Related Work

Text-to-Image Generation. Recent progress in text-to-image (T2I) generation is largely driven by two paradigms: diffusion-based and autoregressive (AR) models. Diffusion models dominate current practice due to their scalability and photorealistic synthesis, progressively denoising Gaussian noise conditioned on text, evolved from early GLIDE [44] and Imagen [18] to powerful variants like Stable Diffusion [2], FLUX [9], and HiDream [12]. In contrast, AR models generate images token by token via VQ-VAE [45] compression and transformer decoding, as seen in DALL·E [4] and CogView [46]. Recent advances [47], [48] enhance AR models with unified multimodal reasoning, while hybrid architectures like Bagel [43] integrate both diffusion and AR to enable explicit reasoning before image generation. With such rapid advances, evaluating T2I models, especially their semantic generation capability, has become a central challenge.

Text-to-Image Benchmarks. Prior studies commonly assess T2I models through compositional generation tests. For example, GenEval [34] leverages object detection to rigorously verify whether generated images accurately reflect the spatial arrangements, numerical counts, and color attributes specified in the textual prompts. T2I-CompBench [35] encompasses four core compositional categories and further extends these evaluations with detection-based metrics for spatial reasoning and numerical consistency. Several studies evaluate T2I models through specific knowledge domains, such as physical reasoning [37] and general commonsense understanding [36]. However, the prompts used in these benchmarks are predominantly short and highly repetitive, which constrains semantic richness and expressiveness. Therefore, DPG-Bench [39] centers on assessing models' capability in dense prompts. TIIF-Bench [40] offers both short and long variants of each prompt while preserving identical core semantics.

Despite their effectiveness, these benchmarks still suffer from coarse evaluation across limited dimensions and provide insufficient sub-dimension coverage. Moreover, the lack of diverse prompt scenarios and multilingual support further limits their ability to assess models in real-world application settings. To this end, we introduce **UNIGENBENCH++**, a unified semantic-generation benchmark designed for fine-grained and comprehensive evaluation of T2I models.

III. BENCHMARK

A. Overview

With the rapid advancement of text-to-image (T2I) models, existing evaluation frameworks [34], [35], [39], [40] have become increasingly insufficient. To be precise, (1) as summarized in Tab. I, they often overlook diversity in prompt scenarios and lack multilingual coverage, both of which are indispensable for real-world applicability. Consequently, their evaluations fall short in capturing a model's true applicability across diverse and contextually complex input conditions; (2) although existing benchmarks effectively assess a few broad dimensions, they still overlook several critical semantic aspects and lack systematic coverage and evaluation at the sub-dimension level, ultimately limiting their fine-grained diagnostic capability.

To this end, we propose **UNIGENBENCH++**, a unified semantic evaluation benchmark for T2I generation. As summarized in Fig. 1 and Tab. I, our benchmark offers several key advantages over existing studies:

- **Rich prompt theme design.** Prompts are hierarchically organized into 5 primary themes and 20 sub-themes, spanning both practical real-world use cases and open-ended imaginative scenarios (Sec. III-B).
- **Comprehensive semantic dimension coverage.** It evaluates 10 primary dimensions and 27 sub-dimensions, enabling systematic diagnosis of diverse model capabilities. Despite its breadth, it requires only 600 prompts, each targeting 1–10 explicit test points, achieving a favorable balance between coverage and efficiency (Sec. III-C).
- **Bilingual and length-variant prompt and streamlined model evaluation.** All prompts are provided in both English and Chinese, each available in both short and long forms (Sec. III-D). Leveraging the world knowledge and fine-grained image understanding capabilities of Multimodal Large Language Models (MLLMs), *i.e.,* Gemini-2.5-pro, we design a fully streamlined pipeline for accurate and efficient model evaluation (Sec. III-E).
- **Reliable evaluation model for offline assessment.** To facilitate community use, we train a robust evaluation model that supports offline assessment of T2I model outputs (Sec. III-F).

B. Prompt Themes and Subject Categories

This work posits that diverse prompt themes better approximate real-world usage scenarios, thereby yielding a more faithful evaluation of model performance. Therefore, we organize prompt scenarios based on common real-world usage needs. Specifically, as illustrated in Fig. 1 (1.a), we structure them into 5 primary categories and 10 finer sub-categories to ensure both breadth and practical relevance:

- **Creative Divergence** covers open-ended imaginative ideation and broader forms of other abstract conceptual composition.
- **Art** encompasses a wide range of visual expression styles, including graphic renderings, photography-inspired depictions, sculptural aesthetics, and other fine-art formats.
- **Illustration** is divided into copywriting-oriented visualization (*e.g.,* , slogans or metaphors) and content-centric narrative illustration.
- **Film & Story** accounts for settings across cinematic realism, speculative or science-fiction narratives, and animation-style storytelling.
- **Design** spans professional and commercial use cases such as advertising and e-commerce graphics, spatial layouts, game and UI prototyping, poster composition, IP and logo/icon creation, fashion concept design, and general-purpose design resource generation.

To facilitate understanding of each theme, we present representative prompts in Tab. VI.

Based on a wide range of prompt themes, we further define a diverse set of subject categories to cover different types of entities. As illustrated in Fig. 1 (1.d), these categories include *animals*, *objects*, *anthropomorphic characters*, *scenes*, as well as an *Other* category for special or atypical entities (e.g., robots appearing in science-fiction prompts). To this end, the benchmark can probe model capabilities on both common and unusual entities, providing insights into model strengths and weaknesses across diverse semantic scenarios.

The distribution of prompt themes and subject categories is illustrated in Fig. 1 (1.a) and (1.d), respectively.

C. Evaluation Dimensions

Existing T2I models have demonstrated strong performance on several primary evaluation dimensions in current benchmarks. However, this surface-level success often masks their underlying weaknesses at the sub-dimension level, as coarse-grained metrics are insufficient to reveal fine-grained limitations in specific sub-aspects.

To address this gap, we decompose each major dimension into explicit and comprehensive sub-dimension-level test points. Specifically, our benchmark organizes evaluation dimensions into 10 major categories, most of which encompass multiple subcategories:

1. Style evaluates the model's ability to generate images with coherent style and artistic expression. It considers both overall visual style and artistic genre, ensuring that the generated images exhibit plausible and consistent artistic characteristics.

2. World Knowledge examines the model's grasp of real-world concepts. It evaluates whether the model can generate content consistent with physical laws, cultural norms, geographical facts, and historical context.

3. Attribute assesses the model's understanding of object and scene characteristics, including:

- **Quantity**: The number of objects or elements in a scene.

Fig. 3. **Qualitative Results of Evaluation Dimensions.** We present qualitative examples of T2I models evaluated across our specified dimensions.

- **Expression**: Emotional states or facial expressions of humans or animals.
- **Material**: Surface properties of objects, such as wood, metal, or glass.
- **Color**: Accuracy and appropriateness of colors and color combinations.
- **Shape**: Geometric form and contour of objects.
- **Size**: Relative dimensions of objects within the scene.

4. Compound evaluates the model's ability to combine multiple concepts or features:

- **Imagination**: Creativity in generating novel or non-realistic combinations.
- **Feature Matching**: Coherent integration of different elements and their attributes.

5. Action focuses on the dynamic behaviors and interactions of characters, animals, or objects:

- **Contact Interaction**: Physical interactions between objects, such as torching and holding.
- **Non-contact Interaction**: Non-physical interactions like gazing.
- **Hand Actions** : Representation of hand gestures or manipulations.
- **Full-body Actions**: Depiction of whole-body movements of characters.
- **State**: Status or posture of objects or characters, such as sleeping, suspending, or running.
- **Animal Actions**: Behaviors specific to animals.

6. Entity Layout evaluates spatial arrangement and composition:

- **Two-Dimensional Space**: Layout and relative positions of objects on a plane.
- **Three-Dimensional Space**: Layout and relative positions of objects in three-dimensional space

7. Relationship assess the semantic and logical connections between objects:

- **Composition**: Integration of multiple elements into a coherent whole.

- **Similarity**: Similarity in shape, color, or material between objects.
- **Comparison**: Differences and contrasts between objects.
- **Inclusion**: Containment or hierarchical relationships among objects.

8. Logical Reasoning measures the model's ability to reason about events, object attributes, understand causality, and contrastive expressions.

9. Grammar evaluates the model's understanding of textual and language-related expressions:

- **Pronoun Reference**: Correct association between pronouns and their referents in the image.
- **Consistency**: Maintenance of coherent attributes, properties, or features across objects as described in the prompt.
- **Negation**: Accurate reflection of negation or exclusion expressions in the generated content.

10. Text Generation evaluates the model's ability to generate text content that is accurate, readable, and aligned with the requirements of the input prompt.

We provide qualitative examples of our evaluation dimensions in Fig. 3. Notably, in our benchmark, the distribution of testpoints differs between short and long prompts. Specifically, long prompts tend to have more attribute-related test points, as they provide more detailed and diverse descriptions of subjects, attributes, and scenes. The test point distribution for both is shown in Fig. 1 (1.b) and (1.c).

D. Bilingual and Length-variant Prompt Construction

Bilingual Short Prompt Generation. Let $\mathcal{T}$ denote the set of prompt *themes*, $\mathcal{S}$ the set of *subject categories*, and $\mathcal{C}$ the set of *evaluation dimensions*. For each prompt construction step, a theme $t \sim \mathcal{T}$ and a subject category $s \sim \mathcal{S}$ are first sampled uniformly at random. Subsequently, a subset of k testpoints $c_1, \ldots, c_k \subset \mathcal{C}$, where $k \in [1, 5]$, is selected to specify the targeted fine-grained testpoints.

Given the input tuple $(t, s, c_1, \ldots, c_k)$, the MLLM produces two outputs: (i) a pair of natural language prompts $(p^{\text{en}}, p^{\text{zh}})$ in English and Chinese, both adhering to the semantic constraints imposed by the selected theme t and subject category s; and (ii) a structured description set $d_1, \ldots, d_k$, where each element explicitly explains how the corresponding testpoint c_i is instantiated within the generated prompts. Formally:

$$\left(p^{\text{en}}, p^{\text{zh}}, \{d_1, \ldots, d_k\}\right) \sim \text{MLLM}_{\text{gen}}\left(t, s, \{c_1, \ldots, c_k\}\right), \quad (1)$$

Expanded to Long Prompt. To enrich the descriptive diversity and specificity of the generated prompts, we further expand each short prompt into a long-form prompt through rewriting strategy. Given a short prompt p^{en} or p^{zh}, we instruct the MLLM to generate an expanded version $\tilde{p}$ that satisfies two constraints: (i) the prompt theme, core subject and its key attributes must be preserved, and (ii) attribute, scene, and background details may be further elaborated to enhance specificity and imagination. Formally,

$$\tilde{p} \sim \text{MLLM}_{\text{expand}}\left(p \mid r\right), \quad (2)$$

where r denotes the rewriting constraint.

Fig. 4. **Pipeline of Benchmark Construction and Offline Evaluation Model Training.** (a) Benchmark construction pipeline; (b) Offline evaluation model training; (c) Offline evaluation cases.

However, expanding a prompt may introduce new semantic elements that are not covered by the original evaluation dimensions, or render some of the initial testpoints no longer applicable. To maintain consistency between the expanded prompt and its associated testpoints, we perform a second refinement step. Given the expanded prompt $\tilde{p}$ and the original testpoints $\{c_1, \ldots, c_k\}$ with their descriptions $\{d_1, \ldots, d_k\}$, we instruct the MLLM to revise the testpoint set by: (i) removing those no longer grounded in $\tilde{p}$; (ii) adding newly emerged testpoints, with a maximum allowance of five additional entries; and (iii) updating the in-context descriptions for all retained or newly added testpoints to reflect the semantics of $\tilde{p}$. Formally, the alignment process is defined as

$$\left\{(\hat{c}_1, \hat{d}_1), \ldots, (\hat{c}_{k'}, \hat{d}_{k'})\right\} \sim \text{MLLM}_{\text{align}}\Big(\cdot \;\Big|\; \tilde{p},\ \{(c_i, d_i)\}_{i=1}^{k} \Big), \quad k' \leq k + 5.$$

where k' is determined dynamically by the updated semantic scope of $\tilde{p}$. The resulting tuple

$$\left(\tilde{p},\ \{(\hat{c}_1, \hat{d}_1), \ldots, (\hat{c}_{k'}, \hat{d}_{k'})\}\right)$$

constitutes a semantically coherent long-prompt entry paired with aligned fine-grained evaluation targets.

Fig. 5. **Evaluation Accuracy Comparison.** Our dedicated evaluation model demonstrates a significant improvement in evaluation accuracy across all test points compared to the commonly used offline evaluation VLM, *i.e.,* Qwen2.5-VL-72b.

The word clouds of both English and Chinese prompts in short and long forms are visualized in Fig. 2 (a). We also present statistics on the length distribution of prompts in Fig. 2 (b), as well as the distribution of test point counts between short and long prompts in Fig. 2 (c).

E. T2I Model Evaluation

To systematically evaluate the quality of model-generated images, we employ a MLLM, *i.e.,* Gemini-2.5-pro, as an automatic evaluator. For each test prompt p_i, the corresponding generated image x_i is paired with a set of fine-grained testpoints $\{c_{i,1}, \ldots, c_{i,k}\}$ and their descriptions $\{d_{i,1}, \ldots, d_{i,k}\}$. Since each test point corresponds uniquely to its description, we henceforth refer only to the descriptions $\{d_{i,j}\}$ for brevity. Then, the MLLM takes $(x_i, p_i, \{d_{i,j}\})$ as input and performs an independent assessment for each testpoint. For each $d_{i,j}$, it returns both a binary decision $r_{i,j} \in \{0, 1\}$, indicating whether the requirement is satisfied, and a natural-language explanation $e_{i,j}$, which articulates the reasoning behind the judgment. This process is formally expressed as:

$$\begin{aligned}&(r_{i,1}, \ldots, r_{i,k}, e_{i,1}, \ldots, e_{i,k}) \\ &\sim \mathrm{MLLM}\Big(\{r_{i,j}, e_{i,j}\} \mid x_i, p_i, \{d_{i,1}, \ldots, d_{i,k}\}\Big).\end{aligned} \tag{3}$$

Compared to scalar-only metrics, this formulation not only quantifies correctness but also reveals failure modes by exposing *why* a testpoint is considered satisfied or violated. The availability of rationales $e_{i,j}$ further facilitates human verification and downstream error attribution. We provide example evaluation case in Fig. 1 (2).

