# MMIG-Bench: Towards Comprehensive and Explainable Evaluation of Multi-Modal Image Generation Models

**Hang Hua**[1]* **Ziyun Zeng**[1]* **Yizhi Song**[2]* **Yunlong Tang**[1]
**Liu He**[2] **Daniel Aliaga**[2] **Wei Xiong**[3] **Jiebo Luo**[1]†
[1] University of Rochester [2] Purdue University [3] NVIDIA
{hhua2, jluo}@cs.rochester.edu, {zzeng24, ytang37}@ur.rochester.edu,
{song630, he425, aliaga}@purdue.edu, wxiong@nvidia.com


Figure 1: Overview of MMIG-Bench. We present a unified multi-modal benchmark which contains 1,750 multi-view reference images with 4,850 richly annotated text prompts, covering both text-only and image-text-conditioned generation. We also propose a comprehensive three-level evaluation framework: low-level of artifacts and identity preservation, mid-level of VQA-based Aspect Matching Score, and high-level of aesthetics and human preferences—delivers holistic and interpretable scores.

## Abstract

Recent multimodal image generators such as GPT-4o, Gemini 2.0 Flash, and Gemini 2.5 Pro excel at following complex instructions, editing images and maintaining concept consistency. However, they are still evaluated by *disjoint* toolkits: text-to-image (T2I) benchmarks that lacks multi-modal conditioning, and customized image generation benchmarks that overlook compositional semantics

*Equal Contribution
†Corresponding Author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

and common knowledge. We propose MMIG-Bench, a *comprehensive* **M**ulti-**M**odal **I**mage **G**eneration **Bench**mark that unifies these tasks by pairing 4,850 richly annotated text prompts with 1,750 multi-view reference images across 380 subjects, spanning humans, animals, objects, and artistic styles. MMIG-Bench is equipped with a three-level evaluation framework: (1) low-level metrics for visual artifacts and identity preservation of objects; (2) novel Aspect Matching Score (AMS): a VQA-based mid-level metric that delivers fine-grained prompt-image alignment and shows strong correlation with human judgments; and (3) high-level metrics for aesthetics and human preference. Using MMIG-Bench, we benchmark 17 state-of-the-art models, including Gemini 2.5 Pro, FLUX, DreamBooth, and IP-Adapter, and validate our metrics with 32k human ratings, yielding in-depth insights into architecture and data design. Resources are available at: https://hanghuacs.github.io/MMIG-Bench/

# 1 Introduction

With the rapid progress in foundational image generation systems, a diverse range of models has emerged at the forefront of research and application. These include commercial models such as GPT-4o [45] and Gemini 2.0 Flash, as well as open-source models like FLUX [32], Hunyuan-DiT [37], Emu3 [69], and DreamO [44]. Currently, the community primarily evaluates them with separate toolkits: text-to-image (T2I) benchmarks that focus on compositionality and world knowledge; and customized generation benchmarks that emphasize identity preservation of the reference images. However, fine-grained semantic alignment and compositional reasoning included in the T2I evaluation are equally critical for the customization task; conversely, providing reference images with text enhances the flexibility and also broadens the expressive scope of generation—enabling style transfer and other capabilities that pure T2I tasks does not contain. Therefore there is a pressing need for a comprehensive and unified benchmark that treats multi-modal input (both text and image) as an integrated entity.

To be more specific, early T2I benchmarks (e.g., PartiPrompts, Gecko) are large sparsely labelled, typically assigning only a single category per prompt. Recent benchmarks (T2I-CompBench++ [24], GenEval [11], GenAI-Bench [34], T2I-FactualBench [26]) incorporate dense tags, evaluating nuanced aspects of generated images such as compositionality, common sense, and world knowledge. However, they focus on evaluating generators only conditioned on text, and thus are limited in evaluating newer multi-modal generation models with both images and text as input, such as GPT-4o and Gemini 2.0 Flash. Customization benchmarks [6, 48] are still scarce, most are tiny and lack enough multiview reference images. In addition, the evaluation metrics in T2I benchmarks mostly score prompt following, overlooking visual fidelity. Customization benchmarks often rely on trivial approaches to assess semantic alignment or identity preservation, lacking fine-grained and effective metrics.

To address these issues, we build the first comprehensive multi-modal benchmark MMIG-Bench for image generation. we summarize our contributions below and illustrate them in Fig. 1 and Fig. 2.

- **Unified task coverage and multi-modal input**. We collect over **380** groups (animal, object, human,, and style) comprising **1,750** multiview object-centric images enabling rigorous reference-based generation. We also construct **4,850** richly annotated prompts across compositionality (attributes, relations, objects, and numeracy), style (fixed pattern, professional, natural, human-written), realism (imaginative) and common sense (comparisons, negations). The proposed benchmark provides future research with the flexibility to conduct any image generation task.

