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

Anonymous authors
Paper under double-blind review

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) it 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 test points. To rigorously assess model robustness to variations in language and prompt length, we provide both English and Chinese versions of each prompt in short and long forms. Leveraging 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, we develop an effective pipeline for reliable benchmark construction and streamlined model assessment. Moreover, to further facilitate community use, we train a robust evaluation model that enables offline assessment of T2I model outputs. Through comprehensive benchmarking of both open- and closed-source models, we systematically reveal their strengths and weaknesses across various aspects.

Keywords: Text-to-image generation, semantic generation evaluation, benchmark.

1 Introduction

Recent progress in text-to-image (T2I) generation Ho et al. (2020); Rombach et al. (2022); Liu et al. (2022); OpenAI. (2023); Podell et al. (2023); Esser et al. (2024); Xie et al. (2025); Wu et al. (2025c); Labs (2024); Gao et al. (2025); Seedream Team et al. (2025); Cai et al. (2025); Google (2025c); OpenAI (2025); Wu et al. (2025a); Li et al. (2024a); Wang et al. (2025d); Google (2025b); Labs. (2025) 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 Rombach et al. (2022); Podell et al. (2023), Playground Li et al. (2024a), and FLUX Labs (2024); Labs. (2025), iteratively refine Gaussian noise using U-Net or Transformer backbones to generate images. (2) Autoregressive (AR) approaches, such as Infinity Han et al. (2025), Janus series Wu et al. (2025b); Chen et al. (2025b); Ma et al. (2024), and BLIP3-o Chen et al. (2025a), treat images as token sequences and synthesize them via next-token prediction or progressive scaling. Recent methods incorporate reinforcement learning Rafailov et al. (2023); Guo et al. (2025); Wang et al. (2025a; 2024b) to improve T2I models' instruction following capability Wang et al. (2025b); Tong et al. (2025) and the visual quality of generated images Liu et al. (2025); Xue et al. (2025); Zhou et al. (2025). 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 Ghosh et al. (2023);

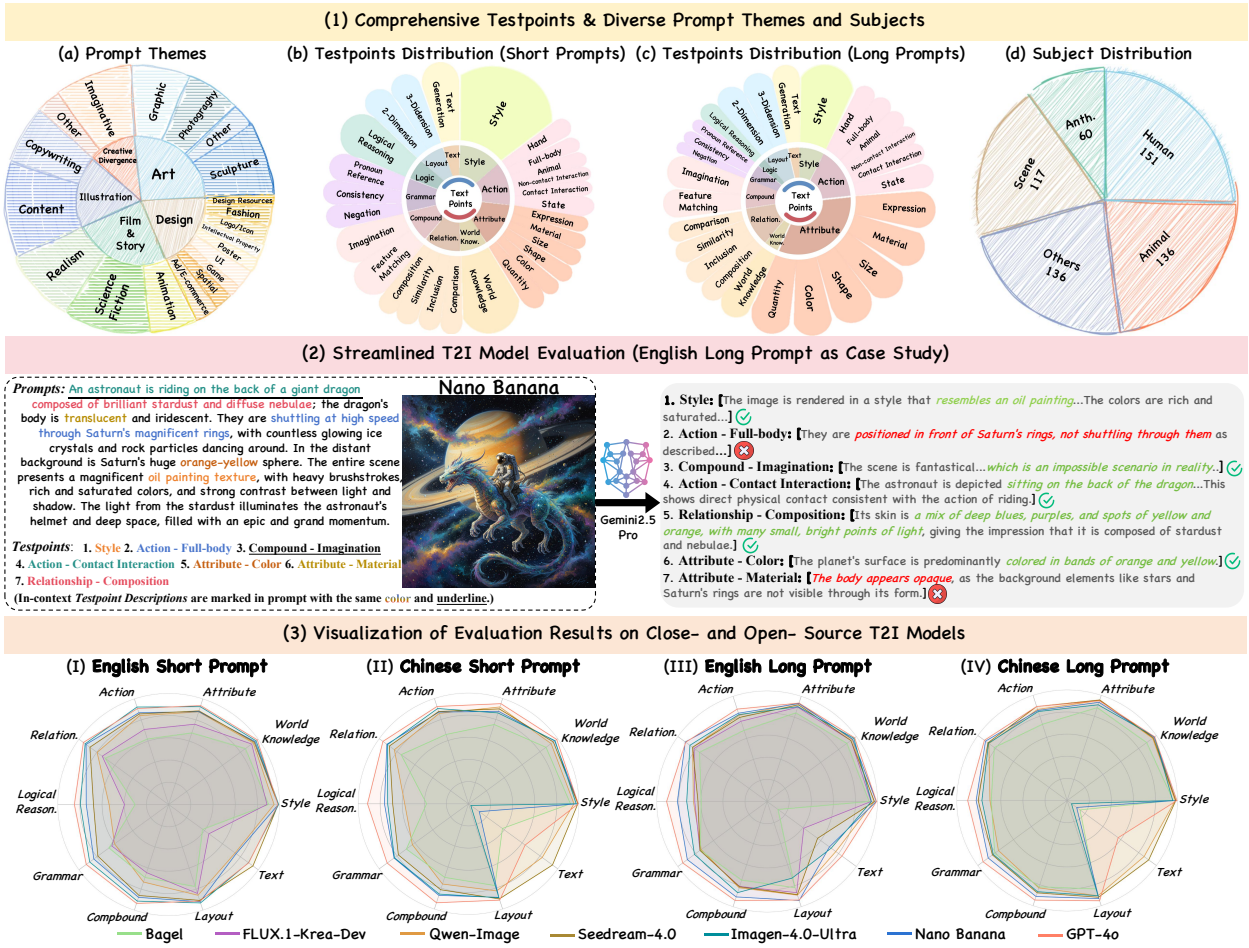

Figure 1: **Benchmark Overview.** (1) Our UniGenBench++ covers diverse prompt themes, subjects, and comprehensive evaluation criteria. (2) Each prompt includes multiple test points and is assessed through a streamlined MLLM-based pipeline for reliable and efficient evaluation. (3) We conduct comprehensive 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.

Huang et al. (2023) 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 Niu et al. (2025); Sun et al. (2025) evaluate the implicit semantic understanding and world knowledge integration capabilities of T2I models using powerful visual-language models (VLMs) Hurst et al. (2024) as the evaluator. Recent efforts broaden T2I evaluation by incorporating long-prompt semantics generation Hu et al. (2024); Wei et al. (2025) and additional evaluation dimensions Wei et al. (2025) such as style and text generation.

Despite effectiveness, as shown in Tab. 1, these benchmarks encounter two key limitations: (1) **Coarse evaluation on limited dimensions:** they 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 are incapable of fine-grained assessment for each sub-dimension; (2) **Lacking diversity of prompt scenarios and multilingual evaluation:** they only focus on evaluation dimension design but neglect the diversity of prompt scenarios and multilingual evaluation support, hindering comprehensive assessment in real-world applicability.

Table 1: **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	✓	✓

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 Ghosh et al. (2023); Hu et al. (2024); Wei et al. (2025). 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. Furthermore, 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 Wei et al. (2025); Huang et al. (2023); Ghosh et al. (2023), 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 explicitly designed test points (each accompanied by its in-context description within the prompt), the evaluation model, *i.e.,* Gemini-2.5-Pro Google (2025a), 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 OpenAI (2025), Nano Banana Pro Google (2025c), Seedream-4.5 Seedream Team et al. (2025), and FLUX-Kontext-Max Labs (2024), as well as leading open-source counterparts, including Z-Image Z-Image Team (2025), Qwen-Image Wu et al. (2025a), FLUX.2-dev Labs (2025), Lumina-DiMOO Xin et al. (2025) and Bagel Deng et al. (2025). 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:

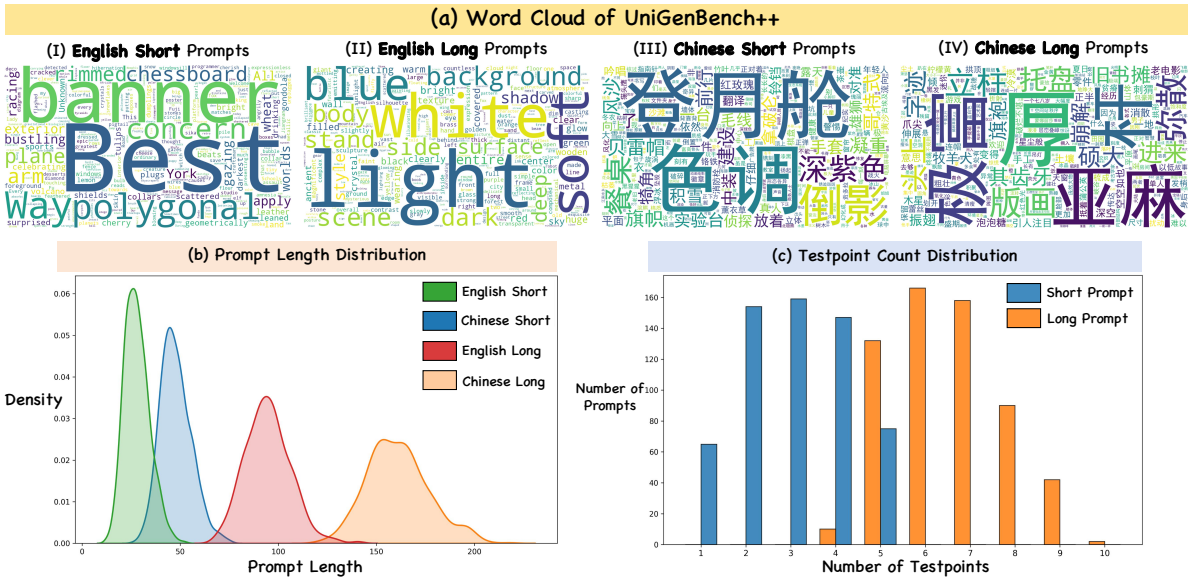

Figure 2: **Benchmark Statistics.** (a) Word clouds for English and Chinese prompts in both short and long forms; (b) overall prompt length distribution; and (c) distribution of testpoint counts per prompt for short versus long versions.