Once all evaluation results are collected, we aggregate them at both the sub-dimension and primary-dimension levels. For each sub-dimension c, which groups semantically related testpoints, its score is defined as the ratio of satisfied instances to the total number of its occurrences across the benchmark:

$$R_c = \frac{\sum_{i,j} \mathbf{1}\{d_{i,j} \in c \text{ and } r_{i,j} = 1\}}{\sum_{i,j} \mathbf{1}\{d_{i,j} \in c\}}, \tag{4}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Higher-level primary dimensions C are then scored by averaging over their constituent sub-dimensions.

This hierarchical aggregation strategy enables multi-granular evaluation: it reflects fine-grained capability trends while also supporting concise reporting at a holistic level. Moreover, by separating binary correctness from explanatory evidence, our protocol provides both *quantitative comparability* and *qualitative interpretability*, which are crucial for diagnosing the strengths and weaknesses of T2I models at scale.

F. Offline Evaluation Model Training

To facilitate convenient and cost-efficient evaluation for the community, we further train an *offline evaluation model* that serves as a lightweight substitute for proprietary MLLMs during inference. Instead of querying a proprietary model online for every evaluation instance, our goal is to distill its scoring behavior into a compact model that can be executed locally without external API calls.

The supervision signals are constructed as described above from the online MLLM evaluator: for each image–prompt pair (x_i, p_i) and testpoint description $\{d_{i,j}\}$, the reference outputs $(r_{i,j}, e_{i,j})$ produced by the MLLM are collected and assembled into target sequences for supervised fine-tuning. Formally, given the tokenized target sequence y_i associated with input $(x_i, p_i, \{d_{i,j}\})$, the training objective is:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T_i} \log P_\theta\big(y_i^{(t)} \mid y_i^{(<t)},\, x_i, p_i, \{d_{i,1}, \ldots, d_{i,k}\}\big), \tag{5}$$

where T_i is the length of y_i. This formulation allows the model to implicitly learn both binary judgment and explanatory reasoning through a unified language modeling objective.

At evaluation time, the offline evaluator can follow the same workflow as the original proprietary models-based assessment pipeline, producing decisions and explanatory rationales in a manner consistent with the online model.

TABLE II
OVERALL BENCHMARKING RESULTS OF T2I MODELS ON UNIGENBENCH++USING ENGLISH SHORT PROMPTS. *Gemini2.5-pro* IS USED AS THE VLM FOR EVALUATION. BEST SCORES ARE IN **BOLD**, SECOND-BEST IN UNDERLINED.

English Short Prompt Evaluation											
Model	**Overall**	Style	World Know.	Attribute	Action	Relation.	Logic.Reason.	Grammar	Compound	Layout	Text
Closed-source Models											
HiDream-v2L	61.64	87.99	89.62	64.38	59.50	66.62	26.73	58.86	49.28	69.06	44.31
Stable-Image-Ultra	61.96	87.20	87.18	66.35	59.22	69.04	31.59	61.10	54.25	64.55	39.08
Recraft	62.63	87.20	90.19	68.16	60.55	62.56	29.55	63.64	44.85	57.84	61.78
Wan2.2-plus	64.82	91.10	87.34	70.19	68.00	73.03	42.05	66.53	61.37	74.77	13.83
Keling-Ketu	65.93	92.27	86.62	71.66	68.73	70.94	43.75	71.26	60.81	77.23	16.03
DALL-E-3	69.18	95.06	93.51	75.97	69.83	78.06	48.18	68.07	70.60	66.67	25.86
Runway-Gen4	69.75	93.44	90.36	74.03	70.21	72.56	49.31	70.08	67.76	76.33	33.43
FLUX-pro-1.1-Ultra	70.67	90.60	91.61	76.50	76.50	77.54	43.18	70.05	67.78	81.53	37.36
Imagen-3.0	71.85	89.25	94.75	77.33	81.46	82.86	48.36	69.84	71.71	81.34	21.55
FLUX-kontext-pro	75.84	94.78	91.61	79.20	77.66	79.34	55.68	72.69	72.68	84.47	50.29
Imagen-4.0-Fast	77.75	92.00	94.78	83.65	79.85	82.36	56.36	76.74	74.10	86.19	51.44
Wan2.5	78.17	93.15	95.22	81.06	74.23	82.23	56.36	73.59	76.23	77.61	71.97
Seedream-3.0	78.95	98.10	95.25	85.58	82.98	80.84	52.73	61.36	73.84	87.31	71.55
FLUX-kontext-max	80.00	96.59	94.19	80.93	77.38	85.08	61.36	78.53	78.99	85.04	61.92
Imagen-4.0	85.84	97.80	96.36	84.94	88.40	89.34	70.45	79.68	85.31	88.81	77.30
Seedream-4.0	87.35	98.80	95.41	88.57	85.65	87.69	67.73	78.88	86.08	90.67	**93.97**
Nano Banana	87.45	98.87	96.32	87.84	86.83	92.00	74.26	83.36	87.83	91.96	75.22
Imagen-4.0-Ultra	91.54	**99.20**	97.47	92.52	**92.20**	93.02	79.55	87.97	91.37	**93.10**	89.08
GPT-4o	**92.77**	98.57	**98.87**	**93.59**	90.79	**94.97**	**84.97**	**91.76**	**93.55**	91.35	89.24
Open-source Models											
SDXL	39.75	87.40	72.63	44.34	34.22	44.92	9.55	47.33	26.68	29.85	1.15
MMaDA	41.35	82.40	56.65	48.93	37.83	50.25	17.95	55.75	32.35	30.22	1.15
Kolors	45.47	84.40	77.22	54.17	48.00	52.79	19.77	46.66	33.63	42.91	1.15
Playground2.5	45.61	89.50	76.11	52.78	42.68	51.52	16.59	53.21	35.44	37.13	1.15
Emu3	46.02	86.80	77.06	51.39	40.11	49.75	19.32	52.94	36.86	44.78	1.15
Janus-flow	46.39	86.20	62.50	47.97	43.35	50.00	21.14	60.29	45.10	46.46	0.86
Janus	51.23	89.90	73.58	54.81	50.38	55.08	26.82	59.09	46.65	54.85	1.15
Hunyuan-DiT	51.38	94.10	80.70	62.71	49.05	59.64	24.55	55.48	41.62	44.78	1.15
X-Omni	53.77	72.70	76.27	60.04	54.47	56.60	29.09	59.09	41.75	62.69	25.00
CogView4	56.30	82.00	83.07	63.25	57.51	62.44	28.18	54.81	44.72	69.22	17.82
OneCAT	58.28	93.30	82.28	63.46	58.56	68.15	33.41	60.83	56.96	64.74	1.15
Infinity	59.81	90.80	87.97	68.06	60.17	69.16	31.36	60.16	51.42	66.60	12.36
BLIP3-o	59.87	92.80	80.22	63.89	63.97	66.50	39.55	68.58	53.74	68.47	1.15
SD-3.5-Medium	60.71	89.80	84.34	66.99	60.65	68.78	37.73	59.89	53.35	70.34	15.23
FLUX.1-dev	61.30	83.90	88.92	67.84	62.17	67.26	30.91	60.96	47.04	71.83	32.18
Bagel	61.53	90.20	85.60	67.74	61.98	70.69	30.23	66.44	58.12	76.49	7.76
Janus-Pro	61.61	90.80	86.71	67.74	64.26	68.40	37.05	64.44	62.11	72.01	2.59
Show-o2	62.73	87.20	86.08	70.51	69.58	70.18	40.91	61.63	64.69	75.37	1.15
SD-3.5-Large	62.99	88.60	88.92	68.59	62.17	69.80	32.27	58.96	58.76	69.03	32.76
OmniGen2	63.09	91.90	86.39	72.12	62.83	68.27	32.50	59.89	56.31	71.64	29.02
UniWorld-V1	63.11	91.10	82.91	70.62	67.21	67.13	38.41	63.77	54.51	69.03	26.44
BLIP3-o-Next	65.15	91.00	86.71	70.94	66.83	73.60	48.64	68.05	64.82	76.31	4.60
Echo-4o	69.12	92.20	90.51	79.06	68.92	76.52	44.77	75.13	71.78	82.28	10.06
FLUX.1-Krea-dev	69.88	88.70	92.56	75.96	71.01	73.98	39.77	63.37	64.43	84.14	44.83
Lumina-DiMOO	71.12	89.70	90.03	81.62	73.76	78.43	45.45	**70.45**	**73.32**	82.84	25.57
HiDream-I1-Full	71.81	92.50	94.15	72.97	73.00	75.38	41.14	63.24	62.63	78.17	64.94
Hunyuan-Image-2.1	74.64	90.88	92.06	79.66	77.81	77.54	46.59	62.83	64.82	84.14	70.11
Qwen-Image	**78.81**	**95.10**	**94.30**	**87.61**	**84.13**	**79.70**	**53.64**	60.29	**73.32**	**85.52**	**76.14**

TABLE III
OVERALL BENCHMARKING RESULTS OF T2I MODELS ON UNIGENBENCH++USING ENGLISH LONG PROMPTS. *Gemini2.5-pro* IS USED AS THE VLM FOR EVALUATION. BEST SCORES ARE IN **BOLD**, SECOND-BEST IN UNDERLINED.

English Long Prompt Evaluation											
Model	**Overall**	Style	World Know.	Attribute	Action	Relation.	Logic.Reason.	Grammar	Compound	Layout	Text
Closed-source Models											
Recraft	60.93	87.13	86.99	73.23	51.77	55.82	34.22	60.28	49.56	63.81	46.47
Stable-Image-Ultra	62.01	85.63	86.71	74.73	58.27	63.63	40.29	65.10	58.28	71.67	15.76
Runway-Gen4	68.29	91.72	88.82	79.83	64.30	69.53	48.28	70.55	68.57	73.79	27.47
Wan2.2-plus	68.76	90.28	87.57	81.08	66.49	72.79	55.58	70.18	71.73	79.13	12.77
DALL-E-3	70.82	95.08	92.71	84.98	68.36	77.90	57.11	68.19	73.88	71.76	18.26
FLUX-pro-1.1-Ultra	75.40	91.36	91.76	84.97	72.43	81.90	60.92	71.94	78.07	82.62	38.04
Imagen-3.0	75.76	92.41	94.19	86.32	75.81	80.76	61.25	77.96	78.70	86.06	24.18
FLUX-kontext-pro	78.58	94.83	93.60	86.24	74.44	78.40	66.26	77.05	79.75	85.46	49.73
FLUX-kontext-max	80.88	96.51	93.35	87.45	75.52	80.78	71.12	79.34	82.24	87.58	54.89
Seedream-3.0	80.99	97.18	93.79	91.90	79.94	83.41	62.62	75.13	81.03	88.41	56.52
Imagen-4.0-Fast	81.54	93.77	93.64	90.33	80.18	84.05	67.72	79.57	84.01	90.48	51.63
Wan2.5	84.34	96.75	95.52	91.40	77.55	86.96	71.32	78.06	85.60	87.18	73.10
Imagen-4.0	85.34	94.44	97.11	90.14	82.62	86.42	72.82	81.35	86.56	90.24	71.74
Nano Banana	88.82	98.83	95.78	93.06	83.93	**91.59**	81.27	89.33	90.63	**94.04**	69.75
Seedream-4.0	89.77	98.42	95.95	**95.06**	86.76	88.69	79.13	82.74	87.79	92.38	**90.76**
Imagen-4.0-Ultra	90.95	97.67	**98.26**	93.21	86.91	90.57	83.50	88.07	91.42	93.49	86.41
GPT-4o	**92.63**	**99.08**	97.95	93.53	**87.78**	91.13	**91.02**	**94.46**	**93.99**	93.59	83.79
Open-source Models											
MMaDA	40.10	75.83	52.75	49.90	32.42	39.06	19.42	50.00	38.37	43.02	0.27
SDXL	41.48	81.81	69.51	54.31	31.18	36.26	19.42	46.83	34.30	40.40	0.82
Emu3	50.95	89.36	76.16	66.81	43.80	51.70	27.43	50.25	46.00	56.67	1.36
Kolors	53.60	86.54	76.01	68.12	49.96	58.51	31.31	55.20	47.24	60.95	2.17
Janus-flow	54.80	88.70	65.90	63.60	48.68	58.24	41.75	63.83	55.16	60.48	1.63
Hunyuan-DiT	54.88	92.94	80.06	69.47	48.80	55.66	29.85	58.76	50.22	61.43	1.63
Janus	60.37	92.03	73.27	70.67	55.78	63.25	54.37	67.26	61.85	64.13	1.09
BLIP3-o	61.01	91.61	74.42	71.28	55.38	62.61	48.30	65.36	65.55	74.21	1.36
OneCAT	62.92	94.93	83.67	74.90	58.95	65.36	48.06	63.58	63.59	74.29	1.90
SD-3.5-Large	64.35	88.12	88.15	78.78	59.63	67.62	44.90	65.23	62.21	71.19	17.66
SD-3.5-Medium	64.67	92.19	86.56	80.24	58.59	69.88	45.87	65.86	62.86	73.25	11.41
X-Omni	67.00	80.15	82.37	79.82	61.96	64.28	51.70	68.78	64.17	73.33	43.48
Infinity	67.28	92.77	88.44	81.06	63.28	70.04	51.46	68.53	66.13	77.54	13.59
CogView4	67.68	88.29	89.45	80.57	64.33	66.97	49.76	71.70	66.86	79.84	19.02
FLUX.1-dev	69.42	89.29	89.45	79.90	64.54	69.40	54.37	70.56	68.46	77.54	30.71
UniWorld-V1	69.60	93.19	84.10	79.94	65.81	68.91	57.04	75.13	71.37	79.60	20.92
Show-o2	70.33	93.11	88.44	86.35	69.02	77.37	59.71	70.30	76.45	80.63	1.90
BLIP3-o-Next	71.03	94.60	88.87	80.57	70.18	74.68	65.53	76.02	74.27	80.71	4.89
Janus-Pro	71.11	94.02	88.15	81.81	69.14	77.96	62.62	74.62	76.53	82.14	4.08
Bagel	71.26	92.44	89.31	84.21	67.62	75.70	59.71	74.75	74.71	81.90	12.23
OmniGen2	71.39	94.35	84.83	83.03	66.57	73.06	56.55	76.40	70.49	80.63	27.99
Lumina-DiMOO	71.81	86.88	88.58	83.71	69.66	73.33	58.01	74.49	74.93	84.84	23.64
HiDream-I1-Full	74.25	93.11	92.63	83.49	68.82	74.30	50.24	72.59	69.77	79.92	57.61
Echo-4o	76.41	96.10	90.17	90.24	73.56	82.81	**69.42**	**82.36**	**84.88**	86.43	8.15
FLUX.1-Krea-dev	78.45	94.10	93.79	89.55	76.28	81.73	65.53	75.25	80.67	86.59	41.03
Hunyuan-Image-2.1	82.19	94.52	93.35	92.81	81.14	**85.13**	68.20	77.41	82.49	88.65	58.15
Qwen-Image	**83.94**	**96.93**	**95.09**	**93.65**	**81.86**	83.41	66.75	73.86	81.98	**88.97**	**76.90**

TABLE IV
OVERALL BENCHMARKING RESULTS OF T2I MODELS ON UNIGENBENCH++USING CHINESE SHORT PROMPTS. *Gemini2.5-pro* IS USED AS THE VLM FOR EVALUATION. BEST SCORES ARE IN **BOLD**, SECOND-BEST IN UNDERLINED.