- **Three-level evaluation suite.** We propose a multilevel scoring framework for comprehensive evaluation. (1) Low-level metrics assess visual artifacts and identity preservation of objects; (2) At mid-level, we propose the **Aspect Matching Score (AMS)** : a novel VQA-based metric that captures fine-grained semantic alignment, showing strong correlation with human perception; (3) high-level metrics measure aesthetics and human preferences. This multi-level framework expands T2I evaluation beyond prompt adherence and provides customized generation the nuanced semantic assessment it lacks.

We validate our metrics with **32k** human ratings and benchmark 17 state-of-the-art models, offering design insights on architecture choices and data curation. We will release MMIG-Bench and the evaluation code to accelerate future research on multi-modal generation.

# 2 Related Work

## 2.1 Text-to-Image Generation

Recent advancements in text-to-image generation have significantly enhanced models' visual synthesis capabilities [15, 16, 58–60, 78–80, 43, 42, 81]. FLUX.1-dev [32] employs a rectified flow transformer integrated with 3D modeling, enabling precise compositional control. Hunyuan-DiT [37] advances diffusion transformers with multilingual support and multimodal dialogue, enhancing caption accuracy. Lumina-Image 2.0 [52] prioritizes efficiency through unified architectures and progressive training, achieving scalability with compact models. Photon-v1 [50] specifically targets photorealism, effectively rendering challenging visual elements. PixArt-$\Sigma$ [3] innovates with attention mechanisms, achieving ultra-high-resolution generation. Stable Diffusion variants, including SDXL [51] and SD3.5 [7], leverage advanced multimodal conditioning to enhance image quality and textual fidelity. Janus Pro [4] offers superior multimodal stability through optimized training and extensive datasets. Finally, CogView4 [85], with its large-scale parameters, sets benchmarks in visual fidelity and resolution, highlighting ongoing innovation in generative image synthesis.

## 2.2 Customized Image Generation

Customized image generation techniques have significantly advanced, enabling precise, context-specific visual content [70]. DreamBooth [54] and HyperDreamBooth [56] established robust frameworks for efficient subject-driven fine-tuning from minimal references. Methods like Imagic [27] and Textual Inversion [10] embed new concepts into pretrained models for semantic editing without extensive retraining. InstantBooth [57] and GroundingBooth [74] streamline personalization, reducing computational costs and training time. With the recent breakthroughs in large language models (LMMs) [64, 13, 1, 82, 40, 38, 17], multimodal models such as Kosmos-G [47], UNIMO-G [36], and Emu3 [69] expand personalization capabilities through multimodal integration and semantic understanding. BLIP-Diffusion [35] and IP-Adapter [76] enhance visual grounding between textual prompts and personalized features. InstantID [67] specializes in identity-aware personalization with high-fidelity identity preservation. Recently, Personalize Anything [8] and DreamO [44] have further advanced the field, enabling versatile, contextually adaptive image synthesis.

## 2.3 Benchmarks and Metrics for Image Generation

Recent benchmarks and metrics comprehensively evaluate generative image models. DreamBench++ [48] and GenAI-Bench [34] systematically assess generative AI across diverse tasks, while PartiPrompts [77] and Gecko [71] provide specialized datasets for prompt-based generation fidelity. T2I-CompBench and T2I-CompBench++ [24] target compositional complexity and context understanding. DPG-Bench [18] focuses on perceptual metrics, whereas GenEval [11] and HEIM [33] offer robust frameworks for systematic comparison. Q-Eval-100K [84] and T2I-FactualBench [26] specifically evaluate factual consistency and quality alignment. Additionally, LMM4LMM [65] assesses multimodal language models for image generation, and EvalMuse-40K [14] provides extensive benchmarking of image quality and model performance.

# 3 Data Curation

## 3.1 Overview

Multi-modal image generation commonly involves both reference images and text prompts as inputs. Accordingly, our benchmark's data collection is structured into two components: grouped image collection and text prompt generation (as shown in Fig. 3). We begin by extracting entities from prompts used in existing text-to-image (T2I) benchmarks (such as [33, 34, 71]). After collecting over 2,000 distinct entities, we retain the 207 most frequent entities for subsequent use.