- 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 test points, 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 test point 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.

2 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 Nichol et al. (2021) and Imagen Google (2025b) to powerful variants like Stable Diffusion Rombach et al. (2022), FLUX Labs (2024), and HiDream Cai et al. (2025). In contrast, AR models generate images token by token via VQ-VAE Van Den Oord et al. (2017) compression and transformer decoding, as seen in DALL · E OpenAI. (2023) and CogView Ding et al. (2021). Recent advances Li et al. (2025); Chameleon Team (2024) enhance AR models with unified multimodal reasoning, while hybrid architectures like Bagel Deng et al. (2025) 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 Ghosh et al. (2023) 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 Huang et al. (2023) 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 Sun et al. (2025) and general commonsense understanding Niu et al. (2025); Luo et al. (2025). However, the prompts used in these benchmarks are predominantly short and highly repetitive, which constrains semantic richness and expressiveness. Therefore, DPG-Bench Hu et al. (2024) centers on assessing models' capability in dense prompts. TIIF-Bench Wei et al. (2025) 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.

3 Benchmark

3.1 Overview

With the rapid advancement of text-to-image (T2I) models, existing evaluation frameworks Ghosh et al. (2023); Huang et al. (2023); Wei et al. (2025); Hu et al. (2024) have become increasingly insufficient. To be precise, (1) as summarized in Tab. 1, 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. 1, 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. 3.2).
- **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. 3.3).
- **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. 3.4). Leveraging the world knowledge and fine-grained image understanding capabilities of Multi-modal Large Language Models (MLLMs), *i.e.,* Gemini-2.5-Pro, we design a fully streamlined pipeline for accurate and efficient model evaluation (Sec. 3.5).
- **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. 3.6).

3.2 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 20 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. 6.

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.

3.3 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.
- **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.

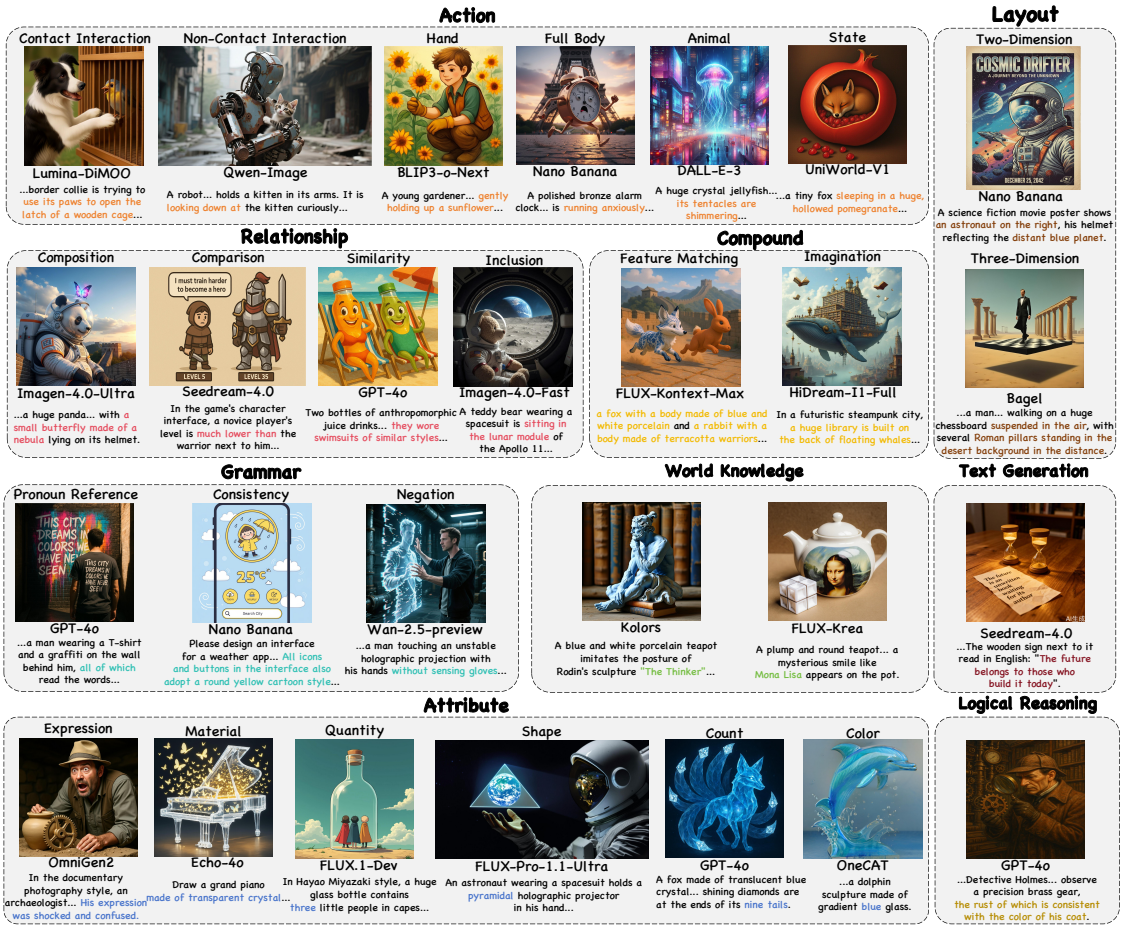

Figure 3: **Qualitative Results of Evaluation Dimensions.** We present qualitative examples of T2I models evaluated across our specified dimensions.

- **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 touching 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 assesses 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 test points 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).

3.4 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 test points $c_1, \ldots, c_k \subset \mathcal{C}$, where $k \in [1, 5]$, is selected to specify the targeted fine-grained test points.

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:

$$\big(p^{\text{en}}, p^{\text{zh}}, \{d_i\}_{i=1}^{k}\big) \sim \text{MLLM}_{\text{gen}}\big(t, s, \{c_i\}_{i=1}^{k}\big), \tag{1}$$

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 a 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 subjects and their key attributes must be preserved, and (ii) attribute, scene, and background details may be further elaborated to enhance specificity and imagination. Formally,

$$\tilde{p} \sim \text{MLLM}_{\text{expand}}\big(p \mid r\big), \tag{2}$$

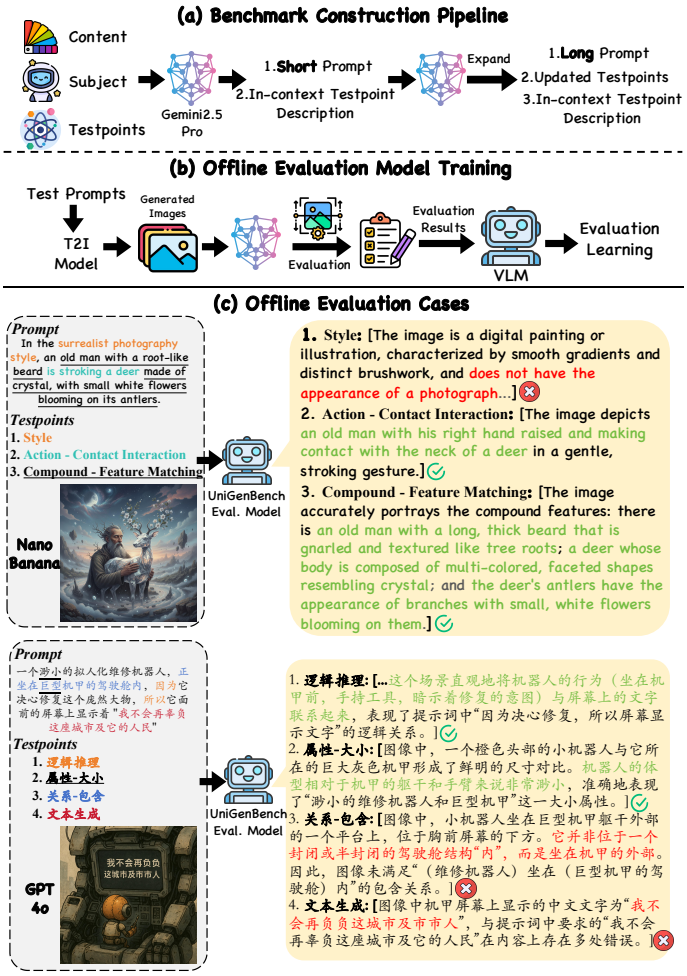

Figure 4: **Pipeline of Benchmark Construction and Offline Evaluation Model Training.** (a) Benchmark construction pipeline; (b) Offline evaluation model training; (c) Offline evaluation cases.

where r denotes the rewriting constraint.