Chinese Short Prompt Evaluation											
Model	**Overall**	Style	World Know.	Attribute	Action	Relation.	Logic.Reason.	Grammar	Compound	Layout	Text
Closed-source Models											
Runway-Gen4	54.93	64.75	71.05	60.43	60.42	65.90	42.03	58.38	61.00	64.71	0.59
Recraft	57.67	87.70	90.03	69.34	63.88	64.47	34.09	60.56	43.94	58.40	4.31
HiDream-v2L	59.95	89.34	91.02	67.87	64.90	72.67	32.01	62.57	53.19	64.77	1.16
Wan2.2-plus	66.96	91.06	84.39	73.93	72.52	76.78	51.82	70.59	64.77	71.83	11.92
DALL-E-3	67.93	95.90	93.04	78.42	72.24	79.95	51.59	71.52	72.94	62.50	1.15
Imagen-4.0-Fast	71.60	93.30	91.30	80.98	79.28	82.49	54.77	77.41	73.97	78.73	3.74
FLUX-kontext-max	71.85	96.38	92.83	76.41	78.59	83.97	56.48	75.68	75.13	81.34	1.72
Wan2.5	78.40	93.30	93.51	83.65	76.62	81.85	63.64	72.58	78.74	75.93	64.22
Imagen-4.0	79.52	97.50	96.84	86.22	90.40	90.74	73.18	82.89	85.70	89.18	2.59
Nano Banana	80.91	99.27	96.47	87.76	86.99	91.39	76.10	83.33	86.89	88.80	12.06
Seedream-3.0	81.68	97.50	93.99	88.03	86.98	84.39	59.09	67.25	76.68	84.14	78.74
Imagen-4.0-Ultra	83.21	98.90	97.94	90.71	**93.82**	92.13	79.32	87.43	89.95	**92.16**	9.77
Seedream-4.0	87.31	99.00	94.94	90.06	87.55	88.58	68.64	78.48	81.57	90.30	**93.97**
GPT-4o	**91.02**	**99.39**	**98.72**	**94.99**	92.34	**95.77**	**91.44**	**91.02**	**93.91**	89.27	63.37
Open-source Models											
UniWorld-V1	15.21	49.40	16.61	15.06	14.64	11.80	2.95	27.81	4.38	9.14	0.29
Janus-flow	20.93	58.50	18.67	19.23	22.05	19.54	10.68	35.03	10.70	14.93	0.00
Janus-Pro	30.83	75.60	39.08	33.12	26.33	32.74	10.23	36.63	24.48	30.04	0.00
Janus	30.98	78.10	27.85	30.88	31.37	30.58	13.41	48.40	17.53	31.72	0.00
Emu3	33.91	78.08	55.54	38.29	31.18	36.68	13.90	41.31	21.65	22.43	0.00
MMaDA	44.00	78.20	52.06	55.24	43.44	56.22	26.14	58.56	32.86	37.31	0.00
BLIP3-o-Next	44.48	74.60	50.00	55.98	47.62	53.55	27.50	54.14	26.55	54.85	0.00
HiDream-I1-Full	50.65	83.30	78.32	62.18	53.71	57.23	23.64	53.88	34.54	59.70	0.00
Hunyuan-DiT	53.36	92.50	84.97	62.93	57.22	59.39	29.55	54.68	44.59	47.76	0.00
X-Omni	53.69	70.07	71.52	63.85	58.37	59.77	34.77	56.28	41.75	59.51	20.98
CogView4	55.14	82.40	84.18	63.35	61.69	61.68	30.23	54.55	45.75	65.30	2.30
Lumina-DiMOO	58.35	80.90	69.46	75.64	61.12	67.13	39.09	64.84	56.06	69.22	0.00
Kolors	58.80	85.20	86.23	69.34	65.02	67.13	36.14	56.68	66.03	62.31	4.89
OneCAT	58.50	94.40	86.55	63.89	63.12	67.39	38.64	59.00	51.55	60.45	0.00
BLIP3-o	59.25	92.60	81.17	66.56	64.35	65.36	41.59	63.37	51.80	65.67	0.00
OmniGen2	63.20	93.00	86.39	75.43	66.54	70.69	44.09	65.64	59.92	69.96	0.29
Bagel	65.69	92.30	86.71	75.21	65.78	75.38	37.95	69.52	69.85	77.61	6.61
Echo-4o	72.40	92.80	87.66	84.29	76.05	82.23	56.82	**75.40**	**77.96**	83.02	7.76
Hunyuan-Image-2.1	77.76	92.20	90.51	84.19	80.51	82.74	50.23	61.50	70.62	**85.45**	79.60
Qwen-Image	**81.04**	**95.50**	**92.41**	**91.88**	**85.74**	**82.99**	**57.73**	62.83	76.16	82.65	**82.47**

IV. EXPERIMENT

A. Implementation Details

1) Benchmarking Models: **Closed-source Models.** GPT-4o [38], Imagen3.0/4.0-ultra/fast [18], Nano Banana [13], Seedream-3.0/4.0 [10], [11], Wan2.2-plus/2.5 [49], Runway-Gen4 [50], Recraft [51], DALL-E-3 [4], FLUX-Pro-Ultra/Kontext-Max [9], HiDream-v2L [52], and Stable-Image-Ultra [53]. **Open-source Models.** Qwen-Image [15], Hunyuan-Image-2.1 [54], Hidream-I1-Full [12], Lumina-DiMOO [42], Show-o2 [7], Infinity [20], OneCAT [47], CogView4 [46], X-Omni [55], MMaDA [56], Flux.1-dev [9], Flux.1-Krea-dev [19], Echo-4o [57], BLIP3-o series [24], UniWorld-V1 [58], OmniGen2 [8], Bagel [43], Hunyuan-DiT [59], Janus serise [21]–[23], Emu3 [60], Playground2.5 [61], Kolors [62], SDXL [5], and SD-3.5-Medium/Large [2].

2) Offline Evaluation Model: We use UnifiedReward-2.0-qwen-72b [63] as the base model and collect approximately 375K evaluation samples from Gemini-2.5-pro. Of this, 300K is used for model training, and 75K is reserved for evaluation.

B. Benchmarking Result Analysis

In this subsection, we will analyze the overall performance of current mainstream closed-source and open-source models on our **UNIGENBENCH++**, focusing on both Chinese and English, as well as long and short prompts.

1) English Short Prompt (Tab. II): **(a)** *Closed-source Models.* GPT-4o is the most well-rounded model, excelling in a broad range of metrics, including logical reasoning and grammar. Besides, Imagen-4.0-Ultra also performs well in visual generation accuracy but lags behind GPT-4o in logical reasoning. In contrast, remaining models like Seedream-3.0 and Wan-2.5 perform strongly in specific areas but struggle with

TABLE V
OVERALL BENCHMARKING RESULTS OF T2I MODELS ON UNIGENBENCH++USING CHINESE LONG PROMPTS. *Gemini2.5-pro* IS USED AS THE VLM FOR EVALUATION. BEST SCORES ARE IN **BOLD**, SECOND-BEST IN UNDERLINED.

Chinese Long Prompt Evaluation											
Model	**Overall**	Style	World Know.	Attribute	Action	Relation.	Logic.Reason.	Grammar	Compound	Layout	Text
Closed-source Models											
Recraft	56.90	86.38	85.55	74.31	54.65	57.44	36.17	57.49	50.00	64.52	2.45
Wan2.2-plus	70.05	91.61	88.73	82.42	70.22	73.65	57.04	70.05	71.51	80.08	15.22
DALL-E-3	71.16	95.85	94.36	85.41	70.59	80.12	61.41	70.81	75.87	73.33	3.80
Imagen-3.0	71.85	89.25	94.75	77.33	81.46	82.86	48.36	69.84	71.71	81.34	21.55
FLUX-kontext-max	75.24	97.59	92.31	86.17	75.71	81.27	68.20	78.77	80.16	87.58	4.62
Imagen-4.0	79.90	95.60	**97.98**	90.94	84.55	88.04	77.18	82.74	86.63	90.48	4.89
Nano Banana	83.17	98.41	97.38	93.29	85.55	91.32	82.40	88.35	91.21	93.15	10.68
Imagen-4.0-Ultra	83.86	97.34	97.40	93.59	88.80	92.35	86.89	88.83	92.51	94.13	6.79
Wan2.5	84.24	98.00	94.30	90.49	78.39	86.64	74.51	80.08	85.13	88.54	66.30
Seedream-3.0	86.14	98.42	95.36	93.93	84.53	87.55	68.45	77.54	83.11	90.16	82.34
Seedream-4.0	90.35	98.42	96.39	**95.54**	89.29	88.69	80.58	83.63	87.72	91.90	**91.30**
GPT-4o	**90.51**	**99.41**	97.96	94.72	**89.33**	**92.59**	**90.05**	**94.11**	**94.59**	**95.21**	57.14
Open-source Models											
UniWorld-V1	21.50	55.48	17.34	27.50	19.34	19.34	8.98	28.68	12.50	24.44	1.36
Janus-flow	23.01	57.39	17.49	23.42	19.46	20.04	17.48	32.23	21.58	21.59	0.27
Janus	33.63	75.00	30.06	35.98	29.74	28.23	20.15	44.04	31.47	40.56	1.09
Emu3	35.95	75.08	53.03	48.82	27.81	32.06	19.66	38.32	28.49	35.40	0.82
MMaDA	50.61	84.05	63.58	61.31	42.98	52.69	31.80	58.76	50.07	60.63	0.27
HiDream-I1-Full	50.70	83.06	78.61	65.05	47.47	49.25	24.27	53.81	42.08	60.40	2.99
BLIP3-o-Next	54.55	87.71	61.85	63.75	51.81	57.76	41.50	60.66	54.00	64.60	1.90
Hunyuan-DiT	55.57	94.10	76.16	69.72	51.04	55.60	33.98	60.06	52.03	61.67	1.36
BLIP3-o	59.25	89.70	77.17	69.24	55.98	60.56	47.09	60.91	60.68	69.29	1.90
Janus-Pro	60.21	91.28	75.87	65.79	54.33	62.61	49.27	68.53	65.62	66.59	2.17
X-Omni	62.18	76.91	74.13	76.51	58.43	60.83	46.60	64.85	61.12	73.02	29.35
Lumina-DiMOO	63.80	84.30	76.45	79.41	61.32	66.70	49.27	71.95	68.90	78.33	1.36
OneCAT	63.88	95.85	85.26	74.79	60.11	65.03	54.37	63.07	62.35	75.79	2.17
Kolors	65.12	90.61	87.14	81.18	64.49	71.23	47.82	63.96	64.17	74.60	5.98
CogView4	68.09	89.62	89.31	80.99	67.94	70.58	51.94	70.94	69.91	81.51	8.15
OmniGen2	70.75	95.35	87.57	85.05	67.17	75.38	62.62	77.03	74.06	81.35	1.90
Bagel	75.75	96.10	89.02	88.25	72.43	81.52	68.69	81.09	82.05	83.97	14.40
Echo-4o	78.31	96.26	91.18	91.82	75.56	85.83	**72.57**	**83.50**	85.25	88.10	13.04
Qwen-Image	86.91	**97.84**	**95.66**	**95.04**	**86.56**	87.61	69.90	76.90	82.99	90.48	86.14
Hunyuan-Image-2.1	**87.01**	95.18	94.08	93.82	83.99	**88.09**	71.36	80.08	**85.61**	**91.43**	**86.41**

logical reasoning and relational understanding. For example, Seedream-3.0 excels in stylistic quality and world knowledge but falls short in complex reasoning and grammar understanding tasks. These results highlight a clear trend where many closed-source models have specialized in most visual generation tasks but still fall short in handling complex reasoning and understanding. **(b)** *Open-source Models.* Qwen-Image stands out as the top performer among open-source models, excelling in generating semantical accurate and contextually relevant images based on English short text descriptions. Besides, HiDream-I1-Full excels in world knowledge, but falls short in attribute generation and logical reasoning. Notably, Lumina-DiMOO performs strongly in relation generation and grammar understanding, but struggles with text generation consistency. The remaining models show promising results in specific areas but exhibit weaknesses in others. For example, Echo-4o and BLIP3-o-Next excel in compound semantic generation and grammar understanding but struggle with complex relationships and logically consistent scenes. **(c)** *Closed- v.s. Open-source Models.* A clear trend emerges where some open-source models are making significant strides in catching up to their closed-source counterparts. To be precise, Qwen-Image, the leading open-source model, surpasses many closed-source models in key areas such as world knowledge, action generation, and logical reasoning. It competes closely with top performers like Seedream-3.0 and FLUX-kontext-max. However, despite the impressive progress, closed-source models still hold a significant advantage in several areas. For instance, Seedream-4.0 and Nano Banana outperform all open-source models in dimensions like style, grammar, and compound feature construction. Overall, while open-source models are making remarkable progress, particularly in the world knowledge and attribute generation domains, closed-source models remain dominant in grammar, logical reasoning, and relation generation.