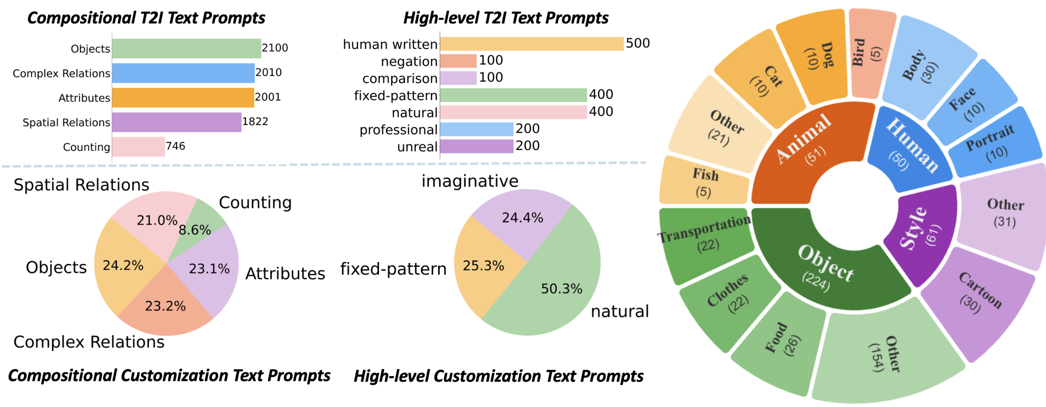


Figure 2: Statistics of the tags in MMIG-Bench. *Top-left*: Data distribution of compositional categories and high-level categories for text in T2I task. *Bottom-left*: Data distribution of text prompts in customization task. *Right*: Statistics of classes for the reference images.

### 3.1.1 Prompting GPT for Text Prompt Generation

To enable scalable and diverse prompt generation, we use GPT-4o with several predefined instruction templates, as illustrated in Fig. 3. By providing entities and instruction templates as inputs, we generate a total of 4,350 synthetic prompts covering both tasks. Furthermore, we manually select 500 human-written prompts from prior work [9, 34]. To ensure broad coverage of semantic aspects, we organize prompts into two main categories: compositional and high-level. The compositional category includes five sub-categories: *object*, *counting*, *attribute*, *spatial relations* (e.g., next to, atop, behind), and *complex relations* (e.g., pour into, toss, chase). The high-level category contains seven sub-categories, including *style* (fixed pattern, natural, professional, human-written), *realism* (imaginative), and *common sense* (negation, comparison).

To better control the aspects, style, and structure of the prompts, we design eight instruction templates, using the T2I task as an example. When prompts require compositionality and adherence to a specific structure, we use the following format: "`[scene description (optional)] + [number][attribute][entity1] + [interaction (spatial or action)] + [number (optional)][attribute][entity2]`". For prompts to resemble natural, human-written language, a more flexible instruction is used: "`Please generate prompts in a NATURAL format. It should contain one or more "entities / nouns", (optional) "attributes / adjective" that describes the entities, (optional) "spatial or action interactions" between entities, and (optional) "background description`". The full set of templates is provided in the Appendix.

To ensure the quality and safety of the generated prompts, we further filter out toxic or low-quality content (see Sec. 3.4), and utilize FineMatch [22] to generate dense labels (see Sec. 3.3.1), making the dataset more flexible and suitable for research applications.

## 3.2 Collecting Grouped Subject Images

Grouped reference images which are object-centric and realistic are usually missing from the previous benchmarks. However, multiple reference images have proven effective across various tasks, including image customization [30, 54, 86], video generation [28] and 3D reconstruction [66]. To address this gap, we collect a large set of grouped reference images.

The target objects are selected from the 207 common entities we previously identified. We employ annotators to curate grouped object images from Pexels [49] following these guidelines: (1) each group contains 3–5 images of the same object; (2) within each group, the object appears in varying poses or views; and (3) objects with complex logos or textures are prioritized. Additionally, we collect artistic images in 12 styles (e.g., sketch, low-poly, oil painting) to support style transfer tasks.

In total, we collect 1,750 images across 386 groups, covering four main categories—animals, humans, objects, and styles—as shown in Fig. 2 (right). To ensure quality, we apply filtering and cropping to remove unrelated content from the images. Based on the entities in the collected images, we generate corresponding text prompts using the aforementioned procedure.