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

$$\begin{aligned} \{(\hat{c}_i, \hat{d}_i)\}_{i=1}^{k'} &\sim \mathrm{MLLM}_{\mathrm{align}}\big(\cdot \mid \tilde{p},\ \{(c_i, d_i)\}_{i=1}^{k}\big), \\ k' &\leq k + 5, \end{aligned} \tag{3}$$

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

$$\big(\tilde{p},\ \{(\hat{c}_i, \hat{d}_i)\}_{i=1}^{k'}\big)$$

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

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).

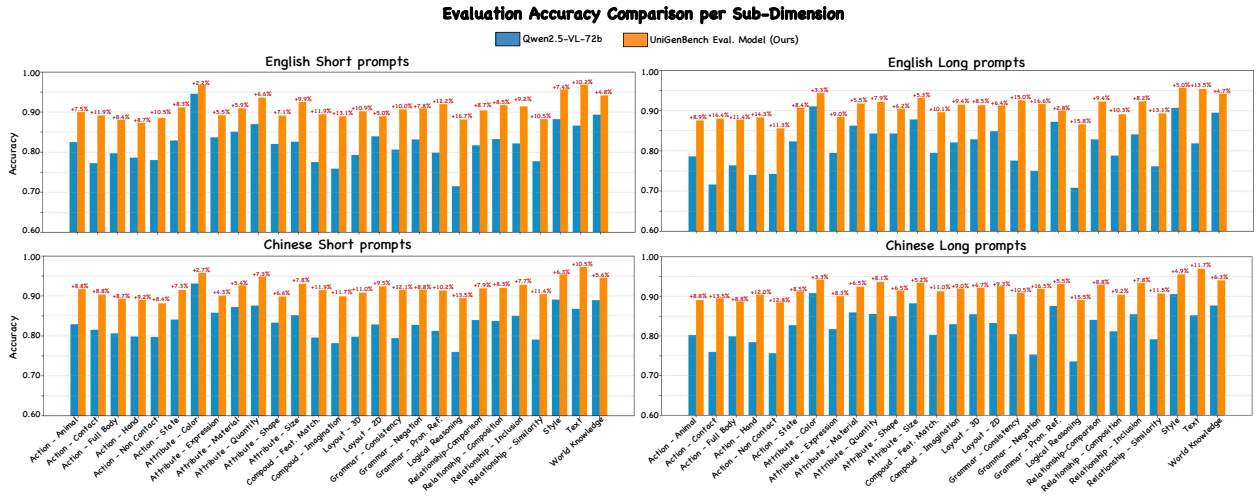

Figure 5: **Evaluation Accuracy Comparison.** Our dedicated evaluation model demonstrates a significant improvement in evaluation accuracy across all test points compared to the commonly used offline evaluation VLM, *i.e.,* Qwen2.5-VL-72b.

3.5 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 test points $\{c_{i,1}, \ldots, c_{i,k}\}$ 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:

$$\big(\{r_{i,j}\}_{j=1}^{k}, \{e_{i,j}\}_{j=1}^{k}\big) \sim \text{MLLM}\big(\{r_{i,j}, e_{i,j}\}_{j=1}^{k} \mid x_i, p_i, \{d_{i,j}\}_{j=1}^{k}\big). \tag{4}$$

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 downstream error attribution. We provide an 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 test points, its score is defined as the ratio of satisfied instances to the total number of its occurrences across the benchmark:

$$R_c = \frac{\sum_{i,j} \mathbf{1}\{d_{i,j} \in c \text{ and } r_{i,j} = 1\}}{\sum_{i,j} \mathbf{1}\{d_{i,j} \in c\}}, \tag{5}$$

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.

3.6 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 evaluation. 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:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T_i} \log P_\theta\big(y_i^{(t)} \mid y_i^{(<t)},\, x_i, p_i, \{d_{i,j}\}_{j=1}^{k}\big), \tag{6}$$

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

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

4 Experiment

4.1 Implementation Details

4.1.1 Benchmarking Models

Closed-source Models. GPT-4o Hurst et al. (2024), Imagen-3.0/4.0-Ultra/Fast Google (2025b), Nano Banana (Pro) Google (2025c), Seedream-3.0/4.0/4.5 Gao et al. (2025); Seedream Team et al. (2025), Wan2.2/2.5 Cloud (2025), Runway-Gen4 Runway (2025), Recraft Recraft (2025), DALL-E-3 OpenAI. (2023), FLUX-Pro-1.1-Ultra/Kontext-Max/Kontext-Pro Labs (2024), HiDream-v2L Hidream (2025), Stable-Image-Ultra Stability (2025), FLUX-2-Pro/Flex/Max Labs (2025).

Open-source Models. Qwen-Image Wu et al. (2025a), Hunyuan-Image-2.1 Tencent (2025), HiDream-I1-Full Cai et al. (2025), Lumina-DiMOO Xin et al. (2025), Show-o2 Xie et al. (2025), Infinity Han et al. (2025), OneCAT Li et al. (2025), CogView4 Ding et al. (2021), X-Omni Geng et al. (2025), MMaDA Yang et al. (2025), FLUX.1-dev Labs (2024), FLUX.1-Krea-dev Labs. (2025), Echo-4o Ye et al. (2025), BLIP3-o (Next) Chen et al. (2025a), UniWorld-V1 Lin et al. (2025), OmniGen2 Wu et al. (2025c), Bagel Deng et al. (2025), Hunyuan-DiT Li et al. (2024b), Janus series Wu et al. (2025b); Ma et al. (2024); Chen et al. (2025b), Emu3 Wang et al. (2024a), Kolors Kolors Team (2024), SDXL Podell et al. (2023), SD-3.5-Large Rombach et al. (2022), GLM-Image Z.ai (2026), Z-Image (Turbo) Z-Image Team (2025), LongCat-Image Meituan LongCat Team et al. (2025), and FLUX.2-dev/Klein Labs (2025).

4.1.2 Offline Evaluation Model

We use UnifiedReward-2.0-qwen-72b Wang et al. (2025c) 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.

4.2 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.

4.2.1 English Short Prompt (Tab. 2)

(a) *Closed-source Models.* GPT-4o-1.5 achieves the best overall performance, showing consistently strong scores across nearly all dimensions. Nano Banana Pro also performs competitively and remains among the

Table 2: **Overall Benchmarking Results of T2I models on UniGenBench++ using English short prompts**. *Gemini-2.5-Pro* is used as the MLLM 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
DALL-E-3	68.85	94.43	92.64	75.76	70.78	78.31	46.22	69.22	71.08	65.65	24.43
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	70.46	90.99	91.30	76.79	71.39	78.05	41.46	68.18	68.17	80.60	37.64
Imagen-3.0	71.34	89.35	93.95	77.92	78.80	82.75	45.09	69.97	72.81	80.04	22.70
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	77.69	91.90	95.73	83.01	80.23	82.61	56.82	76.87	72.68	86.75	50.29
Wan2.5	77.87	92.64	94.75	81.49	74.14	81.98	55.50	72.79	75.45	76.87	73.12
Seedream-3.0	78.41	98.19	94.90	84.62	83.14	80.18	51.83	60.30	72.32	88.74	69.86
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
Nano Banana	87.29	98.59	96.20	87.99	87.36	92.47	73.41	83.82	88.34	91.42	73.28
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
FLUX-2-Pro	88.35	99.29	96.77	88.79	85.50	89.41	74.31	83.15	89.82	94.13	82.35
FLUX-2-Flex	89.35	98.59	97.10	90.41	86.74	92.09	74.77	82.47	90.85	92.23	88.24
Seedream-4.5	89.70	99.20	96.35	91.03	88.21	90.61	73.17	84.09	90.08	92.54	91.67
FLUX-2-Max	90.85	99.09	96.77	90.94	87.30	92.22	78.44	86.82	92.27	**95.26**	89.38
Imagen-4.0-Ultra	91.65	99.10	97.78	92.09	92.10	93.53	80.45	87.83	91.37	92.91	89.37
GPT-4o	92.48	98.98	98.22	94.01	90.78	94.33	83.79	91.21	92.89	91.35	89.24
Nano Banana Pro	92.72	**99.30**	97.47	91.95	91.38	95.43	80.24	89.59	92.91	93.28	95.65
GPT-4o-1.5	**95.77**	99.19	**99.20**	**96.33**	**94.84**	**96.94**	**88.76**	**92.27**	**98.17**	94.56	**97.39**
Open-source Models											
SDXL	40.22	87.45	72.28	44.66	35.10	46.37	10.34	48.48	26.68	30.80	0.00
MMaDA	41.35	82.40	56.65	48.93	37.83	50.25	17.95	55.75	32.35	30.22	1.15
Emu3	45.42	87.50	76.42	50.11	40.40	48.60	19.32	50.67	36.21	43.84	1.15
Kolors	46.07	84.40	77.22	54.17	48.00	52.79	19.77	46.66	33.63	42.91	1.15
Janus-Flow	47.10	86.34	62.98	49.20	43.57	51.45	22.41	62.80	46.49	45.76	0.00
Hunyuan-DiT	51.38	94.10	80.70	62.71	49.05	59.64	24.55	55.48	41.62	44.78	1.15
Janus	51.60	90.08	73.56	55.34	50.92	56.54	28.74	61.74	47.10	52.01	0.00
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	56.00	80.80	81.96	63.14	59.51	60.91	27.95	54.81	44.97	69.03	16.95
OneCAT	58.28	93.30	82.28	63.46	58.56	68.15	33.41	60.83	56.96	64.74	1.15
BLIP3-o	59.57	92.81	79.97	64.77	64.59	65.99	36.78	69.05	54.57	67.19	0.00
Infinity	59.81	90.80	87.97	68.06	60.17	69.16	31.36	60.16	51.42	66.60	12.36
Bagel	59.91	90.08	85.42	67.73	62.14	70.64	23.85	65.85	56.86	76.56	0.00
FLUX.1-dev	60.97	85.00	87.50	67.20	62.26	66.88	29.77	62.30	45.75	70.90	32.18
Janus-Pro	61.36	90.40	86.55	68.59	63.88	69.54	35.68	64.04	60.18	72.76	2.01
Show-o2	61.90	87.40	85.44	69.87	69.01	68.78	39.55	60.83	63.79	73.13	1.15
SD-3.5-Large	62.89	88.60	89.72	68.80	61.98	67.51	32.05	59.89	58.38	67.72	34.20
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
GLM-Image	67.23	84.10	90.82	69.12	60.93	68.15	31.65	64.04	54.38	72.95	76.15
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	71.36	92.30	93.67	73.40	72.53	74.24	40.45	62.43	60.31	77.61	66.67
Z-Image-Turbo	71.40	90.00	92.25	74.57	69.30	71.57	39.68	64.57	63.02	78.36	70.69
LongCat-Image	73.54	90.70	89.72	80.88	75.48	75.13	45.87	65.78	64.43	81.34	66.09
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
Z-Image	78.10	96.80	94.46	82.48	78.90	80.20	49.08	68.98	76.80	84.89	68.39
FLUX.2-Klein-9b	78.28	**97.50**	93.04	84.08	80.80	85.15	57.34	73.26	80.03	88.81	42.82
Qwen-Image	78.36	94.70	94.15	**87.93**	**82.60**	80.08	51.59	60.96	72.94	86.57	72.13
FLUX.2-Klein-base-9b	79.35	95.80	91.13	82.16	76.78	86.42	57.34	**77.51**	78.22	88.62	59.48
FLUX.2-dev	**84.76**	96.60	**95.41**	87.39	82.22	**87.31**	**62.84**	77.41	**83.51**	**89.55**	**85.34**