*2) **English Long Prompt (Tab. III):*** **(a)** *Closed-source Models.* For English long prompt generation, the closed-source models exhibit strong performance across most evaluation metrics. GPT-4o stands out with the highest overall score,

leading in grammar, compound generation, and logical reasoning, though its layout consistency and text generation slightly lag behind several models. Seedream-4.0 and Nano Banana are also notable performers, with Seedream-4.0 achieving exceptional scores in text while Nano Banana shines in style consistency and relation generation. Remaining models like Imagen-4.0 and Wan2.5 offer promising results in specific areas such as attribute and layout generation, but still trail behind in grammar understanding. **(b)** *Open-source Models* show significant progress: Qwen-Image still leads the open-source group with strong results across world knowledge, action, text, and compound generation, but still challenges in grammar understanding. Models such as Hunyuan-Image-2.1 and FLUX-Krea-dev also perform well in relational understanding and logical reasoning. Lumina-DiMOO and OmniGen2 provide solid performance but are weaker in logical reasoning and text generation. **(c)** *Closed- v.s. Open-source Models.* Most closed-source models consistently outperform in areas such as world knowledge, logical reasoning, layout consistency, and text fluency. Open-source models, while showing significant progress, still fall behind in these areas. Hunyuan-Image-2.1 and Qwen-Image are the strongest open-source contenders, achieving competitive results in text generation, style and world knowledge, but lack the relation and grammar understanding seen in closed-source models like GPT-4o and Seedream-4.0.

3) ***Chinese Short Prompt (Tab. IV):*** **(a)** *Closed-source Models.* When using Chinese short prompt evaluation, although GPT-4o leads in most evaluation metrics, it still has room for improvement, particularly in layout generation and the accuracy of Chinese text generation. In contrast, Seedream-4.0 excels in text generation and also performs strongly in style and attribute generation. Besides, Imagen-4.0-Ultra performs strongly, particularly in action generation and logical reasoning. It also achieves high scores in layout consistency but slightly trails GPT-4o in overall performance. Other closed-source models, such as Seedream-3.0, FLUX-kontext-max, and Nano Banana, show promise in areas like style generation and world knowledge but struggle in more complex tasks like logical reasoning and layout generation. **(b)** *Open-source Models.* Qwen-Image and Hunyuan-Image-2.1 stand out as the top-performing open-source models, excelling in relation, action, and attribute generation, though they still face challenges in grammar understanding. In contrast, Echo-4o performs well in grammar and compound tasks, but struggles with Chinese text generation compared to the top models. Models like OmniGen2 and Bagel show balanced performance across multiple metrics, but face limitations in layout generation and text consistency. Specifically, OmniGen2 excels in world knowledge, while Bagel is solid in style and action prediction, but neither matches the best models in complex reasoning or text generation. The remaining models, such as X-Omni, Kolors, show promise in certain areas but generally fall behind in grammar understanding, text, and compound context generation. **(c)** *Closed- v.s. Open-source Models.* Closed-source models, particularly GPT-4o, Seedream-4.0, and Imagen-4.0, dominate the evaluation, excelling in overall performance. In comparison, open-source models such as Qwen-Image and Hunyuan-Image-2.1 also show significant progress, especially in world knowledge and text generation. However, they generally lag behind in grammar understanding, compound generation, and complex logical reasoning tasks.

4) ***Chinese Long Prompt (Tab. V):*** **(a)** *Closed-source Models* Closed-source models still demonstrate strong overall performance in generating Chinese long prompts. Notably, Seedream-4.0 performs exceptionally well in Chinese text generation and attribute generation, achieving an overall performance very close to GPT-4o. Meanwhile, Imagen-4.0-Ultra excels in layout consistency, grammar, relational understanding, compound feature generation, and logical reasoning but trails in world knowledge and fluency. In addition, Wan2.5 demonstrates highly balanced capabilities. While it does not excel in any particular dimension, its overall score remains relatively high. **(b)** *Open-source Models.* Hunyuan-Image-2.1 leading the group, excelling in tasks like layout and text generation. Qwen-Image competes closely with Seedream-4.0 in attribute and layout generation, though it still lags behind in grammar understanding and logical reasoning. Other models like Echo-4o and Bagel perform well in relational understanding and world knowledge but face challenges in handling complex action generation and accuracy Chinese text generation. **(c)** *Closed- v.s. Open-source Models.* Closed-source models outperform in grammar understanding and generating logically consistent images, while open-source models are making significant strides, particularly in world knowledge, attribute generation, and text generation. However, open-source models still need further improvements in handling compound and action generation. Most closed-source and open-source models also have room for improvement in logical reasoning.

Detailed 27 dimensions benchmarking results are provided in Tabs. VII, VIII, IX, and X.

C. Offline Evaluation Model

Existing benchmarks [36], [40] typically use Vision-Language Models (VLMs) like Qwen2.5-VL-72b [64] for offline generalization evaluation. However, compared to closed-source models, the evaluation accuracy of these models often falls short. Specifically, in our benchmark, we observed that Qwen2.5-VL-72b performs reasonably well on relatively simple dimensions such as attribute-color and facial expressions. However, its performance becomes unreliable on more complex dimensions. To address this, we trained a dedicated evaluation model, and the results, compared to Qwen2.5-VL-72b, are shown in Fig. 5. As demonstrated, our model significantly outperforms Qwen2.5-VL-72b across both short and long, as well as Chinese and English prompts evaluations, highlighting a substantial improvement in evaluation accuracy. Both English and Chinese qualitative evaluation results are provided in Fig. 4 (c).

D. Compared with UniGenBench

Compared with the preliminary version [29], this work introduces several significant extensions across the following aspects: (1) **Bilingual and length-variant prompt support**: The prompts are expanded to include varying lengths, as well as both English and Chinese languages, thereby enhancing

the diversity and comprehensiveness of the benchmark. This extension allows for a more in-depth evaluation of T2I model sensitivity and robustness to prompt length and language variations; (2) **Dedicated offline evaluation model**: Due to the inconvenience of accessing closed-source proprietary models via APIs, we provide a dedicated offline evaluation model that enables reliable assessments of T2I model outputs, offering enhanced flexibility and ease of use for the research community; (3) **More comprehensive benchmarking results and detailed analysis**: We extensively tested a wide range of both open-source and closed-source models on English and Chinese prompts of varying lengths. Through thorough comparative analyses, we further identify their strengths and weaknesses, providing a deeper understanding of model performance across a broader set of test points and real-world scenarios.

V. Conclusion

In this work, we introduce **UniGenBench++**, a unified semantic benchmark for evaluating text-to-image (T2I) models. It consists of 600 prompts organized within a hierarchical structure that ensures both coverage and efficiency. Specifically, it covers 5 main themes and 20 subthemes across diverse real-world scenarios, assessing models on 10 primary and 27 sub-evaluation criteria using English and Chinese prompts in both short and long forms. Leveraging the world knowledge and fine-grained image understanding capabilities of the Visual Language Model (VLM), we developed an effective pipeline for benchmark construction and model evaluation. Additionally, to facilitate community usage, we propose a robust offline evaluation model for T2I model assessments. Our comprehensive benchmarking reveals the strengths and weaknesses of both open- and closed-source T2I models, offering valuable insights into their semantic consistency and performance across various aspects.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[3] X. Liu, C. Gong, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," *arXiv preprint arXiv:2209.03003*, 2022.

[4] OpenAI., "Dall·e 3," *https://openai.com/zh-Hans-CN/index/dall-e-3/*, 2023.

[5] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[6] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *ICML*, 2024.

[7] J. Xie, Z. Yang, and M. Z. Shou, "Show-o2: Improved native unified multimodal models," *arXiv preprint arXiv:2506.15564*, 2025.

[8] C. Wu, P. Zheng, R. Yan, S. Xiao, X. Luo, Y. Wang, W. Li, X. Jiang, Y. Liu, J. Zhou *et al.*, "Omnigen2: Exploration to advanced multimodal generation," *arXiv preprint arXiv:2506.18871*, 2025.

[9] B. F. Labs, "Flux," *https://github.com/black-forest-labs/flux*, 2024.

[10] Y. Gao, L. Gong, Q. Guo, X. Hou, Z. Lai, F. Li, L. Li, X. Lian, C. Liao, L. Liu *et al.*, "Seedream 3.0 technical report," *arXiv preprint arXiv:2504.11346*, 2025.

[11] T. Seedream, Y. Chen, Y. Gao, L. Gong, M. Guo, Q. Guo, Z. Guo, X. Hou, W. Huang, Y. Huang *et al.*, "Seedream 4.0: Toward next-generation multimodal image generation," *arXiv preprint arXiv:2509.20427*, 2025.

[12] Q. Cai, J. Chen, Y. Chen, Y. Li, F. Long, Y. Pan, Z. Qiu, Y. Zhang, F. Gao, P. Xu *et al.*, "Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer," *arXiv preprint arXiv:2505.22705*, 2025.

[13] Google, "Nano banana," *https://deepmind.google/models/gemini/image/*, 2025.

[14] OpenAI, "Gpt-image-1," *https://openai.com/index/introducing-4o-image-generation/*, 2025.

[15] C. Wu, J. Li, J. Zhou, J. Lin, K. Gao, K. Yan, S.-m. Yin, S. Bai, X. Xu, Y. Chen *et al.*, "Qwen-image technical report," *arXiv preprint arXiv:2508.02324*, 2025.

[16] D. Li, A. Kamko, E. Akhgari, A. Sabet, L. Xu, and S. Doshi, "Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation," *arXiv preprint arXiv:2402.17245*, 2024.

[17] Y. Wang, W. Zhang, X. Honghui, and C. Jin, "High fidelity scene text synthesis," in *CVPR*, 2025.

[18] Google, "Imagen," *https://deepmind.google/models/imagen/*, 2025.

[19] B. F. Labs., "Flux.1 krea," *https://www.krea.ai/apps/image/flux-krea*, 2025.

[20] J. Han, J. Liu, Y. Jiang, B. Yan, Y. Zhang, Z. Yuan, B. Peng, and X. Liu, "Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis," in *CVPR*, 2025, pp. 15 733–15 744.

[21] C. Wu, X. Chen, Z. Wu, Y. Ma, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, C. Ruan *et al.*, "Janus: Decoupling visual encoding for unified multimodal understanding and generation," in *CVPR*, 2025, pp. 12 966–12 977.

[22] X. Chen, Z. Wu, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, and C. Ruan, "Janus-pro: Unified multimodal understanding and generation with data and model scaling," *arXiv preprint arXiv:2501.17811*, 2025.

[23] Y. Ma, X. Liu, X. Chen, W. Liu, C. Wu, Z. Wu, Z. Pan, Z. Xie, H. Zhang, X. yu, L. Zhao, Y. Wang, J. Liu, and C. Ruan, "Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation," 2024.

[24] J. Chen, Z. Xu, X. Pan, Y. Hu, C. Qin, T. Goldstein, L. Huang, T. Zhou, S. Xie, S. Savarese *et al.*, "Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset," *arXiv preprint arXiv:2505.09568*, 2025.

[25] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *NeurIPS*, vol. 36, pp. 53 728–53 741, 2023.

[26] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[27] Y. Wang, Z. Li, Y. Zang, C. Wang, Q. Lu, C. Jin, and J. Wang, "Unified multimodal chain-of-thought reward model through reinforcement fine-tuning," *arXiv preprint arXiv:2505.03318*, 2025.

[28] Y. Wang, Z. Tan, J. Wang, X. Yang, C. Jin, and H. Li, "Lift: Leveraging human feedback for text-to-video model alignment." *arXiv preprint arXiv:2412.04814*, 2024.

[29] Y. Wang, Z. Li, Y. Zang, Y. Zhou, J. Bu, C. Wang, Q. Lu, C. Jin, and J. Wang, "Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning," *arXiv preprint arXiv:2508.20751*, 2025.

[30] C. Tong, Z. Guo, R. Zhang, W. Shan, X. Wei, Z. Xing, H. Li, and P.-A. Heng, "Delving into rl for image generation with cot: A study on dpo vs. grpo," *arXiv preprint arXiv:2505.17017*, 2025.

[31] J. Liu, G. Liu, J. Liang, Y. Li, J. Liu, X. Wang, P. Wan, D. Zhang, and W. Ouyang, "Flow-grpo: Training flow matching models via online rl," *arXiv preprint arXiv:2505.05470*, 2025.

[32] Z. Xue, J. Wu, Y. Gao, F. Kong, L. Zhu, M. Chen, Z. Liu, W. Liu, Q. Guo, W. Huang *et al.*, "Dancegrpo: Unleashing grpo on visual generation," *arXiv preprint arXiv:2505.07818*, 2025.

[33] Y. Zhou, P. Ling, J. Bu, Y. Wang, Y. Zang, J. Wang, L. Niu, and G. Zhai, "Ggrpo: Granular grpo for precise reward in flow models," *arXiv preprint arXiv:2510.01982*, 2025.

[34] D. Ghosh, H. Hajishirzi, and L. Schmidt, "Geneval: An object-focused framework for evaluating text-to-image alignment," *NIPS*, vol. 36, pp. 52 132–52 152, 2023.

[35] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu, "T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation," *NIPS*, vol. 36, pp. 78 723–78 747, 2023.

[36] Y. Niu, M. Ning, M. Zheng, W. Jin, B. Lin, P. Jin, J. Liao, C. Feng, K. Ning, B. Zhu *et al.*, "Wise: A world knowledge-informed semantic evaluation for text-to-image generation," *arXiv preprint arXiv:2503.07265*, 2025.

[37] K. Sun, R. Fang, C. Duan, X. Liu, and X. Liu, "T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation," *arXiv preprint arXiv:2508.17472*, 2025.

[38] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[39] X. Hu, R. Wang, Y. Fang, B. Fu, P. Cheng, and G. Yu, "Ella: Equip diffusion models with llm for enhanced semantic alignment," *arXiv preprint arXiv:2403.05135*, 2024.

[40] X. Wei, J. Zhang, Z. Wang, H. Wei, Z. Guo, and L. Zhang, "Tiif-bench: How does your t2i model follow your instructions?" *arXiv preprint arXiv:2506.02161*, 2025.

[41] Google, "Gemini2.5-pro," *https://deepmind.google/models/gemini/pro/*, 2025.

[42] Y. Xin, Q. Qin, S. Luo, K. Zhu, J. Yan, Y. Tai, J. Lei, Y. Cao, K. Wang, Y. Wang *et al.*, "Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding," *arXiv preprint arXiv:2510.06308*, 2025.