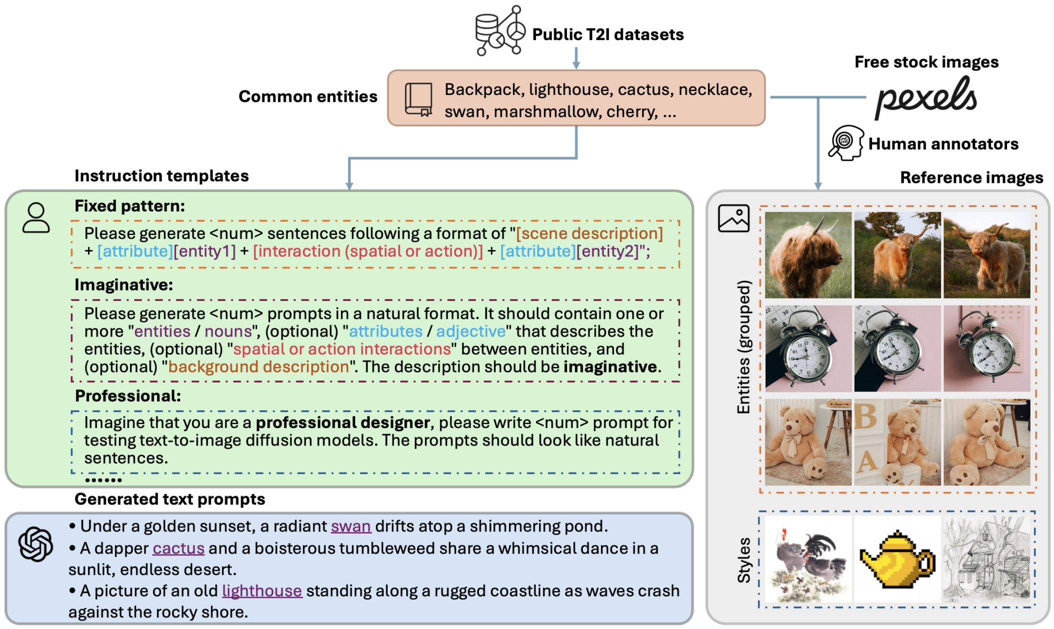


Figure 3: Our data curation pipeline for multi-modal image generation benchmarking. We begin by extracting 207 frequent entities from public T2I datasets. Using these entities, we generate diverse prompts with GPT-4o by prompting it with a set of carefully designed instruction templates, which control the structure and style of the prompts (left). Simultaneously, we collect grouped reference images for each entity from free stock sources, with human annotators selecting 3–5 object-centric images per group that vary in pose or view (right). We further collect artistic images in 12 visual styles to support style transfer. The resulting dataset includes high-quality, structured text-image pairs for both T2I and customization.

### 3.3 Data Curation for Mid-Level Evaluation

The goal of mid-level evaluation is to analyze the text-image alignment in fine-grained aspects, enabling more interpretable assessment on the generated details. To this end, we follow FineMatch [22] to analyze the fine-grained text-image alignment from the perspective of **Object**, **Relation**, **Attribute**, and **Counting**. We conduct specific data curation for these aspects by first using GPT-4o to extract all the aspect-related phrases from input prompts and then using in-context learning to prompt GPT-4o to generate the corresponding QA pairs.

#### 3.3.1 Prompt Parsing

We follow FineMatch [22] to curate aspect phrases from text prompts, employing GPT-4o for aspect graph parsing due to its superior compositional parsing capabilities. Specifically, GPT-4o is guided by explicit instructions and in-context examples to accurately extract and categorize phrases into four categories: objects, relations, attributes, and counting queries.

#### 3.3.2 QA Pair Generation

Following the prior VQA-based evaluation frameworks [5, 19–21, 23, 39, 62, 75], we proceed to generate high-quality question-answer (QA) pairs corresponding to each aspect phrase. Initially, domain experts manually curate a set of exemplar QA pairs for each category (Object, Relation, Attribute, Counting). These manually curated QA pairs serve as contextual examples in the subsequent in-context learning phase. GPT-4o is then prompted with these examples to generate a comprehensive set of QA pairs for the extracted aspect phrases, ensuring alignment with the fine-grained evaluation dimensions. This automated generation process is iteratively refined by adjusting instructions and examples based on preliminary outputs to improve coverage, clarity, and consistency.

### 3.4 Human Verification

To guarantee dataset quality, interpretability, and reliability, we engage trained human annotators in a structured verification process. Annotators perform multiple quality assurance tasks, including: ❶ **Toxicity and Appropriateness Filtering**: Annotators screen generated QA pairs for toxic, offensive, or inappropriate content, ensuring ethical compliance and usability in research settings.❷ **QA Pair Correction and Validation**: Each QA pair generated by GPT-4o undergoes meticulous human validation to confirm the logical coherence, accuracy, and relevance to the original aspect phrase. Annotators refine ambiguous questions, corrected factual inaccuracies, and ensure precise correspondence between questions and answers. ❸ **Aspect Phrase Refinement**: Extract aspect phrases were scrutinized and refined for linguistic clarity and semantic precision. Annotators review each phrase to ensure they correctly and comprehensively represent the intended compositional aspects (Object, Relation, Attribute, Counting).

After these rigorous human verification steps, we obtain a high-quality dataset consisting of 28,668 (16,819 for T2I tasks and 11,849 for Customization tasks) validated QA pairs, explicitly designed to support detailed analyses of fine-grained text-image alignment.

## 4 Proposed Metrics - MMIG-Bench

### 4.1 Low-Level Evaluation Metrics

The goal of low-level evaluation is to assess artifacts in the generated images and to evaluate the low-level feature similarity between the generated images and the prompt, as well as between the generated images and the reference images. To achieve this, we leverage previous evaluation metrics:

- CLIP-Text [53]: measures the semantic alignment between the generated image and input prompt;
- CLIP-Image, DINOv2 [46], and CUTE [29]: measures identity preservation;
- PAL4VST [83]: measures the amount of generative artifacts using a segmentation model.

These metrics collectively provide a comprehensive assessment of the visual quality and consistency.

### 4.2 Mid-Level Evaluation Metrics

The goal of mid-level evaluation is to assess the fine-grained semantic alignment of generated images with text prompts. We use the collected QA pairs corresponding to the four aspects (as described in Section 3.3) to design a new interpretable evaluation framework, **Aspect Matching Score (AMS)**.