most balanced models, especially in grammar and logical reasoning. In contrast, several models such as Seedream-3.0 and Wan2.5 exhibit strong capabilities in style and world knowledge, but their performance drops notably on complex logical reasoning and relational understanding, indicating that robust high-level understanding remains a key differentiator among closed-source systems.

(b) *Open-source Models.* FLUX.2-dev is the strongest open-source model overall under English short prompts, with FLUX.2-Klein variants and Qwen-Image forming the next tier. The FLUX.2-Klein fam-

Table 3: **Overall Benchmarking Results of T2I models on UniGenBench++ using Chinese short prompts**. *Gemini-2.5-Pro* is used as the MLLM 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	59.73	89.55	91.36	67.87	64.52	72.15	31.54	62.02	51.33	65.53	1.45
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	78.86	93.80	93.04	83.97	76.33	84.14	63.99	72.45	78.74	76.12	65.98
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	80.45	98.95	96.32	88.31	86.03	90.87	77.26	83.90	86.09	89.75	7.06
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	83.08	99.20	97.63	91.13	93.54	92.89	79.55	88.64	89.95	91.04	7.18
FLUX-2-Pro	85.40	99.20	96.47	89.69	87.50	90.69	75.93	82.84	89.13	93.98	48.53
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
FLUX-2-Flex	87.62	98.09	95.99	90.76	89.67	91.57	77.08	85.68	92.09	94.54	60.77
FLUX-2-Max	88.14	99.10	97.28	92.26	90.55	94.26	80.00	87.57	93.65	**94.92**	51.76
Seedream-4.5	89.58	98.90	96.20	92.31	89.54	90.48	71.10	84.22	88.66	91.04	93.39
GPT-4o	91.02	99.39	98.72	94.99	92.34	95.77	**91.44**	91.02	93.91	89.27	63.37
Nano Banana Pro	93.82	**99.50**	97.47	94.55	94.96	96.07	82.34	89.04	94.20	94.40	**95.69**
GPT-4o-1.5	**95.62**	99.49	**99.68**	**96.55**	**95.52**	**97.83**	90.60	**91.98**	**97.13**	93.80	93.60
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
OneCAT	56.77	94.90	87.34	64.32	57.13	61.80	34.32	60.83	46.78	60.26	0.00
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	55.03	62.31	4.89
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
GLM-Image	70.57	85.80	90.51	71.15	65.11	69.29	42.89	63.37	57.86	74.07	**85.63**
Echo-4o	72.40	92.80	87.66	84.29	76.05	82.23	56.82	75.40	77.96	83.02	7.76
FLUX.2-Klein-base-9b	73.81	96.70	88.77	85.79	78.99	84.90	60.09	**77.94**	78.09	83.96	2.87
Z-Image-Turbo	74.18	91.70	90.98	76.92	74.71	72.08	50.69	65.51	65.85	80.97	72.41
FLUX.2-Klein-9b	75.19	**98.60**	93.67	86.11	83.08	86.68	58.03	77.01	82.35	84.89	1.44
LongCat-Image	75.97	87.60	92.09	79.17	77.00	79.95	49.31	65.64	66.62	79.29	83.05
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
FLUX.2-dev	81.44	95.70	93.20	90.49	**87.55**	**89.34**	**68.35**	76.20	**84.02**	**90.49**	39.08
Z-Image	**81.69**	96.30	**94.62**	86.11	82.60	84.64	54.82	71.26	79.51	86.57	80.46

ily tends to be particularly strong in relation modeling and grammar consistency, while Qwen-Image is competitive in attribute/action understanding and layout-related dimensions but still trails the best open-source models on grammar and logical reasoning. Other models such as Lumina-DiMOO, HiDream-I1-Full, and Echo-4o show strengths on specific dimensions (*e.g.,* relation/world knowledge), yet remain less stable on logical consistency and compositional robustness.

(c) *Closed- v.s. Open-source Models.* A clear trend emerges where open-source models are making meaningful progress in closing the gap with closed-source systems. The best open-source model in this setting (FLUX.2-dev) can match or surpass several mid-tier closed-source models and is already competitive on many visual-understanding dimensions. However, closed-source models remain clearly ahead at the frontier (*e.g.,* GPT-4o-1.5 and Nano Banana Pro), especially on grammar consistency, logical reasoning, and robust compositional generation. Overall, the gap is now more concentrated on high-level reasoning and consistency rather than purely visual fidelity.

Table 4: **Overall Benchmarking Results of T2I models on UniGenBench++ using English long prompts**. *Gemini-2.5-Pro* is used as the MLLM 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	84.56	96.50	96.24	91.17	78.98	87.01	72.28	77.68	86.22	87.26	72.28
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
FLUX-2-Pro	90.10	99.08	96.89	94.37	84.38	90.86	80.15	87.83	90.81	93.98	82.69
FLUX-2-Flex	90.43	98.73	97.02	94.55	85.74	90.33	75.74	86.78	91.20	93.82	90.38
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
Seedream-4.5	91.38	98.67	96.24	96.15	88.20	89.92	83.09	86.80	89.57	93.33	91.85
FLUX-2-Max	92.18	99.24	96.73	94.78	86.97	92.90	83.00	89.40	92.84	95.02	90.93
GPT-4o	92.63	99.08	97.95	93.53	87.78	91.13	**91.02**	94.46	93.99	93.59	83.79
Nano Banana Pro	94.20	**99.58**	97.83	95.94	89.19	94.29	87.75	93.15	94.10	93.73	**96.47**
GPT-4o-1.5	**95.41**	**99.58**	**98.98**	**97.20**	**92.90**	**95.79**	90.15	**94.84**	**96.45**	**96.70**	91.46
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	62.80	94.93	83.96	74.98	59.41	65.46	47.55	62.18	62.97	74.37	2.17
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
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	64.53	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
GLM-Image	75.48	87.38	93.93	82.55	67.77	73.87	51.47	71.83	67.71	84.37	73.91
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
Z-Image-Turbo	80.72	93.19	93.93	89.34	74.20	80.44	66.18	76.65	76.46	86.67	70.11
LongCat-Image	81.28	92.11	93.50	90.01	77.69	81.30	66.91	75.89	79.15	87.22	69.02
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
FLUX.2-Klein-9b	85.06	98.67	94.65	94.11	82.40	89.92	75.25	86.68	88.70	**93.17**	47.01
FLUX.2-Klein-base-9b	86.45	97.92	95.38	92.79	80.83	88.85	77.94	**87.82**	88.48	92.78	61.68
Z-Image	86.77	97.26	94.36	93.25	83.72	89.06	76.72	80.46	86.52	91.11	75.27
FLUX.2-dev	**90.31**	**99.17**	**96.39**	**94.57**	**86.17**	**91.70**	**79.90**	84.52	**90.16**	92.22	**88.32**

4.2.2 English Long Prompt (Tab. 4)

(a) *Closed-source Models.* For English long prompt generation, closed-source models remain strong across nearly all evaluation dimensions. GPT-4o-1.5 reaches the best overall performance, while Nano Banana Pro and GPT-4o also remain top-tier, with particularly strong results in logical reasoning and grammar consistency. Seedream-4.5 further stands out in text generation quality. Models such as Imagen-4.0-Ultra

and Wan2.5 can be competitive on visual semantics and layout-related dimensions, yet they generally lag behind the strongest models on the most demanding reasoning- and grammar-focused evaluations.