[43] C. Deng, D. Zhu, K. Li, C. Gou, F. Li, Z. Wang, S. Zhong, W. Yu, X. Nie, Z. Song *et al.*, "Emerging properties in unified multimodal pretraining," *arXiv preprint arXiv:2505.14683*, 2025.

[44] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[45] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[46] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang *et al.*, "Cogview: Mastering text-to-image generation via transformers," *Advances in neural information processing systems*, vol. 34, pp. 19 822–19 835, 2021.

[47] H. Li, X. Peng, Y. Wang, Z. Peng, X. Chen, R. Weng, J. Wang, X. Cai, W. Dai, and H. Xiong, "Onecat: Decoder-only auto-regressive model for unified understanding and generation," *arXiv preprint arXiv:2509.03498*, 2025.

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

[49] A. Cloud, "Wan-t2i," *https://www.alibabacloud.com/help/en/model-studio/text-to-image-v2-api-reference*, 2025.

[50] Runway, "Runway-gen4," *https://docs.dev.runwayml.com*, 2025.

[51] Recraft, "Recraft," *https://www.recraft.ai*, 2025.

[52] Hidream, "Hidream-v2l," *https://hidreamai.com/studio*, 2025.

[53] Stability, "Stable image ultra," *https://platform.stability.ai/*, 2025.

[54] Tencent, "Hunyuan-image-2.1," *https://github.com/Tencent-Hunyuan/HunyuanImage-2.1*, 2025.

[55] Z. Geng, Y. Wang, Y. Ma, and et. al, "X-omni: Reinforcement learning makes discrete autoregressive image generative models great again," *arXiv preprint arXiv:2507.22058*, 2025.

[56] L. Yang, Y. Tian, B. Li, X. Zhang, K. Shen, Y. Tong, and M. Wang, "Mmada: Multimodal large diffusion language models," *arXiv preprint arXiv:2505.15809*, 2025.

[57] J. Ye, D. Jiang, Z. Wang, L. Zhu, and et. al, "Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation," *https://arxiv.org/abs/2508.09987*, 2025.

[58] B. Lin, Z. Li, X. Cheng, Y. Niu, Y. Ye, X. He, S. Yuan, W. Yu, S. Wang, Y. Ge *et al.*, "Uniworld: High-resolution semantic encoders for unified visual understanding and generation," *arXiv preprint arXiv:2506.03147*, 2025.

[59] Z. Li, J. Zhang, Q. Lin, J. Xiong, Y. Long, and et. al, "Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding," 2024.

[60] X. Wang, X. Zhang, Z. Luo, Q. Sun, Y. Cui, J. Wang, F. Zhang, Y. Wang, Z. Li, Q. Yu *et al.*, "Emu3: Next-token prediction is all you need," *arXiv preprint arXiv:2409.18869*, 2024.

[61] D. Li, A. Kamko, E. Akhgari, A. Sabet, L. Xu, and S. Doshi, "Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation," *arXiv preprint arXiv:2402.17245*, 2024.

[62] K. Team, "Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis," *arXiv preprint*, 2024.

[63] Y. Wang, Y. Zang, H. Li, C. Jin, and J. Wang, "Unified reward model for multimodal understanding and generation," *arXiv preprint arXiv:2503.05236*, 2025.

[64] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

TABLE VI
OVERVIEW OF PROMPT THEMES. WE PROVIDE AN EXAMPLE PROMPT FOR EACH OF THE PROMPT THEMES TO ILLUSTRATE THE SCOPE AND DIVERSITY OF GENERATION SCENARIOS IN OUR BENCHMARK.

Prompt Themes	Sub-Themes	Example Prompt
Creative Divergence	Imaginative	*"An astronaut rides a dragon made of star dust, shuttling through the rings of Saturn. The picture presents a magnificent oil painting texture."*
	Others	*"In the ink painting style, a lonely swordsman stood on the edge of a cliff, facing the strong wind. His face had no expression, but his eyes were filled with endless sadness."*
Art	Graphic Art	*"Please generate a graphic art poster: On the left side of the picture is a towering city silhouette, on the right side is a peaceful forest, and on the top is the text 'We build the future and cherish the green earth'."*
	Photography	*"A golden Labrador retriever is leaping excitedly on the green grass, chasing a soap bubble that glows with a rainbow in the sun, National Geographic photography style."*
	Sculpture	*"A giant elephant sculpture carved from transparent crystal is crystal clear and stands quietly in the center of the museum."*
	Others	*"Please generate a painting: an ancient magic hourglass is being turned upside down. Due to the passage of time, a line of English words appears on the stone platform below it: ' Time reveals all hidden truths and lies'."*
Illustration	Copywriting Illustration	*"A little fox successfully built a cabin. It looked proudly at its masterpiece. The wooden sign next to it read in English: 'The future belongs to those who build it today'."*
	Content Illustration	*"There was an open retro wooden jewelry box with an exquisite sapphire necklace lying quietly inside, shining with a glimmer."*
Film & Story	Realistic	*"The texture of the movie. An elderly historian wearing white cotton gloves carefully examined a yellowed sheepskin scroll map with a magnifying glass, with a solemn expression."*
	Science Fiction	*"An astronaut wearing a spacesuit holds a pyramidal holographic projector in his hand, projecting an image of the earth."*
	Animation	*"Pixar animation style, a clumsy young wizard whose robe is emitting colorful smoke due to a failed spell, and he himself has a panicked expression."*
Design	Ad / E-commerce Design	*"Please generate an advertisement for a fashionable assault coat: A young man is standing in the heavy rain, but he does not have an umbrella, but his clothes and hair are not wet at all, and his face shows a confident smile."*
	Spatial Design	*"A modern library that incorporates elements of the Forbidden City. Its dome is a golden caisson structure, presenting a grand new Chinese style as a whole."*
	Game Design	*"The game character design shows a mechanical wolf whose body is joined by multiple sharp triangles. The joints exude blue light and have a low polygonal style."*
	UI Design	*"Design the UI interface of a pet health App with a cat. Because of its high health index, this kitten is happily wagging its tail. The overall is a flat illustration style."*
	Poster Design	*"Advertising posters, two bottles of anthropomorphic juice drinks, one bottle of orange juice and one bottle of apple juice, they wore swimsuits of similar styles but different colors, lying side by side on beach chairs."*
	IP Design	*"A cute anthropomorphic alarm clock IP, with a line of words "Every second is a brand new start" engraved on the bell above its head, is running happily."*
	Logo / Icon Design	*"A logo design has two similar mechanical phoenixes symmetrical left and right, with the same metallic texture in the middle."*
	Fashion Design	*"A model with long-chestnut hair wore a beige linen suit consisting of a long-sleeved top and wide-leg pants, with a pen stained with blue ink inserted in the chest pocket of the top."*
	Design Resources	*"A huge blue gear and a much smaller red gear mesh with each other, and the latter drives it to rotate slowly, in a flat illustration style."*

TABLE VII
DETAILED BENCHMARKING RESULTS OF T2I MODELS ON UNIGENBENCH++USING ENGLISH SHORT PROMPTS. *Gemini2.5-pro* IS USED AS THE VLM FOR EVALUATION. BEST SCORES ARE IN **BOLD**, SECOND-BEST IN UNDERLINED.