#### 4.2.1 Aspect Matching Score

Formally, given a prompt $P$, we extract a set of $n$ aspect phrases $\{A_1, A_2, \ldots, A_n\}$ and generate a corresponding set of VQA pairs $\{(Q_1, Ans_1), (Q_2, Ans_2), \ldots, (Q_n, Ans_n)\}$. These questions are designed to probe whether the generated image $I$ faithfully reflects the semantics of each aspect.

To compute the alignment score, we use Qwen2.5-VL-72B [2] to answer each question $Q_i$ based on the generated image $I$, resulting in predicted answers $\{\hat{Ans}_1, \hat{Ans}_2, \ldots, \hat{Ans}_n\}$. We then compare each prediction $\hat{Ans}_i$ with the ground truth answer $Ans_i$ to assess correctness. We define the **Aspect Matching Score** as the proportion of correctly answered VQA questions:

$$\text{AMS}(I, P) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\hat{Ans}_i = Ans_i), \tag{1}$$

where $\mathbf{1}(\cdot)$ is an indicator function that returns 1 if the predicted answer exactly matches the ground truth and 0 otherwise.

**AMS** provides a direct and interpretable measure of how well the generated image aligns with each semantic component of the prompt. A higher **AMS** indicates better fine-grained alignment, capturing failures that coarse-level metrics often miss.

### 4.3 High-Level Evaluation Metrics

The goal of high-level evaluation is to evaluate image aesthetics and human preference in the generated images. To achieve this, we leverage previous evaluation metrics, such as Aesthetic, HPSv2 and

PickScore. These metrics offer a comprehensive assessment of the visual appeal and alignment with human preferences in the generated outputs.

# 5 Experiments

Table 1: Quantitative comparison is conducted across images generated by 12 different text-to-image models using 2,100 well-designed prompts. Most models generate images at the default resolution of 1024 × 1024, except for the two autoregressive models, which produce outputs at 384 × 384, and GPT-4o and Gemini-2.0-Flash produce images with variable, non-fixed resolutions. ↑ indicates higher is better and ↓ indicates lower is better. The **best** and second-best results are in bold and underlined, respectively.

| | Low Level | | Mid Level | | High Level | | |
|---|---|---|---|---|---|---|---|
| **Method** | **CLIP-T ↑** | **PAL4VST ↓** | **AMS ↑** | **Human ↑** | **Aesthetic ↑** | **HPSv2 ↑** | **PickScore ↑** |
| **Diffusion Models** | | | | | | | |
| SDXL [51] | 33.529 | 14.340 | 79.08 | 72.29 | 6.337 | 0.277 | 0.120 |
| Photon-v1 [50] | 33.296 | 2.947 | 77.12 | 69.49 | 6.391 | 0.284 | 0.088 |
| Lumina-2 [52] | 33.281 | 15.531 | 84.11 | 73.18 | 6.048 | 0.287 | 0.116 |
| HunyuanDit-v1.2 [37] | 33.701 | 8.024 | 83.61 | 74.89 | 6.379 | 0.300 | 0.144 |
| Pixart-Sigma-xl2 [3] | 33.682 | 9.283 | 83.18 | 76.65 | 6.409 | 0.304 | 0.165 |
| Flux.1-dev [31] | 33.017 | 2.171 | 84.44 | 76.44 | 6.433 | 0.307 | 0.210 |
| SD 3.5-large [7] | 33.873 | 6.359 | 85.33 | 77.04 | 6.318 | 0.294 | 0.157 |
| HiDream-I1-Full [63] | **33.876** | **1.522** | **89.65** | **83.18** | 6.457 | **0.321** | **0.450** |
| **Autoregressive Models** | | | | | | | |
| JanusFlow [41] | 31.498 | 365.663 | 70.25 | 75.69 | 5.221 | 0.209 | 0.031 |
| Janus-Pro-7B [4] | 33.358 | 31.954 | 85.35 | 80.36 | 6.038 | 0.275 | 0.129 |
| **API-based Models** | | | | | | | |
| Gemini-2.0-Flash [12] | 32.433 | 11.053 | 85.35 | 81.98 | 6.102 | 0.275 | 0.110 |
| GPT-4o [45] | 32.380 | 3.497 | 82.57 | 81.02 | **6.719** | 0.279 | 0.263 |

Table 2: Quantitative comparison is conducted across imagees generated by 9 different multi-modal image generation models using 1,690 samples. Most models generate images 3 times per multi-modal input except GPT-4o at the default resolution of 1024 × 1024, except for Blip Diffusion, which produce outputs at 512 × 512, and GPT-4o produce images with variable, non-fixed resolutions. ↑ indicates higher is better and ↓ indicates lower is better. The **best** and second best results are in bold and underlined, respectively.