(b) *Open-source Models* Open-source models also show substantial progress. FLUX.2-dev leads the open-source group and is competitive across most dimensions under long prompts. Z-Image and the FLUX.2-Klein variants form a strong second tier, typically showing better relation modeling, grammar, and logical reasoning than earlier open-source baselines. Qwen-Image and Hunyuan-Image-2.1 provide solid overall performance with strong world knowledge and text-related capabilities, while some models (*e.g.,* Echo-4o) may remain unstable on text generation under long prompts.

(c) *Closed- v.s. Open-source Models.* Compared to English short prompts, the closed- vs. open-source gap is noticeably smaller under long prompts. The best open-source models can approach the performance of top closed-source models on several dimensions, particularly for attribute/action understanding and layout. Nevertheless, closed-source systems still maintain an advantage on the most challenging aspects, including logical reasoning and grammar consistency, and they provide more reliable performance across a wider range of long-prompt compositions.

4.2.3 Chinese Short Prompt (Tab. 3)

(a) *Closed-source Models.* For Chinese short prompts, GPT-4o-1.5 achieves the strongest overall performance, followed by Nano Banana Pro and GPT-4o. Notably, Nano Banana Pro and Seedream-4.0 excel in Chinese text rendering, while GPT-4o-1.5 remains the most balanced across reasoning- and composition-heavy dimensions. We also observe that some models (*e.g.,* Imagen-4.0-Ultra) can score very high on action/relation/logic-related aspects but still struggle with Chinese text generation, highlighting that accurate multilingual text rendering remains a non-trivial bottleneck.

(b) *Open-source Models.* Among open-source models, Z-Image, FLUX.2-dev, and Qwen-Image form the leading group for Chinese short prompts. GLM-Image is particularly strong on Chinese text generation, but is less balanced on other dimensions compared to the top overall open-source models. Other models (*e.g.,* OmniGen2, Bagel) show relatively balanced performance but still lag behind the leaders on complex reasoning, composition, and grammar consistency. The remaining models, such as X-Omni and 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 dominate the overall evaluation, with GPT-4o-1.5 and Nano Banana Pro setting a strong upper bound across nearly all dimensions. Open-source models have progressed substantially, with the leading group (*e.g.,* Z-Image, FLUX.2-dev, Qwen-Image) showing competitive performance on many non-text visual dimensions. However, a clear gap remains in Chinese grammar consistency, complex compositional reasoning, and stable Chinese text rendering across diverse scenarios.

4.2.4 Chinese Long Prompt (Tab. 5)

(a) *Closed-source Models.* Closed-source models still demonstrate strong overall performance in generating Chinese long prompts. GPT-4o-1.5 achieves the best overall score, with Nano Banana Pro and Seedream-4.5 also ranking among the top models. Seedream-4.0 remains highly competitive, and its overall performance is close to GPT-4o, largely supported by its strong Chinese text generation. Similar to the short-prompt setting, some models can be strong on reasoning- and relation-related dimensions yet remain weak in Chinese text rendering, indicating that long-form multilingual text generation is still challenging.

(b) *Open-source Models.* For open-source models, Z-Image leads the group on Chinese long prompts, followed by Hunyuan-Image-2.1 and Qwen-Image, which show strong performance in attribute/layout/text-related dimensions. We also observe a clear split in Chinese text capability: some models (*e.g.,* GLM-Image, Qwen-Image, Hunyuan-Image-2.1) can generate Chinese text reliably, while several others (including some FLUX.2 variants) remain weak on the text dimension, even when their non-text visual understanding is strong.

(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

Table 5: **Overall Benchmarking Results of T2I models on UniGenBench++ using Chinese long prompts**. *Gemini-2.5-Pro* is used as the MLLM 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
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	84.36	97.42	94.15	91.04	77.75	87.23	73.28	81.09	85.53	89.01	67.12
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
FLUX-2-Pro	87.11	98.83	95.91	94.66	86.00	92.42	79.26	86.47	91.96	93.12	52.50
FLUX-2-Flex	89.19	98.33	96.78	95.71	87.60	92.84	81.73	86.98	92.11	95.03	64.80
FLUX-2-Max	89.80	99.25	97.37	96.12	88.05	94.54	85.96	90.72	93.21	94.97	57.78
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
Seedream-4.5	93.12	99.00	97.83	96.49	90.55	92.29	86.76	89.96	90.88	94.20	**93.21**
Nano Banana Pro	95.42	**99.42**	98.84	97.14	92.97	95.64	91.91	93.27	95.85	96.27	92.93
GPT-4o-1.5	**96.12**	98.73	**99.27**	**98.18**	**94.31**	**96.79**	**94.36**	**96.01**	**98.08**	**96.47**	89.01
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	23.09	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	60.03	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
OneCAT	61.40	96.01	80.35	72.01	56.90	61.85	49.76	63.20	58.50	73.49	1.90
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
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
GLM-Image	79.11	89.62	93.35	83.92	71.78	77.16	58.09	73.48	74.85	85.95	82.88
FLUX.2-Klein-base-9b	81.41	97.67	92.63	94.09	82.76	90.95	**80.64**	86.42	90.60	92.38	5.98
FLUX.2-Klein-9b	81.74	**99.09**	92.92	94.03	83.20	90.73	80.15	**86.55**	**91.69**	93.33	5.71
LongCat-Image	83.14	90.20	93.35	90.96	81.11	82.60	65.69	77.79	81.27	86.35	82.07
Z-Image-Turbo	83.69	96.26	94.80	90.96	78.74	84.38	70.83	78.30	80.03	87.94	74.73
FLUX.2-dev	86.12	98.42	95.52	**95.29**	**88.46**	**92.40**	79.66	84.26	89.50	**94.44**	43.21
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
Z-Image	**89.17**	97.67	95.52	94.32	86.13	89.12	79.90	81.73	86.15	92.62	**88.59**

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 benchmarking results for 27 dimensions are provided in Tabs. 7, 8, 9, and 10.

4.3 Offline Evaluation Model

Existing benchmarks Niu et al. (2025); Wei et al. (2025) typically use Vision-Language Models (VLMs) like Qwen2.5-VL-72b Bai et al. (2025) 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 like grammar-consistency and action-contact. To address this, we train 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 cases are provided in Fig. 4 (c).

5 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 Multi-modal Large Language Model (MLLM), we develop 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

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

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025.

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

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

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

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

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *NeurIPS*, 34: 19822–19835, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

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

Zigang Geng, Yibing Wang, Yeyao Ma, et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *NeurIPS*, 36:52132–52152, 2023.

Google. Gemini2.5-pro. *https://deepmind.google/models/gemini/pro/*, 2025a.

Google. Imagen. *https://deepmind.google/models/imagen/*, 2025b.

Google. Nano banana. *https://deepmind.google/models/gemini/image/*, 2025c.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *CVPR*, pp. 15733–15744, 2025.

Hidream. Hidream-v2l. *https://hidreamai.com/studio*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.

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

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *NeurIPS*, 36:78723–78747, 2023.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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

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

Black Forest Labs. Flux.1 krea. *https://www.krea.ai/apps/image/flux-krea*, 2025.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024a.

Han Li, Xinyu Peng, Yaoming Wang, Zelin Peng, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, Wenrui Dai, and Hongkai Xiong. Onecat: Decoder-only auto-regressive model for unified understanding and generation. *arXiv preprint arXiv:2509.03498*, 2025.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024b.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Yuxuan Luo, Yuhui Yuan, Junwen Chen, Haonan Cai, Ziyi Yue, Yuwei Yang, Fatima Zohra Daha, Ji Li, and Zhouhui Lian. Mmmg: A massive, multidisciplinary, multi-tier generation benchmark for text-to-image reasoning. In *NeurIPS*, 2025.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, 2024.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, Xunliang Cai, Yayong Guan, and Jie Hu. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

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

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36:53728–53741, 2023.

Recraft. Recraft. *https://www.recraft.ai*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Runway. Runway-gen4. *https://docs.dev.runwayml.com*, 2025.

Seedream Team, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Stability. Stable image ultra. *https://platform.stability.ai/*, 2025.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, and Xihui Liu. T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation. *arXiv preprint arXiv:2508.17472*, 2025.

Tencent. Hunyuan-image-2.1. *https://github.com/Tencent-Hunyuan/HunyuanImage-2.1*, 2025.

Chengzhuo Tong, Ziyu Guo, Renrui Zhang, Wenyu Shan, Xinyu Wei, Zhenghao Xing, Hongsheng Li, and Pheng-Ann Heng. Delving into rl for image generation with cot: A study on dpo vs. grpo. *arXiv preprint arXiv:2505.17017*, 2025.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 30, 2017.

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

Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*, 2024b.

Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*, 2025a.

Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025b.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025c.

Yibin Wang, Weizhong Zhang, Xu Honghui, and Cheng Jin. High fidelity scene text synthesis. In *CVPR*, 2025d.

Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. Tiif-bench: How does your t2i model follow your instructions? *arXiv preprint arXiv:2506.02161*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *CVPR*, pp. 12966–12977, 2025b.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025c.

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

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025.

Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025.