English Short Prompt Evaluation																													
Models	Overall	Style	World Know.	Attribute						Action						Relationship				Compound		Grammar			Layout		Logic. Reason.	Text
				Quant.	Express.	Materi.	Size	Shape	Color	Hand	Full Body	Animal	Non Contact	Contact	State	Compos.	Sim.	Inclus.	Compare.	Imagin.	Feat Match.	Pron Ref.	Consist.	Neg.	2D	3D		
Closed-source Models																													
HiDream-v2L	61.64	87.99	89.62	65.71	44.87	57.82	74.26	59.87	94.92	51.28	58.56	67.65	61.98	51.52	65.09	71.23	64.20	65.93	60.32	53.75	44.76	72.35	60.00	44.23	70.41	67.68	26.73	44.31
Stable-Image-Ultra	61.96	87.20	87.18	67.36	48.08	64.15	69.44	64.38	91.67	55.77	58.15	63.24	61.22	51.79	64.15	72.64	66.67	70.11	62.50	60.97	47.40	78.68	58.33	45.00	67.28	61.74	31.59	39.08
Recraft	62.63	87.20	90.19	68.06	56.41	70.75	65.97	57.50	95.83	50.00	70.65	76.47	55.61	48.81	63.21	64.53	59.44	59.24	67.19	43.37	46.35	73.16	58.33	58.08	58.82	56.82	29.55	61.78
Wan2.2-plus	64.82	91.10	87.34	76.39	55.77	66.51	71.53	64.38	94.17	58.33	75.82	69.12	68.88	57.74	75.00	70.27	67.98	77.72	76.69	66.92	55.73	73.90	56.74	66.92	77.49	71.97	42.05	13.83
Keling-Ketu	65.93	92.27	86.62	75.00	56.41	78.77	79.17	53.12	91.38	54.49	76.09	72.79	69.90	58.93	76.89	68.92	70.56	74.46	71.09	66.24	55.26	77.21	67.59	68.08	80.97	73.36	43.75	16.03
DALL-E-3	69.18	95.06	93.51	62.14	59.87	87.74	87.50	65.00	92.50	60.90	75.00	76.47	66.84	63.41	75.47	82.43	69.44	87.78	66.41	76.79	64.21	74.24	74.07	56.64	57.72	76.17	48.18	25.86
Runway-Gen4	69.75	93.44	90.36	72.86	51.97	89.42	68.06	65.62	95.00	62.18	79.35	82.35	66.15	60.37	71.70	74.32	62.22	77.84	75.78	71.65	63.71	71.21	67.59	71.03	77.61	75.00	49.31	33.43
FLUX-pro-1.1-Ultra	70.67	90.60	91.61	75.69	59.62	78.77	77.78	74.38	96.67	57.69	68.48	77.21	76.53	64.29	76.89	80.41	72.78	82.07	71.09	74.74	60.68	84.56	68.98	55.77	80.15	82.95	43.18	37.36
Imagen-3.0	71.85	89.25	94.75	75.78	64.67	80.66	82.84	70.00	93.10	80.00	83.89	85.29	77.37	74.40	87.38	83.90	73.33	88.64	83.90	79.23	64.06	79.04	70.75	59.13	82.72	79.92	48.36	21.55
FLUX-kontext-pro	75.84	94.78	91.61	75.00	71.62	76.89	84.72	74.38	97.50	75.00	79.35	80.88	71.94	73.21	84.91	81.42	75.56	83.33	74.22	75.00	70.31	84.23	76.85	57.69	85.98	82.95	55.68	50.29
Imagen-4.0-Fast	77.75	92.00	94.78	77.08	75.00	85.85	89.58	78.75	98.33	73.72	84.24	81.62	76.53	76.79	84.91	83.45	73.89	89.13	82.03	80.10	67.97	86.03	75.00	68.46	88.24	84.09	56.36	51.44
Wan2.5-t2i-preview	78.17	93.15	95.22	75.00	67.95	91.04	85.29	77.50	87.50	61.18	75.00	76.47	75.00	72.02	82.55	85.14	75.00	82.07	85.94	79.38	73.04	84.07	73.15	63.08	75.74	79.55	56.36	71.97
Seedream-3.0	78.95	98.10	95.25	80.56	82.05	90.57	85.42	78.12	97.50	75.00	89.67	85.29	75.51	80.95	90.09	82.77	73.89	84.24	81.25	78.57	69.01	79.78	69.91	35.00	86.76	87.88	52.73	71.55
FLUX-kontext-max	80.00	96.59	94.19	75.69	74.32	82.55	86.81	74.38	94.17	67.95	83.15	77.94	77.04	70.83	84.43	87.50	78.89	90.00	81.25	83.93	73.96	84.23	78.70	72.69	86.74	88.33	61.36	61.92
Imagen-4.0	85.84	97.80	96.36	84.03	76.92	90.57	89.58	71.88	98.33	86.54	**94.02**	88.97	85.71	83.33	91.04	93.58	78.89	95.11	85.94	90.31	80.21	86.76	77.31	74.23	88.24	89.39	70.45	77.30
Seedream-4.0	87.35	98.80	95.41	86.81	85.90	**97.17**	84.03	76.88	**100.00**	77.56	87.50	88.24	80.10	83.93	94.81	88.18	80.56	94.02	87.50	88.27	83.85	84.93	79.17	72.31	90.81	90.53	67.73	93.97
Nano Banana	87.45	98.87	96.32	85.00	83.33	88.50	95.74	78.21	99.17	82.05	93.41	86.03	82.47	83.33	91.98	94.76	86.52	91.26	**94.53**	89.66	86.02	90.71	82.08	76.59	**92.65**	91.25	74.26	75.22
Imagen-4.0-Ultra	91.54	**99.20**	97.47	**93.06**	81.41	94.34	**95.83**	91.88	**100.00**	**90.38**	93.44	**91.91**	90.31	89.29	**96.70**	95.27	84.44	**98.37**	92.19	92.86	89.84	**94.12**	87.04	82.31	**92.65**	**93.56**	79.55	89.08
GPT-4o	**92.77**	98.57	**98.87**	90.00	**94.70**	94.20	91.61	**92.50**	99.17	89.74	92.22	87.12	**90.43**	**89.82**	93.75	**96.23**	**95.00**	94.89	92.19	**95.64**	**91.40**	92.91	**91.67**	**90.57**	91.04	91.67	**84.97**	**89.24**
Open-source Models																													
SDXL	39.75	87.40	72.63	44.44	25.00	52.83	44.44	33.75	68.33	19.23	35.33	43.38	26.53	24.40	53.30	53.72	38.33	39.67	41.41	33.93	19.27	50.37	42.59	48.08	26.47	33.33	9.55	1.15
MMaDA	41.35	82.40	56.65	45.83	29.49	54.25	49.31	44.38	74.17	15.38	40.22	52.94	33.16	25.60	56.60	55.07	57.22	47.28	33.59	40.56	23.96	59.19	40.28	65.00	30.15	30.30	17.95	1.15
Kolors	45.47	84.40	77.22	62.50	33.33	51.89	62.50	40.62	83.33	42.95	42.39	56.62	45.92	39.88	59.43	55.41	53.89	51.63	46.88	41.33	25.78	56.62	47.22	35.77	43.01	42.80	19.77	1.15
Playground2.5	45.61	89.50	76.11	58.33	43.59	57.08	44.44	41.25	75.83	28.85	50.00	52.21	35.20	29.17	58.02	60.14	49.44	48.37	39.06	43.88	26.82	58.82	50.00	50.00	34.56	39.77	16.59	1.15
Emu3	46.02	86.80	77.06	44.44	45.51	53.77	43.06	46.25	80.00	25.00	47.28	50.74	35.20	27.98	52.36	56.76	46.67	48.37	39.84	41.33	32.29	59.56	53.70	45.38	45.22	44.32	19.32	1.15
Janus-flow	46.39	86.20	62.50	43.06	30.77	55.19	55.56	30.00	78.33	23.08	48.37	58.82	36.73	36.31	55.66	59.80	38.89	51.63	40.62	57.65	32.29	66.18	48.61	63.85	49.26	43.56	21.14	0.86
Janus	51.23	89.90	73.58	37.50	37.82	58.96	65.97	47.50	86.67	32.69	51.63	61.76	48.47	38.10	66.51	56.76	53.89	59.24	46.88	58.16	34.90	66.18	51.39	58.08	57.72	51.89	26.82	1.15
Hunyuan-DiT	51.38	94.10	80.70	67.36	44.23	71.70	61.81	47.50	86.67	35.90	54.89	54.41	46.94	35.71	62.74	60.14	64.44	60.33	50.78	46.68	36.46	62.87	57.87	45.77	39.34	50.38	24.55	1.15
X-Omni	53.77	72.70	76.27	63.19	53.21	58.96	55.56	53.75	80.83	46.79	56.52	62.50	56.63	42.26	60.85	61.82	56.11	51.09	53.12	47.45	35.94	66.91	54.17	55.00	69.49	55.68	29.09	25.00
CogView4	56.30	82.00	83.07	71.53	44.23	55.19	72.22	57.50	89.17	53.85	59.78	68.38	50.51	51.19	62.74	60.47	60.00	69.57	60.16	47.19	42.19	69.49	56.02	38.46	77.21	60.98	28.18	17.82
OneCAT	58.28	93.30	82.28	59.42	58.33	67.45	65.97	42.50	92.50	35.90	65.22	69.12	57.65	48.81	71.23	78.04	69.44	62.50	51.56	66.33	47.40	70.59	59.72	51.54	64.34	65.15	33.41	1.15
Infinity	59.81	90.80	87.97	66.67	53.21	66.04	77.78	58.75	93.33	55.13	65.22	72.06	58.16	49.40	62.26	73.31	65.00	67.39	67.97	55.87	46.88	73.16	65.74	41.92	71.69	61.36	31.36	12.36
BLIP3-o	59.87	92.80	80.22	51.39	60.26	64.62	75.00	54.37	81.67	58.33	70.11	70.59	60.20	51.79	71.70	70.61	60.00	67.39	64.84	61.73	45.57	79.04	61.11	63.85	72.79	64.02	39.55	1.15
SD-3.5-Medium	60.71	89.80	84.34	59.72	51.92	67.92	70.83	63.75	93.33	50.00	63.04	69.12	55.61	52.98	71.70	74.66	61.67	73.37	58.59	58.16	48.44	73.53	61.57	44.23	72.06	68.56	37.73	15.23
FLUX.1-dev	61.30	83.90	88.92	72.22	53.85	58.96	75.00	65.00	91.67	51.28	67.39	69.85	59.69	58.93	65.57	62.50	66.67	72.83	62.50	47.96	46.09	73.16	63.43	46.15	74.26	69.32	30.91	32.18
Bagel	61.53	90.20	85.60	59.03	50.00	72.64	76.39	59.38	93.33	52.56	60.87	69.12	62.24	58.93	67.45	76.35	70.56	69.57	59.38	67.35	48.70	71.69	68.52	59.23	79.04	73.86	30.23	7.76
Janus-Pro	61.61	90.80	86.71	56.25	55.77	71.70	73.61	61.88	90.83	50.64	63.04	75.00	62.24	56.55	76.42	76.01	56.11	75.00	58.59	69.64	54.43	75.37	66.20	51.54	74.63	69.32	37.05	2.59
Show-o2	62.73	87.20	86.08	59.03	63.46	73.58	72.92	63.12	95.00	56.41	77.72	72.79	70.41	52.38	83.02	79.05	61.11	70.11	62.50	69.90	59.38	75.37	65.28	44.23	77.94	72.73	40.91	1.15
SD-3.5-Large	62.99	88.60	88.92	71.53	51.92	68.87	68.06	65.62	90.83	57.05	61.96	63.24	62.24	59.52	67.45	75.34	68.33	68.48	60.94	64.80	52.60	74.63	61.11	40.77	70.96	67.05	32.27	32.76
OmniGen2	63.09	91.90	86.39	67.36	73.08	66.04	72.22	66.25	95.00	55.77	69.02	68.38	62.24	54.17	66.51	68.24	67.78	71.20	64.84	62.24	50.26	71.32	60.65	47.31	78.31	64.77	32.50	29.02
UniWorld-V1	63.11	91.10	82.91	70.14	64.74	61.32	72.22	66.25	**99.17**	55.13	72.28	73.53	63.78	61.90	75.00	72.30	63.33	64.67	64.06	58.16	50.78	74.26	64.35	52.31	73.90	64.02	38.41	26.44
BLIP3-o-Next	65.15	91.00	86.71	67.36	73.72	70.28	76.39	60.62	80.00	57.69	75.00	73.53	67.35	57.74	68.87	76.01	65.00	77.17	75.00	73.72	55.73	76.47	67.13	60.00	80.15	72.35	48.64	4.60
Echo-4o	69.12	92.20	90.51	70.14	71.15	84.91	83.33	68.75	98.33	66.03	66.30	77.94	67.86	59.52	75.94	81.76	70.56	77.72	71.09	76.79	66.67	80.51	74.54	**70.00**	**87.13**	77.27	44.77	10.06
FLUX.1-Krea-dev	69.88	88.70	92.56	70.83	60.90	77.36	79.17	73.12	**99.17**	64.74	70.11	77.94	72.96	67.26	73.11	76.35	66.11	77.17	75.00	67.35	61.46	77.21	67.13	45.77	86.76	81.44	39.77	44.83
Lumina-DiMOO	71.12	89.70	90.03	69.44	85.90	81.60	76.39	80.00	**99.17**	64.10	78.80	75.74	73.98	64.88	82.08	**83.45**	**74.44**	81.52	67.97	**78.83**	67.71	81.99	**77.78**	52.31	84.93	80.68	45.45	25.57
HiDream-I1-Full	71.81	92.50	94.15	73.61	59.62	72.17	79.17	61.88	98.33	62.18	76.09	73.53	74.49	70.24	78.77	79.05	68.33	78.26	72.66	64.29	60.94	83.09	65.74	40.38	82.72	73.48	41.14	64.94
Hunyuan-Image-2.1	74.64	90.88	92.06	**86.62**	72.44	78.77	78.47	68.12	**99.17**	75.00	80.98	82.35	73.71	72.02	82.55	78.38	70.56	84.78	75.00	64.54	65.10	77.94	66.20	44.23	86.76	81.44	46.59	70.11
Qwen-Image	**78.81**	**95.10**	**94.30**	81.94	**84.62**	**91.98**	**84.03**	**84.38**	**99.17**	**82.05**	**88.59**	**88.24**	**80.61**	**77.38**	**87.74**	81.76	67.78	**86.96**	**81.25**	73.21	**73.44**	**83.82**	70.37	27.31	86.40	**85.23**	**53.64**	**76.14**

TABLE VIII
DETAILED BENCHMARKING RESULTS OF T2I MODELS ON UNIGENBENCH++USING ENGLISH LONG PROMPTS. *Gemini2.5-pro* IS USED AS THE VLM FOR EVALUATION. BEST SCORES ARE IN **BOLD**, SECOND-BEST IN UNDERLINED.