| | Low Level | | | | | Mid Level | | High Level | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | **CLIP-T ↑** | **CLIP-I ↑** | **DINOv2 ↑** | **CUTE ↑** | **PAL4VST ↓** | **BLIPVQA ↑** | **AMS ↑** | **Aesthetic ↑** | **HPSv2 ↑** | **PickScore ↑** |
| **Diffusion Models** | | | | | | | | | | |
| BLIP Diffusion[35] | 26.137 | 80.286 | 26.232 | 69.681 | 56.780 | 0.247 | 41.59 | 5.830 | 0.213 | 0.032 |
| DreamBooth [55] | 24.227 | **88.758** | **38.961** | **79.780** | 43.535 | 0.108 | 28.00 | 5.368 | 0.179 | 0.019 |
| Emu2 [61] | 28.410 | 79.026 | 31.831 | 71.132 | 10.461 | 0.378 | 53.13 | 5.639 | 0.243 | 0.066 |
| Ip-Adapter-XL [76] | 28.577 | 85.297 | 34.177 | 74.995 | 8.531 | 0.290 | 51.10 | 5.840 | 0.233 | 0.073 |
| MS Diffusion [68] | 31.446 | 77.827 | 23.600 | 71.306 | 4.748 | 0.496 | 71.40 | 5.979 | 0.271 | 0.143 |
| UNO [72] | 31.439 | 75.194 | 23.079 | 65.808 | 3.060 | 0.539 | 67.94 | 6.156 | 0.271 | 0.087 |
| RealCustom [25] | 31.596 | 73.236 | 22.678 | 67.132 | 2.517 | 0.533 | 63.77 | 6.133 | 0.291 | 0.094 |
| OmniGen [73] | 33.178 | 72.327 | 21.380 | 56.666 | 2.908 | 0.588 | 73.52 | 6.086 | **0.296** | 0.152 |
| **API-based Models** | | | | | | | | | | |
| GPT-4o [45] | **33.527** | 75.152 | 25.174 | 64.776 | **1.973** | **0.672** | **90.90** | **6.368** | 0.289 | **0.550** |

## 5.1 Human Evaluation

To evaluate the semantic preservation of state-of-the-art generation models and compare the human correlation of VQA-based metrics, we conduct five user studies. We assess 12 text-to-image (T2I) models across five aspects: attribute, relation, counting, object, and general prompt following. For each of the first four aspects, 150 prompts are randomly selected; for the last, 300 prompts are used. In each study, users are shown a prompt and a generated image, and asked to rate semantic alignment on a 1–5 scale based on the target aspect (see Appendix for details). In total, we collect 32.4k ratings from over 8,000 Amazon Mechanical Turk users. Results are reported in Table 3.


Figure 4: A qualitative study of text-only (top) and text-image-conditioned (bottom) generation methods on MMIG-Bench.

Table 3: Comparison of VQA-based metrics: BLIPVQA [26], VQ2 [75], DSG [5], and our AMS.

| Method | BLIPVQA ↑ | VQ2 ↑ | DSG ↑ | AMS ↑ | Human ↑ |
|---|---|---|---|---|---|
| **Diffusion Models** | | | | | |
| SDXL [51] | 0.433 | 69.07 | 87.63 | 79.08 | 72.29 |
| Photon-v1 [50] | 0.440 | 66.84 | 86.26 | 77.12 | 69.49 |
| Lumina-2 [52] | 0.517 | 72.51 | 90.12 | 84.11 | 73.18 |
| HunyuanDiT-v1.2 [37] | 0.513 | 73.13 | 89.77 | 83.61 | 74.89 |
| Pixart-Sigma-xl2 [3] | 0.521 | 71.51 | 89.69 | 83.18 | 76.65 |
| Flux.1-dev [32] | 0.511 | 71.41 | 83.33 | 84.44 | 76.44 |
| SD 3.5-large [7] | 0.525 | 73.28 | 91.41 | 85.33 | 77.04 |
| HiDream-I1-Full [63] | 0.572 | 75.09 | 92.43 | 89.65 | 83.18 |
| **Autoregressive Models** | | | | | |
| JanusFlow [41] | 0.390 | 57.24 | 85.43 | 70.25 | 75.69 |
| Janus-Pro [4] | 0.530 | 67.41 | 92.15 | 85.35 | 80.36 |
| **API-based Models** | | | | | |
| Gemini-2.0-Flash [12] | 0.495 | 72.01 | 92.93 | 85.40 | 81.98 |
| GPT-4o [45] | 0.497 | 70.34 | 89.99 | 82.57 | 81.02 |

### 5.2 Correlation of Automated Metrics with Human Annotations

To assess the alignment of automated metrics with human, we compute Spearman correlations against human annotations. As shown in Table 3, our proposed AMS achieves the highest correlation ($\rho = \mathbf{0.699}$), surpassing DSG ($\rho = 0.692$), VQ2 ($\rho = 0.399$), and BLIPVQA ($\rho = 0.147$). This demonstrates the effectiveness of AMS as a reliable metric for compositional T2I evaluation.