Z-Image Team. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Z.ai. Glm-image. *https://z.ai/blog/glm-image*, 2026.

Yujie Zhou, Pengyang Ling, Jiazi Bu, Yibin Wang, Yuhang Zang, Jiaqi Wang, Li Niu, and Guangtao Zhai. Ggrpo: Granular grpo for precise reward in flow models. *arXiv preprint arXiv:2510.01982*, 2025.

Table 6: **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."*
Creative Divergence	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'."*
Art	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."*
Art	Sculpture	*"A giant elephant sculpture carved from transparent crystal is crystal clear and stands quietly in the center of the museum."*
Art	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'."*
Illustration	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."*
Film & Story	Science Fiction	*"An astronaut wearing a spacesuit holds a pyramidal holographic projector in his hand, projecting an image of the earth."*
Film & Story	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."*
Design	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."*
Design	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."*
Design	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."*
Design	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."*
Design	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."*
Design	Logo / Icon Design	*"A logo design has two similar mechanical phoenixes symmetrical left and right, with the same metallic texture in the middle."*
Design	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	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 7: **Detailed Benchmarking Results of T2I models on UniGenBench++ using English short prompts**. *Gemini-2.5-Pro* is used as the MLLM 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	79.69	66.92	55.73	73.90	56.74	66.92	77.49	71.97	42.05	13.83
DALL-E-3	68.85	94.43	92.64	60.14	63.16	87.20	84.72	66.25	91.60	60.78	76.67	77.94	68.72	63.19	76.19	82.99	71.51	85.47	66.93	78.01	63.95	76.34	72.09	59.45	54.78	77.25	46.22	24.43
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	70.46	90.99	91.30	72.92	60.65	79.25	75.00	78.12	98.33	58.97	69.02	76.47	78.06	65.48	77.83	81.08	74.44	80.98	71.88	77.30	58.85	83.46	65.74	54.23	81.25	79.92	41.46	37.64
Imagen-3.0	71.34	89.35	93.95	71.09	64.00	85.85	89.78	64.38	93.28	75.00	83.89	80.15	75.65	71.43	85.29	83.22	76.14	88.27	83.06	80.36	65.10	80.88	70.28	57.94	82.35	77.65	45.09	22.70
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	77.69	91.90	95.73	77.08	75.00	83.02	89.58	80.00	96.67	76.92	84.24	83.09	76.02	75.60	84.91	84.12	75.56	87.50	82.03	78.32	66.93	83.82	77.31	69.23	88.97	84.47	56.82	50.29
Wan2.5	77.87	92.64	94.75	75.00	70.51	91.04	83.09	78.75	88.33	59.87	74.46	77.94	76.04	72.02	81.60	85.47	74.44	81.52	85.16	78.09	72.77	83.70	72.69	61.54	75.74	78.03	55.50	73.12
Seedream-3.0	78.41	98.19	94.90	79.02	81.94	89.62	83.80	77.22	96.67	75.97	89.56	86.03	75.38	81.93	89.10	81.57	74.16	83.61	80.47	76.92	67.62	77.94	68.40	35.14	88.15	89.35	51.83	69.86
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	83.33	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
Nano Banana	87.29	98.59	96.20	86.43	80.77	88.46	95.83	80.77	98.33	80.13	93.48	88.24	83.67	80.95	95.28	93.49	86.67	94.02	**96.09**	90.21	86.46	90.44	83.33	77.31	93.01	89.77	73.41	73.28
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
FLUX-2-Pro	88.35	99.29	96.77	84.72	75.00	96.23	90.28	86.25	99.17	76.92	92.78	80.88	87.76	80.36	90.57	90.88	82.22	93.33	90.62	92.86	86.72	90.38	83.33	75.77	92.05	**96.21**	74.31	82.35
FLUX-2-Flex	89.35	98.59	97.10	88.19	79.05	95.75	92.36	86.88	**100.00**	78.21	90.00	87.50	83.67	84.52	94.34	92.23	87.22	96.11	92.97	92.60	89.06	90.77	79.17	76.92	90.15	94.32	74.77	88.24
Seedream-4.5	89.70	99.20	96.35	87.50	87.82	97.64	86.81	85.62	**100.00**	80.77	90.22	91.91	84.69	86.90	93.87	92.57	85.00	94.57	88.28	90.05	90.10	90.07	85.65	76.54	91.54	93.56	73.17	91.67
FLUX-2-Max	90.85	99.09	96.77	90.28	77.70	97.64	93.06	86.79	99.17	82.69	93.30	86.03	85.20	84.52	90.57	94.93	84.44	95.56	92.19	95.15	89.32	90.77	87.50	82.31	**97.35**	93.16	78.44	89.38
Imagen-4.0-Ultra	91.65	99.10	97.78	**94.44**	80.77	95.28	94.44	88.75	**100.00**	89.74	93.41	93.38	88.78	87.50	**98.58**	96.28	87.78	96.20	91.41	92.86	89.84	91.91	90.28	81.54	93.75	92.05	80.45	89.37
GPT-4o	92.48	98.98	98.22	89.29	**96.00**	94.66	92.96	92.50	99.17	88.46	93.33	87.88	**92.02**	89.16	92.31	96.58	91.11	94.89	92.97	94.07	91.67	91.04	**93.06**	**89.75**	92.16	90.53	83.79	89.24
Nano Banana Pro	92.72	**99.30**	97.47	90.28	85.53	97.64	93.75	85.00	99.17	89.47	91.11	90.44	89.80	94.05	92.92	**96.96**	96.11	92.39	95.31	95.15	90.62	94.49	87.96	85.71	92.65	93.94	80.24	95.65
GPT-4o-1.5	**95.77**	99.19	**99.20**	90.71	92.31	**99.03**	**97.92**	**97.50**	**100.00**	**95.51**	**95.63**	**95.59**	91.15	**97.02**	94.81	**96.96**	**96.67**	**99.44**	93.75	**98.45**	**97.87**	**97.77**	90.28	88.10	93.31	95.83	**88.76**	**97.39**
Open-source Models																														
SDXL	40.22	87.45	72.28	41.67	25.00	54.90	44.85	36.11	68.52	19.74	38.10	45.31	26.74	24.34	52.40	55.38	41.22	38.75	43.33	33.75	19.94	54.58	41.67	47.46	25.00	36.40	10.34	0.00
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
Emu3	45.42	87.50	76.42	42.36	45.51	52.83	40.28	46.25	77.50	23.08	49.46	54.41	34.69	29.17	50.47	55.41	44.44	46.74	41.41	41.33	30.99	58.09	49.07	44.23	42.28	45.45	19.32	1.15
Kolors	46.07	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
Janus-Flow	47.10	86.34	62.98	43.18	30.77	55.39	57.35	33.33	82.41	22.37	48.81	57.81	38.95	36.84	54.81	62.69	36.49	53.75	42.50	60.00	33.63	70.00	51.11	64.41	46.82	44.74	22.41	0.00
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
Janus	51.60	90.08	73.56	35.61	37.82	60.29	66.18	48.61	90.74	31.58	52.38	62.50	50.00	39.47	65.87	58.85	52.70	61.25	50.00	59.38	35.42	70.00	52.22	60.59	51.82	52.19	28.74	0.00
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	56.00	80.80	81.96	70.83	46.79	55.66	68.75	58.75	87.50	57.69	59.78	69.85	52.55	53.57	65.09	58.11	60.00	66.30	60.94	49.23	40.62	69.49	54.17	40.00	76.84	60.98	27.95	16.95
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
BLIP3-o	59.57	92.81	79.97	48.48	60.26	66.67	76.47	56.94	83.33	57.24	71.43	71.09	63.95	50.66	71.15	70.77	57.43	66.25	65.83	64.06	45.54	81.67	61.11	62.29	69.55	64.91	36.78	0.00
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
Bagel	59.91	90.08	85.42	56.82	50.00	73.53	77.94	59.03	94.44	51.32	64.88	67.19	64.53	56.58	66.83	77.31	68.92	70.00	59.17	67.50	46.73	74.17	64.44	58.47	77.73	75.44	23.85	0.00
FLUX.1-dev	60.97	85.00	87.50	71.53	51.92	58.96	74.31	65.62	90.00	50.00	69.02	69.12	60.20	61.90	63.21	66.89	65.56	72.83	60.16	46.17	45.31	76.47	61.57	48.08	74.63	67.05	29.77	32.18
Janus-Pro	61.36	90.40	86.55	56.25	57.69	74.06	73.61	61.88	90.83	47.44	65.22	72.79	60.71	59.52	75.47	76.01	58.33	73.91	64.06	67.35	52.86	76.10	64.81	50.77	74.63	70.83	35.68	2.01
Show-o2	61.90	87.40	85.44	59.03	64.10	70.75	74.31	61.25	95.00	54.49	75.00	75.00	72.45	50.60	82.08	76.35	60.56	71.20	59.38	66.84	60.68	77.57	63.43	41.15	75.37	70.83	39.55	1.15
SD-3.5-Large	62.89	88.60	89.72	69.44	51.28	70.28	70.83	64.38	91.67	57.69	63.04	62.50	59.69	58.93	68.40	73.99	65.00	66.30	57.81	68.37	48.18	77.21	60.19	41.54	70.96	64.39	32.05	34.20
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
GLM-Image	67.23	84.10	90.82	76.39	58.97	74.06	71.53	48.12	90.00	58.97	66.30	65.44	54.08	52.98	67.45	67.91	67.22	67.93	70.31	53.06	55.73	76.84	58.33	55.38	79.78	65.91	31.65	76.15
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	71.36	92.30	93.67	73.61	61.54	72.17	79.17	62.50	98.33	60.90	76.09	74.26	73.98	68.45	78.77	76.69	67.78	78.26	71.88	61.99	58.59	81.62	63.89	41.15	82.72	72.35	40.45	66.67
Z-Image-Turbo	71.40	90.00	92.25	75.00	58.97	79.25	77.78	64.38	95.83	62.82	73.37	78.68	69.90	61.31	70.28	75.68	65.00	75.54	65.62	64.29	61.72	79.78	62.04	50.77	83.09	73.48	39.68	70.69
LongCat-Image	73.54	90.70	89.72	74.31	79.49	87.74	81.94	67.50	95.00	64.10	82.61	82.35	71.94	72.02	79.25	75.34	70.56	83.15	69.53	68.11	60.68	76.10	59.72	60.00	84.93	77.65	45.87	66.09
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
Z-Image	78.10	96.80	94.46	81.25	69.87	91.98	81.25	73.12	97.50	74.36	82.61	82.35	77.04	71.43	84.43	84.80	70.00	86.41	75.00	82.14	71.35	80.88	75.00	51.54	86.76	82.95	49.08	68.39
FLUX.2-Klein-9b	78.28	**97.50**	93.04	75.00	83.33	89.62	82.64	76.88	97.50	**78.21**	85.33	83.82	79.08	73.21	84.43	88.51	72.22	**95.65**	80.47	80.87	79.17	81.99	76.39	61.54	90.81	86.74	57.34	42.82
Qwen-Image	78.36	94.70	94.15	84.03	85.26	91.98	86.11	81.88	99.17	**78.21**	**86.96**	**86.76**	77.55	76.79	**88.68**	82.09	71.11	86.96	78.12	73.21	72.66	84.93	70.37	28.08	87.13	85.98	51.59	72.13
FLUX.2-Klein-base-9b	79.35	95.80	91.13	72.92	76.28	89.15	81.94	75.62	97.50	71.15	78.80	80.15	78.06	68.26	82.55	88.85	**77.78**	92.93	83.59	79.34	77.08	85.29	77.67	69.23	**91.54**	85.61	57.34	59.48
FLUX.2-dev	**84.76**	96.60	**95.41**	73.61	73.72	**96.23**	**91.67**	**88.12**	**100.00**	74.36	86.41	83.82	**82.14**	**80.95**	84.43	**93.24**	**77.78**	88.59	**85.16**	**84.95**	**82.03**	**88.60**	**78.70**	64.62	87.87	**91.29**	**62.84**	**85.34**