English Long Prompt Evaluation																													
Models	Overall	Style	World Know.	Attribute						Action						Relationship				Compound		Grammar			Layout		Logic. Reason.	Text
				Quant.	Express.	Materi.	Size	Shape	Color	Hand	Full Body	Animal	Non Contact	Contact	State	Compos.	Sim.	Inclus.	Compare.	Imagin.	Feat Match.	Pron Ref.	Consist.	Neg.	2D	3D		
Closed-source Models																													
Recraft	60.93	87.13	86.99	56.38	57.22	72.82	76.89	63.64	83.07	40.06	54.37	55.07	45.09	37.36	60.08	51.79	46.47	66.09	61.89	50.21	48.13	73.41	55.56	52.82	65.96	61.05	34.22	46.47
Stable-Image-Ultra	62.01	85.63	86.71	66.49	55.69	76.43	77.27	67.48	83.02	58.33	49.38	59.42	52.23	45.98	66.30	64.92	56.73	67.53	63.11	62.66	48.60	76.19	61.11	58.80	74.86	67.57	40.29	15.76
Runway-Gen4	68.29	91.72	88.82	70.65	65.43	85.33	81.01	67.38	85.64	55.33	63.92	70.65	56.82	56.10	69.76	70.05	59.09	76.76	70.39	69.47	66.50	76.23	62.70	72.76	72.56	75.37	48.28	27.47
Wan2.2-plus	68.76	90.28	87.57	78.19	69.17	80.42	82.77	73.60	88.10	64.10	60.94	70.29	59.38	55.46	73.32	69.13	66.67	81.03	77.43	74.16	66.36	86.90	61.11	63.38	82.34	75.00	55.58	12.77
DALL-E-3	70.82	95.08	92.71	64.67	72.59	88.72	89.48	77.14	90.15	63.49	63.96	67.03	59.55	60.17	76.29	80.57	70.51	83.53	73.76	77.67	65.00	82.92	66.27	56.99	69.22	75.00	57.11	18.26
FLUX-pro-1.1-Ultra	75.40	91.36	91.76	79.26	68.58	82.98	89.96	80.59	93.01	67.31	66.25	73.19	66.96	62.07	80.53	81.89	74.04	90.52	80.58	80.40	72.88	84.52	68.55	63.73	81.78	83.70	60.92	38.04
Imagen-3.0	75.76	92.41	94.19	75.58	71.41	88.34	88.52	78.27	93.13	73.63	77.12	76.81	69.44	65.48	80.62	80.15	74.17	90.59	78.54	81.14	73.22	91.67	76.61	66.67	83.97	88.69	61.25	24.18
FLUX-kontext-pro	78.58	94.83	93.60	74.47	75.00	85.47	89.58	80.63	92.89	73.05	73.12	75.00	67.73	70.40	77.98	73.85	72.08	89.08	82.77	83.58	71.23	90.32	75.40	66.90	84.09	87.23	66.26	49.73
FLUX-kontext-max	80.88	96.51	93.35	79.79	76.68	87.35	88.83	81.51	93.74	73.08	75.94	74.28	66.82	71.55	79.76	77.30	73.05	89.94	85.44	84.75	76.65	90.08	76.61	72.18	85.73	89.96	71.12	54.89
Seedream-3.0	80.99	97.18	93.79	83.51	81.25	93.07	88.26	90.03	97.48	77.88	84.69	78.26	74.11	71.84	83.60	81.63	79.17	87.64	86.41	80.49	82.24	90.48	80.56	56.69	87.85	89.13	62.62	56.52
Imagen-4.0-Fast	81.54	93.77	93.64	78.72	78.89	91.11	90.15	86.89	96.33	82.05	84.06	81.88	75.00	74.71	80.93	82.53	80.13	92.82	82.52	86.18	79.21	91.27	81.35	67.61	90.11	90.94	67.72	51.63
Wan2.5	84.34	96.75	95.52	85.64	81.01	94.03	88.17	87.50	96.11	73.08	82.91	77.21	71.76	69.83	81.27	85.26	81.41	94.48	88.11	87.55	81.31	92.86	77.42	65.49	88.28	85.77	71.32	73.10
Imagen-4.0	85.34	94.44	97.11	82.45	77.64	90.96	92.23	86.36	95.60	83.65	82.81	78.62	85.27	78.74	84.09	86.48	80.13	91.38	86.89	86.81	85.98	94.05	80.56	70.77	90.40	90.04	72.82	71.74
Nano Banana	88.82	98.83	95.78	88.24	86.09	93.05	93.70	88.73	97.31	84.57	84.95	81.16	83.41	78.16	86.28	**90.98**	**91.32**	92.80	91.91	92.15	87.23	94.84	89.24	84.51	**94.77**	93.12	81.27	69.75
Seedream-4.0	89.77	98.42	95.95	**92.02**	89.31	**95.26**	94.70	**92.48**	**98.27**	83.01	**87.50**	81.52	88.39	**83.62**	**89.82**	87.37	80.77	93.97	**92.72**	88.19	86.92	95.63	83.33	70.77	92.94	91.67	79.13	**90.76**
Imagen-4.0-Ultra	90.95	97.67	**98.26**	89.84	83.17	94.20	94.69	89.86	97.22	89.10	86.56	85.14	86.61	81.84	88.63	90.05	84.62	94.52	**92.72**	92.82	88.32	**96.83**	87.70	80.63	92.64	**94.57**	83.50	86.41
GPT-4o	**92.63**	**99.08**	97.95	86.70	**93.44**	92.45	**94.89**	**92.48**	94.95	**89.94**	87.19	**90.94**	**89.29**	83.05	87.75	89.18	90.71	**96.84**	90.29	**94.39**	**93.10**	95.97	**91.67**	**95.65**	94.29	92.70	**91.02**	83.79
Open-source Models																													
MMaDA	40.10	75.83	52.75	50.53	37.22	47.52	54.55	40.56	57.81	16.67	30.63	38.77	19.64	17.24	44.17	39.16	33.97	48.56	34.71	45.99	21.50	53.97	39.29	55.99	47.46	37.32	19.42	0.27
SDXL	41.48	81.81	69.51	39.36	44.03	58.89	58.14	43.01	58.81	19.23	29.69	29.35	17.41	16.67	43.87	41.07	27.88	42.24	28.40	41.24	18.93	53.57	37.70	48.94	39.12	42.03	19.42	0.82
Emu3	50.95	89.36	76.16	44.68	48.47	68.65	73.24	54.29	76.61	28.85	46.25	43.48	30.49	25.57	56.92	53.77	42.31	59.48	53.77	51.69	33.41	55.95	42.46	52.11	56.36	57.07	27.43	1.36
Kolors	53.60	86.54	76.01	61.17	50.42	72.67	71.97	58.74	74.06	39.74	38.44	50.36	44.64	34.20	63.24	58.04	58.01	62.36	56.55	52.11	36.45	72.22	53.57	41.55	61.02	60.87	31.31	2.17
Janus-flow	54.80	88.70	65.90	42.55	43.89	63.18	71.59	45.98	76.47	26.60	50.94	53.26	39.29	35.92	59.98	58.55	52.88	60.34	59.95	62.34	39.25	71.03	50.00	69.72	60.03	61.05	41.75	1.63
Hunyuan-DiT	54.88	92.94	80.06	65.43	52.22	72.14	75.19	58.22	76.31	39.10	46.25	47.46	41.07	34.48	59.58	56.89	55.45	57.18	52.18	55.49	38.55	64.68	59.52	52.82	60.45	62.68	29.85	1.63
Janus	60.37	92.03	73.27	42.55	48.61	71.31	79.17	57.69	82.86	39.42	57.19	64.86	51.34	40.23	64.23	62.76	60.26	67.82	62.62	69.73	44.39	74.21	59.52	67.96	62.85	65.76	54.37	1.09
BLIP3-o	61.01	91.61	74.42	54.26	61.81	70.93	78.22	57.87	78.88	48.08	54.69	61.23	46.88	35.92	64.82	60.97	57.69	62.36	69.66	70.89	53.74	74.60	62.30	59.86	77.40	70.11	48.30	1.36
OneCAT	62.92	94.93	83.67	61.70	66.39	78.09	82.58	62.24	78.88	37.82	59.06	62.32	50.89	43.97	71.44	67.47	62.82	63.22	65.05	72.57	43.69	74.21	67.46	50.70	75.28	73.01	48.06	1.90
SD-3.5-Large	64.35	88.12	88.15	68.62	62.22	81.85	78.79	70.63	86.32	57.69	52.81	57.25	50.89	48.85	68.68	70.15	62.18	70.11	64.81	65.82	54.21	75.79	61.51	59.15	73.45	68.30	44.90	17.66
SD-3.5-Medium	64.67	92.19	86.56	61.70	62.64	83.73	82.01	73.60	87.79	58.01	56.56	54.35	42.86	46.55	68.18	70.15	62.82	75.86	69.66	65.61	56.78	79.37	61.11	58.10	73.59	72.83	45.87	11.41
X-Omni	67.00	80.15	82.37	66.49	70.83	81.33	81.44	69.93	86.01	58.97	63.44	62.68	56.25	48.56	68.08	59.69	58.97	67.53	74.27	65.51	61.21	82.14	61.90	63.03	78.25	67.03	51.70	43.48
Infinity	67.28	92.77	88.44	70.74	66.67	82.83	82.95	71.15	88.73	58.65	60.13	67.75	58.48	52.87	69.07	66.20	67.63	78.45	72.09	68.57	60.75	76.59	71.43	58.80	80.93	73.19	51.46	13.59
CogView4	67.68	88.29	89.45	74.47	66.53	79.74	83.14	74.30	88.21	68.91	60.31	65.94	53.12	56.32	68.97	61.86	64.10	76.44	70.87	68.99	62.15	86.51	67.46	62.32	83.62	75.00	49.76	19.02
FLUX.1-dev	69.42	89.29	89.45	73.94	64.44	80.05	84.47	71.50	87.47	63.78	62.50	65.94	56.70	56.32	69.57	65.05	66.03	79.60	71.60	71.10	62.62	83.33	67.46	61.97	81.21	72.83	54.37	30.71
UniWorld-V1	69.60	93.19	84.10	66.49	72.64	77.11	81.06	72.38	87.95	63.78	64.38	67.03	62.95	55.17	70.85	66.96	67.31	72.99	70.39	74.16	65.19	84.13	69.44	72.18	83.33	74.82	57.04	20.92
Show-o2	70.33	93.11	88.44	59.04	71.53	88.10	87.31	81.12	94.71	53.85	80.00	69.20	60.27	55.75	76.68	77.42	68.59	80.17	81.55	77.64	73.83	87.30	66.67	58.45	80.08	81.34	59.71	1.90
BLIP3-o-Next	71.03	94.60	88.87	70.74	80.00	81.93	86.36	71.85	81.81	65.71	68.44	73.55	60.71	60.63	76.58	72.32	70.19	81.03	77.18	78.80	64.25	83.33	73.02	72.18	82.20	78.80	65.53	4.89
Janus-Pro	71.11	94.02	88.15	62.23	66.39	83.43	85.42	75.87	89.20	57.69	73.44	76.09	62.95	61.21	73.52	77.42	71.15	82.18	80.58	80.59	67.52	87.30	73.81	64.08	81.78	82.61	62.62	4.08
Bagel	71.26	92.44	89.31	69.68	70.28	85.17	86.17	76.92	91.88	68.59	67.19	68.48	58.48	59.77	71.94	72.19	72.12	85.92	76.46	77.32	68.93	87.30	70.63	67.25	83.47	79.89	59.71	12.23
OmniGen2	71.39	94.35	84.83	66.49	73.89	81.78	81.63	77.80	90.93	67.31	64.06	65.22	64.29	54.60	72.13	67.73	72.76	81.90	75.97	72.47	66.12	84.52	75.79	69.72	82.20	78.62	56.55	27.99
HiDream-I1-Full	74.25	93.11	92.63	73.40	68.47	83.51	84.47	75.70	92.19	65.06	68.44	62.32	71.43	57.47	75.20	72.07	73.40	78.74	75.49	73.63	61.21	86.51	69.84	62.68	82.63	76.45	50.24	57.61
FLUX.1-Krea-dev	78.45	94.10	93.79	81.38	76.81	91.34	88.64	85.31	95.44	75.00	76.25	72.46	69.20	72.99	80.43	80.87	73.08	88.22	84.47	80.59	80.84	**91.27**	74.21	61.97	85.45	**86.59**	65.53	41.03
Lumina-DiMOO	71.81	86.88	88.58	74.47	76.11	80.80	84.47	78.67	90.83	67.63	71.56	72.46	65.18	57.18	74.21	69.77	72.76	82.18	73.06	77.00	70.33	89.68	66.67	67.96	90.11	78.08	58.01	23.64
Echo-4o	76.41	96.10	90.17	73.40	82.08	92.39	89.20	84.44	95.49	72.12	76.56	73.19	66.96	65.23	77.47	83.80	78.21	84.77	82.77	**85.44**	**83.64**	86.11	**83.33**	**78.17**	88.70	83.51	**69.42**	8.15
Hunyuan-Image-2.1	82.19	94.52	93.35	86.17	85.56	93.75	**90.34**	**87.24**	**97.90**	**82.05**	**81.88**	79.71	76.79	**75.00**	84.09	**83.93**	78.53	**92.82**	**85.92**	82.28	82.94	**91.27**	75.79	66.55	90.25	**86.59**	68.20	58.15
Qwen-Image	**83.94**	**96.93**	**95.09**	**92.02**	**89.86**	**94.50**	89.58	86.71	97.85	78.53	**81.88**	**83.70**	**83.04**	71.84	**85.57**	81.76	**79.17**	88.79	85.19	82.38	81.07	90.48	78.57	54.93	**91.24**	86.05	66.75	**76.90**

TABLE IX
DETAILED BENCHMARKING RESULTS OF T2I MODELS ON UNIGENBENCH++USING CHINESE SHORT PROMPTS. *Gemini2.5-pro* IS USED AS THE VLM FOR EVALUATION. BEST SCORES ARE IN **BOLD**, SECOND-BEST IN UNDERLINED.

Chinese Short Prompt Evaluation																													
Models	Overall	Style	World Know.	Attribute						Action						Relationship				Compound		Grammar			Layout		Logic. Reason.	Text
				Quant.	Express.	Materi.	Size	Shape	Color	Hand	Full Body	Animal	Non Contact	Contact	State	Compos.	Sim.	Inclus.	Compare.	Imagin.	Feat Match.	Pron Ref.	Consist.	Neg.	2D	3D		
Closed-source Models																													
Runway-Gen4	54.93	64.75	71.05	54.29	46.05	72.60	57.64	50.62	81.90	52.63	65.22	75.00	51.56	54.37	65.09	66.89	51.11	74.43	72.66	68.22	53.49	55.38	55.09	64.29	59.93	69.62	42.03	0.59
Recraft	57.67	87.70	90.03	66.67	59.62	66.51	73.61	61.25	95.83	50.64	72.28	77.94	63.78	45.24	72.17	65.54	58.89	65.22	68.75	45.92	41.93	62.87	59.26	59.23	55.15	61.74	34.09	4.31
HiDream-v2L	59.95	89.34	91.02	71.43	42.31	70.59	70.00	64.52	94.17	48.72	65.22	75.00	71.88	55.95	71.15	78.82	65.00	75.56	65.32	62.63	43.55	75.38	68.75	44.53	66.29	63.26	32.01	1.16
Wan2.2-plus	66.96	91.06	84.39	75.00	67.31	74.06	74.31	66.25	90.83	69.23	80.00	84.56	65.31	61.90	75.94	71.28	72.78	85.87	82.03	74.23	55.00	77.21	63.43	69.62	73.16	70.45	51.82	11.92
DALL-E-3	67.93	95.90	93.04	60.42	68.59	91.04	90.28	65.00	94.17	69.87	77.17	82.35	66.33	61.90	76.89	81.76	77.78	87.50	67.97	82.14	63.54	79.78	76.39	58.85	54.41	70.83	51.59	1.15
Imagen-4.0-Fast	71.60	93.30	91.30	76.39	66.03	83.49	88.19	78.75	95.83	74.36	79.35	83.82	73.47	75.60	88.21	82.09	78.33	88.04	81.25	83.67	64.06	83.82	78.24	70.00	80.51	76.89	54.77	3.74
FLUX-kontext-max	71.85	96.38	92.83	65.97	69.44	80.19	84.72	66.67	93.33	76.32	83.15	83.33	69.90	73.17	85.78	85.14	74.43	91.67	83.59	82.65	67.12	79.85	75.46	71.48	81.62	81.06	56.48	1.72
Wan2.5	78.40	93.30	93.51	78.47	75.64	90.09	84.72	76.88	96.67	73.72	72.28	81.62	77.04	73.81	81.13	80.07	73.33	88.04	89.06	84.95	72.40	82.72	70.37	63.67	76.10	75.76	63.64	64.22
Imagen-4.0	79.52	97.50	96.84	83.33	77.56	92.92	93.75	72.50	98.33	89.10	89.67	93.38	86.73	90.48	93.40	91.55	83.33	94.57	93.75	92.60	78.65	92.65	82.87	72.69	91.54	86.74	73.18	2.59
Nano Banana	80.91	99.27	96.47	81.62	80.79	89.66	**95.74**	82.05	98.33	86.54	91.38	90.44	81.96	81.44	90.64	92.33	83.89	93.44	**96.88**	90.40	83.42	87.27	84.69	78.12	91.82	85.66	76.10	12.06
Seedream-3.0	81.68	97.50	93.99	84.03	82.69	94.34	89.58	80.00	97.50	85.26	90.76	89.71	85.20	80.36	90.09	86.82	74.44	90.22	84.38	82.14	71.09	84.19	79.17	39.62	89.34	78.79	59.09	78.74
Imagen-4.0-Ultra	83.21	98.90	97.94	88.89	79.49	94.81	93.75	88.12	**100.00**	**94.87**	92.93	**95.59**	87.76	**95.24**	**97.17**	91.22	87.22	**97.83**	92.97	94.90	84.90	93.01	85.65	83.08	**93.75**	**90.53**	79.32	9.77
Seedream-4.0	87.31	99.00	94.94	86.81	85.90	**97.64**	86.81	83.12	99.17	82.69	90.22	91.91	84.69	82.74	92.45	85.14	84.44	95.65	92.19	85.20	77.86	89.71	75.00	69.62	90.81	89.77	68.64	**93.97**
GPT-4o	**91.02**	**99.39**	**98.72**	**93.62**	**94.59**	96.19	93.06	**92.95**	**100.00**	94.08	**97.28**	90.91	**90.31**	88.34	92.65	**97.30**	**93.18**	96.69	94.53	**95.92**	**91.74**	**95.15**	**89.35**	**88.05**	89.18	89.35	**91.44**	63.37
Open-source Models																													
UniWorld-V1	15.21	49.40	16.61	14.58	19.87	8.02	13.19	5.00	37.50	9.62	17.93	18.38	9.69	6.55	24.06	16.55	6.67	12.50	7.03	6.63	2.08	19.85	16.20	45.77	8.09	10.23	2.95	0.29
Janus-flow	20.93	58.50	18.67	22.92	10.90	21.70	24.31	8.12	30.00	4.49	31.52	22.06	14.80	19.05	35.85	23.65	16.11	20.11	14.06	19.13	2.08	32.72	16.67	52.69	12.13	17.80	10.68	0.00
Janus-Pro	30.83	75.60	39.08	24.31	19.23	43.87	45.14	18.75	47.50	13.46	26.09	34.56	22.45	20.83	38.68	38.85	35.56	26.09	24.22	33.42	15.36	36.76	31.94	40.38	29.78	30.30	10.23	0.00
Janus	30.98	78.10	27.85	29.17	17.31	35.85	45.83	14.37	17.31	14.10	38.59	42.65	24.49	23.21	43.40	32.43	32.22	27.72	28.12	25.26	9.64	48.53	33.33	60.77	31.25	32.20	13.41	0.00
Emu3	33.91	78.08	55.54	27.78	30.13	44.34	32.64	27.67	71.67	16.67	36.96	49.26	26.02	17.86	40.57	43.58	31.67	38.04	25.78	29.85	13.28	41.91	38.89	42.69	17.71	27.27	13.90	0.00
MMaDA	44.00	78.20	52.06	52.78	33.97	58.49	61.11	45.00	86.67	24.36	54.35	47.06	31.63	29.17	67.92	59.80	52.22	60.87	46.88	39.29	26.30	59.93	46.30	67.31	38.97	35.61	26.14	0.00
BLIP3-o-Next	44.48	74.60	50.00	44.44	57.69	56.13	63.89	48.12	68.33	37.82	61.41	45.59	45.41	36.90	54.72	54.05	48.33	50.00	64.84	32.14	20.83	65.07	49.54	46.54	58.82	50.76	27.50	0.00
HiDream-I1-Full	50.65	83.30	78.32	69.44	45.51	55.66	70.14	55.00	86.67	44.23	57.61	55.88	53.06	47.62	61.32	57.77	52.78	63.04	53.91	38.01	30.99	62.13	51.85	46.92	63.60	55.68	23.64	0.00
Hunyuan-DiT	53.36	92.50	84.97	63.19	46.15	72.17	63.89	49.38	85.00	45.51	67.93	61.76	48.47	47.02	69.81	65.88	64.44	56.52	41.41	52.04	36.98	59.93	62.04	43.08	39.71	56.06	29.55	0.00
X-Omni	53.69	70.07	71.52	61.81	52.56	63.51	67.36	57.50	85.83	48.72	68.48	63.97	56.53	43.45	66.51	60.14	60.00	62.50	54.69	48.72	34.64	63.97	53.70	50.38	66.91	51.89	34.77	20.98
CogView4	55.14	82.40	84.18	68.75	44.87	56.60	72.92	53.75	94.17	61.54	66.30	64.71	52.04	54.76	70.28	61.82	62.22	63.59	57.81	51.02	40.36	67.65	57.41	38.46	75.00	55.30	30.23	2.30
Lumina-DiMOO	58.35	80.90	69.46	62.50	71.79	77.83	78.47	70.00	96.67	42.95	61.41	76.47	58.67	51.79	74.06	68.58	62.78	76.09	57.03	56.96	52.34	76.10	70.37	48.46	73.53	64.77	39.09	0.00
OneCAT	58.50	94.40	86.55	56.94	66.03	73.58	65.28	38.75	84.17	42.31	75.00	80.88	61.22	44.05	73.58	72.64	61.67	69.57	60.16	63.52	39.32	64.34	60.19	52.69	61.76	59.09	38.64	0.00
Kolors	58.80	85.20	86.23	70.14	51.92	73.11	77.78	56.25	91.67	58.33	59.24	71.32	63.78	57.54	77.83	71.96	69.44	67.39	52.34	64.80	45.05	67.28	59.26	43.46	58.82	65.91	36.14	4.89
BLIP3-o	59.25	92.60	81.17	57.64	65.38	67.92	77.08	47.50	89.17	57.69	73.37	68.38	59.18	55.95	70.28	69.26	58.33	63.04	69.53	61.99	41.41	70.22	57.41	61.16	69.12	62.12	41.59	0.00
OmniGen2	63.20	93.00	86.39	67.36	69.87	78.30	77.78	68.75	93.33	64.10	69.57	74.26	61.73	55.95	73.58	77.03	66.67	71.74	60.16	66.33	53.39	71.69	71.30	54.62	76.84	62.88	44.09	0.29
Bagel	65.69	92.30	86.71	64.58	63.46	83.49	79.86	66.25	95.00	61.54	63.59	75.74	65.31	61.90	67.92	77.70	67.78	82.07	71.09	79.59	59.90	73.16	75.00	61.15	82.72	72.35	37.95	6.61
Echo-4o	72.40	92.80	87.66	72.92	77.56	89.15	88.19	80.00	**99.17**	73.08	83.15	85.29	75.00	65.48	75.47	**85.81**	75.00	88.04	75.78	**82.91**	**72.92**	80.15	**77.31**	**68.85**	84.19	81.82	56.82	7.76
Hunyuan-Image-2.1	77.76	92.20	90.51	87.50	80.77	82.55	86.11	75.00	97.50	76.28	84.24	85.29	78.06	79.17	80.66	80.74	**80.56**	87.50	**83.59**	71.68	69.53	80.15	67.13	37.31	**88.24**	**82.58**	50.23	79.60
Qwen-Image	**81.04**	**95.50**	**92.41**	**88.89**	**91.03**	**96.23**	**90.28**	**86.25**	98.33	**83.33**	**87.50**	**89.71**	**81.63**	**82.14**	**90.09**	85.47	73.33	**90.76**	79.69	80.10	72.14	**83.46**	74.07	31.92	84.93	80.30	**57.73**	**82.47**