### 5.3 Leaderboard

We compare the performance across state-of-the-art models in T2I task (Tab. 1) and customization task (Tab. 2) using our multi-level evaluation framework. Based on the scores, we can derive the following insights:

In T2I task: (1) Compared with diffusion models, autoregressive models (JanusFlow and Janus-Pro-7B) perform significantly worse in visual quality, as they are more likely to generate artifacts, and have the lowest aesthetic and human preference scores. (2) HiDream-I1, the largest model with 17B parameters, excels all the other generators; it takes advantage of rectified flow and the VAE from FLUX.1-schnell. (3) FLUX.1-dev (the second largest model with 12B parameters) stands at the second place for most metrics. (4) The performance of HiDream-I1 and FLUX.1-dev suggests the importance of scaling generative models. (5) Although GPT-4o is not the best model in all metrics, it shows very robust generation abilities competitive to the best model in each category.

In customization task, we draw the following conclusions: (1) In most low-level metrics that evaluates identity preservation, DreamBooth is the strongest model; its multi-view inputs and test-time finetuning greatly enhances the identity learning. (2) GPT-4o cannot preserve the identity well, this ability is even worse than some early models like Emu2 and the two encoder-based models (BLIP Diffusion and IP-Adapter). (3) GPT-4o comes at the first place in visual quality and semantic alignment. (4) RealCustom and Omnigen are often the second best in terms of generation quality. However, they show an unsatisfactory ability on identity preservation.

### 5.4 Qualitative Analysis

We present qualitative results for multi-modal image generation in Fig. 4. The top six rows illustrate generations conditioned on text only; the bottom three rows show generations conditioned on both image and text. Key observations are as follows:

In the T2I task, (1) Hunyuan-DiT-V1.2 struggles with entity generation, frequently missing objects, duplicating them, or generating incorrect ones; (2) Pixart-Sigma-XL2 exhibits stronger visual artifacts (e.g., around benches, chairs, and computers), consistent with its lower PAL4VST scores from Tab. 1. In customization task, (1) Non-rigid objects (e.g., dogs) tend to appear in more diverse poses; (2) MS-Diffusion performs worst in preserving object identity, while DreamBooth performs best; This highly aligns with the CLIP-I and DINOv2 scores in Tab. 2. (3) Despite its strength in identity preservation, DreamBooth often fails to generate the correct scene, actions, or additional entities, indicating poor compositional alignment.

# 6 Discussions and Conclusions

We present MMIG-Bench, the first benchmark to treat multi-modal image generation as a single task rather than two disjoint tasks. We demonstrate that by pairing 1,750 multi-view reference images with 4,850 densely annotated prompts, MMIG-Bench enables side-by-side evaluation of pure text-to-image, image-conditioned customization, and every hybrid in between. The proposed three-level evaluation framework provides a comprehensive, interpretable assessment that addresses the evaluation gaps in both T2I and customization tasks. The evaluation metrics prove to be well aligned with human preferences by comparing them with 32k human ratings across 17 state-of-the-art models. The in-depth assessments of the image generators on our benchmark provide insights on how the model capacity, model architecture, and other factors influence the image quality. One limitation is that the human ratings do not yet cover visual quality; we plan to expand future studies to such dimensions. We will publicly release the data, code, and leaderboard to encourage transparent comparison and guide future advances in architecture design, data curation, and training strategy.

# References


[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[2] 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.

[3] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-$\sigma$: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, 2024. URL `https://api.semanticscholar.org/CorpusID:268264262`.

[4] 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*, abs/2501.17811, 2025. URL `https://api.semanticscholar.org/CorpusID:275954151`.

[5] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *ArXiv*, abs/2310.18235, 2023. URL `https://api.semanticscholar.org/CorpusID:264555374`.

[6] dreambench. Dreambench, 2022. `https://github.com/nousr/dream-bench`.

[7] Patrick Esser, Sumith Kulal, A. Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *ArXiv*, abs/2403.03206, 2024. URL `https://api.semanticscholar.org/CorpusID:268247980`.

[8] Haoran Feng, Zehuan Huang, Lin Li, Hairong Lv, and Lu Sheng. Personalize anything for free with diffusion transformer. *arXiv preprint arXiv:2503.12590*, 2025.

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

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

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

[12] Google. Gemini 2.0 flash, 2025. `https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#gemini-2-0-flash`.

[13] 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.

[14] Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, and Chongyi Li. Evalmuse-40k: A reliable and fine-grained benchmark with comprehensive human annotations for text-to-image generation model evaluation. *ArXiv*, abs/2412.18150, 2024. URL `https://api.semanticscholar.org/CorpusID:274992412`.

[15] Liu He, Yijuan Lu, John Corring, Dinei Florencio, and Cha Zhang. Diffusion-based document layout generation. In *International Conference on Document Analysis and Recognition*, pages 361–378. Springer, 2023.