Table 8: **Detailed Benchmarking Results of T2I models on UniGenBench++ using English long prompts**. *Gemini-2.5-Pro* is used as the MLLM 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	84.56	96.50	96.24	85.64	79.61	93.73	88.36	87.68	96.11	78.21	82.91	78.68	74.07	72.13	81.50	86.03	79.17	94.77	88.35	87.61	83.18	93.25	75.81	65.49	88.42	85.77	72.28	72.28
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
FLUX-2-Pro	90.10	99.08	96.89	86.70	86.93	96.67	92.94	91.86	97.41	81.00	85.76	83.70	82.35	80.00	87.06	89.92	87.70	95.35	91.26	91.67	88.92	94.49	87.90	82.14	96.14	91.19	80.15	82.69
FLUX-2-Flex	90.43	98.73	97.02	90.43	88.65	95.15	93.65	91.33	97.93	85.67	86.86	84.78	83.71	82.06	87.36	88.72	88.20	93.60	92.23	91.95	89.52	96.61	82.66	82.14	93.86	93.76	75.74	90.38
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
Seedream-4.5	91.38	98.67	96.24	90.43	91.38	95.56	94.89	**96.50**	**99.16**	83.97	90.31	88.73	87.89	83.00	90.57	88.39	84.89	95.98	91.50	89.71	89.25	97.62	88.89	75.35	93.36	93.28	83.09	91.85
FLUX-2-Max	92.18	99.24	96.73	88.30	86.06	96.36	94.06	93.11	98.25	84.67	88.78	88.77	84.23	84.12	88.17	92.31	90.58	96.51	92.72	93.04	92.38	96.61	88.31	84.29	95.57	94.32	83.00	90.93
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
Nano Banana Pro	94.20	**99.58**	97.83	89.36	90.69	**97.52**	**96.97**	91.43	98.53	86.22	89.69	90.94	89.29	89.37	89.38	94.39	91.99	98.28	92.48	94.92	92.29	**99.60**	90.08	90.14	93.79	93.66	87.75	**96.47**
GPT-4o-1.5	**95.41**	**99.58**	**98.98**	**93.41**	**95.19**	97.25	95.39	95.98	99.15	**92.33**	**93.99**	**95.29**	**91.71**	**92.51**	**92.46**	**96.11**	**94.16**	**99.71**	**93.15**	**97.54**	**93.98**	95.16	**95.24**	94.18	**96.71**	**96.69**	90.15	91.46
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	48.30	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	62.80	94.93	83.96	61.70	67.92	77.48	83.14	62.06	78.83	38.46	61.56	63.77	49.11	45.98	70.93	68.11	62.18	63.79	64.32	72.35	42.29	73.81	64.68	49.65	75.56	72.83	47.55	2.17
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
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	60.31	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
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
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
GLM-Image	75.48	87.38	93.93	79.26	71.11	83.66	85.04	74.83	88.05	70.83	68.12	65.58	63.84	58.33	71.43	69.01	68.27	82.76	79.85	69.07	64.72	87.30	69.05	60.56	86.44	81.70	51.47	73.91
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
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	88.04	65.53	41.03
Z-Image-Turbo	80.72	93.19	93.93	82.98	76.11	91.72	87.50	80.77	96.38	75.64	74.06	71.01	71.43	66.38	77.98	78.32	73.08	87.93	83.74	77.75	73.60	91.27	69.84	69.72	87.71	85.33	66.18	70.11
LongCat-Image	81.28	92.11	93.50	77.13	82.22	91.79	88.64	80.94	96.07	73.72	73.44	79.35	74.11	66.95	84.33	79.85	75.00	89.08	82.28	79.98	77.34	91.27	71.83	65.85	89.55	84.24	66.91	69.02
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
FLUX.2-Klein-9b	85.06	98.67	94.65	80.85	88.06	95.56	91.48	89.69	**98.74**	80.13	81.56	83.70	78.57	76.44	85.91	90.69	83.33	94.83	89.32	89.51	86.92	93.65	**86.11**	80.99	**93.22**	**93.12**	75.25	47.01
FLUX.2-Klein-base-9b	86.45	97.92	95.38	79.79	86.67	94.20	90.53	87.24	97.69	81.73	80.94	80.43	76.34	73.56	84.13	88.14	87.50	**95.69**	85.44	88.98	87.38	93.65	85.71	**84.51**	92.66	92.93	77.94	61.68
Z-Image	86.77	97.26	94.36	85.11	87.08	94.95	90.15	87.41	97.80	81.73	79.38	83.33	**86.16**	77.87	87.30	88.52	84.94	93.39	89.56	86.12	87.38	91.67	80.16	70.77	92.51	89.31	76.72	75.27
FLUX.2-dev	**90.31**	**99.17**	**96.39**	82.98	88.47	**95.78**	**92.42**	**91.43**	98.69	**84.94**	**85.94**	**85.51**	**86.16**	**82.47**	**88.10**	**91.33**	**89.42**	**95.69**	**90.78**	**89.94**	**90.65**	**94.84**	82.94	76.76	92.94	91.30	**79.90**	**88.32**