TABLE X
DETAILED BENCHMARKING RESULTS OF T2I MODELS ON UNIGENBENCH++USING CHINESE LONG PROMPTS. *Gemini2.5-pro* IS USED AS THE VLM FOR EVALUATION. BEST SCORES ARE IN **BOLD**, SECOND-BEST IN UNDERLINED.

Chinese Long Prompt Evaluation																												
Models	Overall	Style	World Know.	Attribute						Action						Relationship				Compound		Grammar			Layout		Logic. Reason.	Text
				Quant.	Express.	Materi.	Size	Shape	Color	Hand	Full Body	Animal	Non Contact	Contact	State	Compos.	Sim.	Inclus.	Compare.	Imagin.	Feat Match.	Pron Ref.	Consist.	Neg.	2D	3D		
Closed-source Models																												
Recraft	56.90	86.38	85.55	61.70	60.56	73.72	79.92	65.03	82.39	44.23	57.81	60.87	42.86	43.39	61.66	54.72	49.68	63.22	63.59	50.95	47.90	71.83	55.95	46.13	64.12	65.04	36.17	2.45
Wan2.2-plus	70.05	91.61	88.73	78.19	66.94	82.15	84.09	77.10	89.99	67.95	69.06	72.46	64.29	63.79	74.21	70.15	70.83	80.17	76.94	74.26	65.42	83.73	62.70	64.44	81.50	78.26	57.04	15.22
DALL-E-3	71.16	95.85	94.36	64.36	71.11	88.93	90.72	77.62	91.30	61.22	65.94	74.28	67.41	62.64	77.37	81.63	73.72	85.63	77.43	80.38	65.89	80.16	74.21	59.51	70.48	76.99	61.41	3.80
Imagen-3.0	71.85	89.25	94.75	75.78	64.67	80.66	82.84	70.00	93.10	80.00	83.89	85.29	77.37	74.40	87.38	83.90	73.33	88.64	83.90	79.23	64.06	79.04	70.75	59.13	82.72	79.92	48.36	21.55
FLUX-kontext-max	75.24	97.59	92.31	72.34	71.41	87.48	88.83	81.64	92.80	76.28	70.22	79.35	69.20	74.43	78.16	78.95	73.40	87.25	86.65	84.60	70.33	88.76	76.19	72.24	87.01	88.32	68.20	4.62
Imagen-4.0	79.90	95.60	**97.98**	82.45	80.42	92.24	91.29	85.84	96.28	81.09	84.69	82.25	83.48	85.63	86.07	87.24	82.05	93.97	89.08	88.71	82.01	92.06	81.75	75.35	90.25	90.76	77.18	4.89
Nano Banana	83.17	98.41	97.38	**90.37**	85.06	93.11	94.29	87.99	98.10	84.42	88.09	84.06	87.05	82.90	86.07	90.59	86.50	96.83	91.71	92.14	89.13	94.78	88.10	82.86	93.19	93.10	82.40	10.68
Imagen-4.0-Ultra	83.86	97.34	97.40	88.30	83.75	94.13	95.27	90.91	97.80	83.97	90.94	88.41	87.50	**88.79**	90.02	**92.22**	87.82	**96.84**	92.23	93.99	89.25	**96.83**	90.08	80.63	94.77	93.30	86.89	6.79
Wan2.5	84.24	98.00	94.30	83.51	80.90	91.77	91.41	87.24	94.59	72.12	78.16	83.82	74.55	75.29	80.85	85.59	77.56	91.95	91.02	86.18	82.78	91.67	79.37	70.42	89.91	86.78	74.51	66.30
Seedream-3.0	86.14	98.42	95.36	85.64	83.98	**96.39**	90.53	93.36	97.90	81.41	89.06	86.13	85.71	79.19	85.18	84.57	83.01	93.10	91.99	83.83	81.54	88.89	82.14	63.38	90.68	89.49	68.45	82.34
Seedream-4.0	90.35	98.42	96.39	86.70	90.69	96.08	**95.45**	93.71	**98.43**	84.94	91.56	**92.03**	**92.41**	86.21	**89.53**	86.35	83.01	93.39	93.45	87.66	87.85	94.44	82.14	75.35	92.66	90.94	80.58	**91.30**
GPT-4o	**90.51**	**99.41**	97.96	85.87	**92.56**	94.43	95.23	**94.23**	96.59	**91.12**	**92.50**	89.49	91.52	86.78	88.14	91.93	**89.10**	95.64	**93.93**	**95.36**	**92.87**	96.37	**92.86**	**93.24**	**95.01**	**95.47**	**90.05**	57.14
Open-source Models																												
UniWorld-V1	21.50	55.48	17.34	12.23	30.28	19.80	27.27	19.76	35.69	12.18	20.31	23.19	9.38	8.05	26.28	16.20	21.47	23.56	20.15	15.30	6.31	23.81	21.03	39.79	24.15	24.82	8.98	1.36
Janus-flow	23.01	57.39	17.49	11.70	11.39	23.72	32.20	15.91	28.72	3.85	18.75	19.20	9.38	9.48	30.24	18.62	18.91	24.43	19.90	28.80	5.61	29.76	13.89	50.70	18.64	25.36	17.48	0.27
Janus	33.63	75.00	30.06	25.53	25.97	39.16	45.83	22.20	39.99	11.54	35.31	32.25	16.96	14.08	41.11	26.02	26.60	30.46	31.80	38.92	14.95	46.43	24.60	59.15	38.98	42.57	20.15	1.09
Emu3	35.95	75.08	53.03	23.40	38.33	49.17	57.77	36.19	56.34	10.58	22.81	25.36	12.05	17.53	42.39	33.29	29.17	35.06	29.37	33.02	18.46	42.86	26.59	44.72	30.37	41.85	19.66	0.82
MMaDA	50.61	84.05	63.58	46.81	40.00	58.96	67.80	52.62	73.22	23.40	39.06	40.58	29.02	30.75	58.20	48.09	49.04	60.63	57.52	56.65	35.51	61.11	50.79	63.73	65.54	54.35	31.80	0.27
HiDream-I1-Full	50.70	83.06	78.61	63.30	55.97	62.50	69.70	56.12	71.80	38.14	45.00	44.93	38.39	36.21	57.71	46.30	45.83	59.20	49.03	45.99	33.41	59.52	49.60	52.46	62.99	57.07	24.27	2.99
BLIP3-o-Next	54.55	87.71	61.85	50.00	64.58	67.85	67.61	55.94	63.21	37.50	56.25	50.72	45.98	37.36	61.36	55.36	53.53	60.34	63.35	59.49	41.82	65.48	58.73	58.10	67.80	60.51	41.50	1.90
Hunyuan-DiT	55.57	94.10	76.16	66.49	54.03	71.76	76.14	58.57	76.10	41.03	51.56	57.25	41.52	37.36	59.09	59.69	48.08	56.90	52.43	57.49	39.95	63.49	60.71	56.34	60.73	62.86	33.98	1.36
BLIP3-o	59.25	89.70	77.17	53.19	59.03	71.31	79.36	54.02	75.00	42.63	59.38	60.87	45.98	43.97	64.03	58.29	54.81	60.63	69.17	67.72	45.09	72.22	53.17	57.75	72.60	65.04	47.09	1.90
Janus-Pro	60.21	91.28	75.87	44.15	52.92	69.80	78.22	56.99	69.18	37.82	51.25	63.04	48.21	51.72	60.28	62.50	57.05	66.38	63.83	72.47	50.47	72.22	61.11	71.83	66.38	66.85	49.27	2.17
X-Omni	62.18	76.91	74.13	72.34	59.72	77.79	82.20	67.83	83.39	50.00	61.56	61.96	49.55	42.82	66.40	57.02	55.45	65.52	68.20	65.51	51.40	76.19	58.33	60.56	76.84	68.12	46.60	29.35
Lumina-DiMOO	63.80	84.30	76.45	64.36	68.06	77.18	82.01	72.73	88.00	54.81	57.50	61.96	60.27	49.43	68.68	62.24	61.22	78.74	69.17	72.57	60.75	76.98	67.06	71.83	84.18	70.83	49.27	1.36
OneCAT	63.88	95.85	85.26	57.98	65.56	78.92	81.25	59.79	79.77	35.26	69.69	64.13	55.36	42.24	70.85	63.65	63.14	65.52	68.69	70.78	43.69	69.05	63.49	57.39	76.13	75.36	54.37	2.17
Kolors	65.12	90.61	87.14	63.83	64.86	82.98	83.52	70.80	90.25	58.97	57.19	63.41	65.18	50.57	73.42	69.90	74.68	74.43	68.45	67.83	56.07	81.35	62.30	50.00	72.46	77.36	47.82	5.98
CogView4	68.09	89.62	89.31	73.40	65.69	80.35	85.98	73.43	88.84	67.31	68.75	71.01	58.04	63.79	70.65	66.07	64.10	80.17	75.97	71.94	65.42	83.33	69.05	61.62	77.72	84.46	51.94	8.15
OmniGen2	70.75	95.35	87.57	74.47	73.33	84.94	85.23	79.90	92.09	63.46	67.81	63.41	63.39	60.34	72.33	70.79	70.51	87.64	77.43	76.05	69.63	85.71	76.59	69.72	84.89	76.81	62.62	1.90
Bagel	75.75	96.10	89.02	71.81	73.47	88.93	90.53	83.39	95.81	71.47	75.62	76.09	66.96	63.22	75.10	80.87	76.60	86.78	82.04	83.97	77.80	84.92	83.33	75.70	87.29	79.71	68.69	14.40
Echo-4o	78.31	96.26	91.18	71.81	82.22	94.50	90.72	88.64	96.80	73.72	81.56	74.28	67.41	66.38	79.55	86.99	81.09	89.08	84.47	**86.08**	83.41	87.70	**83.73**	**79.58**	90.54	84.96	**72.57**	13.04
Qwen-Image	86.91	**97.84**	**95.66**	**89.36**	**91.11**	**96.23**	**93.56**	**90.91**	**97.90**	83.33	**90.62**	**89.86**	**86.61**	**79.60**	**87.75**	85.59	**84.29**	91.67	**90.53**	83.44	82.01	**94.05**	**83.73**	55.63	92.09	88.41	69.90	86.14
Hunyuan-Image-2.1	**87.01**	95.18	94.08	87.77	87.08	95.41	91.67	89.69	97.69	**85.58**	84.69	85.51	83.48	79.02	84.68	**87.88**	81.41	**92.24**	90.05	85.97	**84.81**	92.86	83.33	65.85	**93.50**	**88.77**	71.36	**86.41**