Table 9: **Detailed Benchmarking Results of T2I models on UniGenBench++ using Chinese short prompts**. *Gemini-2.5-Pro* is used as the MLLM 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	59.73	89.55	91.36	71.43	43.59	68.14	72.86	63.87	94.17	47.44	66.85	70.45	67.71	58.33	73.56	80.56	63.89	76.67	58.06	59.47	43.01	72.69	68.75	45.70	64.77	66.29	31.54	1.45
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	78.86	93.80	93.04	79.86	75.64	91.04	84.72	75.62	97.50	72.44	76.09	81.62	72.45	75.00	80.66	83.78	75.56	88.59	90.62	84.69	72.66	83.09	68.52	64.45	77.94	74.24	63.99	65.98
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	80.45	98.95	96.32	83.09	82.78	91.13	95.74	80.13	98.33	83.33	89.14	89.71	78.87	82.63	92.61	90.94	83.33	94.54	96.09	88.53	83.68	89.18	85.17	77.34	92.19	87.21	77.26	7.06
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	83.08	99.20	97.63	89.58	80.13	93.40	94.44	90.62	**100.00**	94.87	91.85	96.32	88.78	93.45	**96.70**	91.89	87.22	98.37	95.31	94.90	84.90	94.85	87.96	82.69	92.65	89.39	79.55	7.18
FLUX-2-Pro	85.40	99.20	96.47	84.72	77.63	97.17	89.58	88.12	**100.00**	83.33	90.56	88.97	88.27	80.95	91.51	91.89	82.22	96.67	91.41	92.35	85.83	91.29	84.72	72.69	93.28	**94.70**	75.93	48.53
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
FLUX-2-Flex	87.62	98.09	95.99	87.50	80.26	95.28	93.06	88.68	**100.00**	89.74	92.18	88.24	87.76	82.63	95.73	93.58	86.59	94.44	89.84	94.90	89.18	94.32	85.65	76.92	94.40	94.68	77.08	60.77
FLUX-2-Max	88.14	99.10	97.28	90.28	80.26	97.64	95.14	90.00	**100.00**	89.10	94.44	90.44	89.80	85.71	92.92	95.95	88.89	98.89	91.41	95.41	91.84	94.32	88.89	79.62	**96.64**	93.18	80.00	51.76
Seedream-4.5	89.58	98.90	96.20	87.50	87.82	**99.53**	89.58	88.12	**100.00**	85.26	94.02	91.91	82.14	86.31	**96.70**	88.85	89.44	94.02	90.62	91.84	85.42	90.81	84.26	77.31	91.54	90.53	71.10	93.39
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
Nano Banana Pro	93.82	**99.50**	97.47	90.97	**96.15**	95.75	95.14	91.25	98.33	94.23	94.57	**97.06**	92.35	**95.24**	**96.70**	96.96	91.67	97.83	**97.66**	96.68	91.67	94.49	90.74	81.92	96.32	92.42	82.34	**95.69**
GPT-4o-1.5	**95.62**	99.49	**99.68**	92.14	94.23	98.08	**99.31**	**95.62**	**100.00**	**96.15**	**98.91**	96.32	**93.81**	92.86	95.28	**97.97**	**97.22**	**100.00**	95.31	**99.23**	**94.95**	**95.90**	**92.13**	87.70	93.28	94.32	90.60	93.60
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	45.83	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	56.63	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
OneCAT	56.77	94.90	87.34	62.50	71.79	68.40	63.89	36.88	86.67	37.18	69.02	76.47	57.14	39.29	63.21	68.58	57.78	60.33	53.91	58.16	35.16	66.91	62.50	53.08	63.24	57.20	34.32	0.00
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	59.69	52.34	76.10	70.37	48.46	73.53	64.77	39.09	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	57.74	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
GLM-Image	70.57	85.80	90.51	77.08	63.46	74.53	73.61	51.88	90.83	66.03	71.74	66.91	56.63	57.14	71.70	70.95	68.33	69.57	66.41	62.50	53.12	75.00	62.50	51.92	79.04	68.94	42.89	**85.63**
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
FLUX.2-Klein-base-9b	73.81	96.70	88.77	75.00	79.49	91.51	87.50	81.88	**100.00**	72.44	82.07	88.97	70.92	77.38	83.49	88.85	79.44	88.04	78.91	83.67	72.40	84.56	81.02	68.46	88.24	79.55	60.09	2.87
Z-Image-Turbo	74.18	91.70	90.98	75.69	66.03	88.21	77.78	60.00	94.17	71.15	79.89	80.15	69.39	72.02	76.42	75.00	61.11	77.17	73.44	69.39	62.24	79.04	64.35	52.31	82.72	79.17	50.69	72.41
FLUX.2-Klein-9b	75.19	**98.60**	93.67	75.69	81.41	93.40	86.11	80.00	**100.00**	76.28	86.41	88.97	78.57	80.36	87.74	90.88	75.56	92.93	**83.59**	**87.24**	77.34	86.76	79.17	65.00	88.60	81.06	58.03	1.44
LongCat-Image	75.97	87.60	92.09	71.53	77.56	88.21	77.78	63.75	96.67	76.92	78.80	85.29	71.43	67.26	83.02	80.41	70.56	86.96	82.03	69.13	64.06	79.78	63.43	52.69	78.31	80.30	49.31	83.05
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
FLUX.2-dev	81.44	95.70	93.20	86.81	83.97	**96.23**	89.58	**86.25**	**100.00**	**87.18**	**91.30**	87.50	**82.14**	**86.90**	**90.09**	**94.26**	**82.78**	**93.48**	81.25	86.73	**81.25**	**90.81**	**82.41**	55.77	**91.54**	**89.39**	**68.35**	39.08
Z-Image	**81.69**	96.30	**94.62**	83.33	74.36	95.28	85.42	79.38	98.33	81.41	85.33	83.82	81.63	76.19	86.32	88.51	75.00	90.22	81.25	83.16	75.78	84.19	73.61	55.77	86.76	86.36	54.82	80.46

Table 10: **Detailed Benchmarking Results of T2I models on UniGenBench++ using Chinese long prompts**. *Gemini-2.5-Pro* is used as the MLLM 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
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	84.36	97.42	94.15	82.98	82.72	92.22	91.79	87.59	94.96	71.15	75.32	83.46	75.91	73.28	80.98	85.97	79.17	91.95	91.75	86.65	83.02	92.46	82.54	69.72	89.63	88.22	73.28	67.12
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
FLUX-2-Pro	87.11	98.83	95.91	83.51	86.16	97.27	93.90	90.99	98.42	83.44	87.34	86.76	85.14	82.85	87.43	91.71	88.60	96.22	93.45	93.09	89.49	95.49	83.73	81.07	94.18	91.76	79.26	52.50
FLUX-2-Flex	89.19	98.33	96.78	90.96	88.70	96.89	93.14	94.14	99.15	84.09	87.90	88.97	89.14	85.17	88.71	91.56	89.87	94.48	96.12	92.88	90.42	95.49	84.92	81.43	95.60	94.29	81.73	64.80
FLUX-2-Max	89.80	99.25	97.37	89.36	89.55	96.97	95.06	96.13	98.95	87.99	88.61	89.34	89.73	84.88	88.25	94.26	90.65	97.09	95.87	94.16	91.12	96.31	90.87	85.71	94.89	95.07	85.96	57.78
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
Seedream-4.5	93.12	99.00	97.83	89.89	91.81	97.06	95.08	96.14	99.00	87.82	90.31	93.48	89.24	87.07	92.15	90.18	88.10	97.13	95.38	90.67	91.36	98.81	90.84	81.34	95.76	92.20	86.76	**93.21**
Nano Banana Pro	95.42	**99.42**	98.84	89.36	94.31	**98.27**	96.78	94.23	99.16	93.59	92.81	97.46	91.52	91.95	**92.26**	95.79	93.91	97.99	94.66	95.87	95.79	**99.21**	90.87	90.14	96.47	**96.01**	91.91	92.93
GPT-4o-1.5	**96.12**	98.73	**99.27**	**95.11**	**95.93**	98.18	**97.14**	**98.25**	**99.58**	**96.05**	**96.23**	**98.55**	**94.14**	**95.40**	91.67	**96.77**	**94.52**	**99.42**	**96.36**	**98.72**	**96.63**	97.18	**96.03**	**94.93**	**97.14**	95.60	**94.36**	89.01
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	23.09	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
OneCAT	61.40	96.01	80.35	60.11	63.75	73.87	79.17	58.57	77.04	32.69	64.06	58.33	46.43	40.52	69.66	63.65	58.01	60.06	62.86	68.99	35.28	69.05	63.89	57.39	76.27	69.93	49.76	1.90
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
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	84.46	77.72	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
GLM-Image	79.11	89.62	93.35	79.26	73.89	85.62	88.83	76.92	87.74	78.21	71.25	73.19	66.07	62.07	74.21	73.09	73.72	85.34	80.58	76.38	71.50	86.90	73.41	61.62	87.15	84.42	58.09	82.88
FLUX.2-Klein-base-9b	81.41	97.67	92.63	82.45	86.94	95.86	93.18	89.69	98.27	81.09	83.75	80.07	79.46	79.60	85.52	90.94	87.18	95.98	89.56	90.78	90.19	92.46	86.11	**81.34**	92.94	91.67	**80.64**	5.98
FLUX.2-Klein-9b	81.74	**99.09**	92.92	77.66	86.67	96.54	92.05	88.99	**98.74**	77.88	85.00	85.87	81.25	78.74	85.52	90.56	83.65	**96.55**	91.50	**92.69**	89.49	92.86	**88.89**	78.87	**94.92**	91.30	80.15	5.71
LongCat-Image	83.14	90.20	93.35	78.72	82.08	91.79	89.39	86.19	96.80	76.28	85.00	86.96	75.00	74.14	83.53	81.25	72.76	90.52	85.92	81.89	79.91	90.87	75.00	68.66	88.84	83.15	65.69	82.07
Z-Image-Turbo	83.69	96.26	94.80	77.66	79.31	93.37	87.88	85.49	97.48	75.64	78.75	77.90	75.45	73.85	82.34	84.69	75.00	87.93	87.86	79.87	80.37	89.68	77.78	68.66	90.40	84.78	70.83	74.73
FLUX.2-dev	86.12	98.42	95.52	84.57	89.44	**97.59**	92.80	**93.01**	98.32	85.58	87.81	**91.67**	**87.95**	**88.51**	**88.79**	**92.73**	**88.46**	95.40	**92.23**	90.25	87.85	**97.62**	82.94	73.59	93.93	**95.11**	79.66	43.21
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
Z-Image	**89.17**	97.67	95.52	86.70	86.39	97.21	92.42	89.51	98.01	**86.22**	88.44	84.42	83.93	81.32	88.00	88.90	83.01	92.53	91.26	86.02	86.45	93.65	80.16	72.54	93.79	91.12	79.90	**88.59**