[16] Liu He, Yizhi Song, Hejun Huang, Pinxin Liu, Yunlong Tang, Daniel Aliaga, and Xin Zhou. Kubrick: Multimodal agent collaborations for synthetic video generation. *arXiv preprint arXiv:2408.10453*, 2024.

[17] Liu He, Xiao Zeng, Yizhi Song, Albert YC Chen, Lu Xia, Shashwat Verma, Sankalp Dayal, Min Sun, Cheng-Hao Kuo, and Daniel Aliaga. Advancing multimodal llms by large-scale 3d visual instruction dataset generation. *arXiv preprint arXiv:2507.08513*, 2025.

[18] 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.

[19] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022.

[20] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20349–20360, 2023. URL `https://api.semanticscholar.org/CorpusID:257636562`.

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

[22] Hang Hua, Jing Shi, Kushal Kafle, Simon Jenni, Daoan Zhang, John Collomosse, Scott Cohen, and Jiebo Luo. Finematch: Aspect-based fine-grained image and text mismatch detection and correction. In *European Conference on Computer Vision*, pages 474–491. Springer, 2024.

[23] Hang Hua, Yunlong Tang, Ziyun Zeng, Liangliang Cao, Zhengyuan Yang, Hangfeng He, Chenliang Xu, and Jiebo Luo. Mmcomposition: Revisiting the compositionality of pre-trained vision-language models. *arXiv preprint arXiv:2410.09733*, 2024.

[24] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:3563–3579, 2023. URL `https://api.semanticscholar.org/CorpusID:259847295`.

[25] Mengqi Huang, Zhendong Mao, Mingcong Liu, Qian He, and Yongdong Zhang. Realcustom: Narrowing real text word for real-time open-domain text-to-image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7476–7485, 2024.

[26] Ziwei Huang, Wanggui He, Quanyu Long, Yandi Wang, Haoyuan Li, Zhelun Yu, Fangxun Shu, Long Chan, Hao Jiang, Leilei Gan, et al. T2i-factualbench: Benchmarking the factuality of text-to-image models with knowledge-intensive concepts. *arXiv preprint arXiv:2412.04300*, 2024.

[27] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6007–6017, 2023.

[28] Xianghao Kong, Qiaosong Qi, Yuanbin Wang, Anyi Rao, Biaolong Chen, Aixi Zhang, Si Liu, and Hao Jiang. Profashion: Prototype-guided fashion video generation with multiple reference images. *arXiv preprint arXiv:2505.06537*, 2025.

[29] Klemen Kotar, Stephen Tian, Hong-Xing Yu, Daniel L. K. Yamins, and Jiajun Wu. Are these the same apple? comparing images based on object intrinsics. *ArXiv*, abs/2311.00750, 2023. URL `https://api.semanticscholar.org/CorpusID:264935263`.

[30] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023.

[31] Black Forest Labs. Flux.1, 2024. `https://bfl.ai/announcements/24-08-01-bfl`.

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

[33] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36:69981–70011, 2023.

[34] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation. *ArXiv*, abs/2406.13743, 2024. URL `https://api.semanticscholar.org/CorpusID:270619531`.

[35] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023.

[36] Wei Li, Xue Xu, Jiachen Liu, and Xinyan Xiao. Unimo-g: Unified image generation through multimodal conditional diffusion. *arXiv preprint arXiv:2401.13388*, 2024.

[37] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiao-Ting Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Mengxi Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yang-Dan Tao, Jianchen Zhu, Kai Liu, Si-Da Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Dingyong Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *ArXiv*, abs/2405.08748, 2024. URL `https://api.semanticscholar.org/CorpusID:269761491`.

[38] Yuan-Hong Liao, Sven Elflein, Liu He, Laura Leal-Taixé, Yejin Choi, Sanja Fidler, and David Acuna. Longperceptualthoughts: Distilling system-2 reasoning for system-1 perception. *arXiv preprint arXiv:2504.15362*, 2025.

[39] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024.

[40] Jiayi Liu, Tinghan Yang, and Jennifer Neville. Cliqueparcel: An approach for batching llm prompts that jointly optimizes efficiency and faithfulness. *arXiv preprint arXiv:2402.14833*, 2024.

[41] 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. *ArXiv*, abs/2411.07975, 2024. URL https://api.semanticscholar.org/CorpusID:273969525.

[42] Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, and Zicheng Liu. Training diffusion models towards diverse image generation with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10844–10853, 2024.

[43] Zichen Miao, Zhengyuan Yang, Kevin Lin, Ze Wang, Zicheng Liu, Lijuan Wang, and Qiang Qiu. Tuning timestep-distilled diffusion model using pairwise sample optimization. *arXiv preprint arXiv:2410.03190*, 2024.

[44] Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. *arXiv preprint arXiv:2504.16915*, 2025.

[45] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

[46] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. URL https://api.semanticscholar.org/CorpusID:258170077.

[47] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. *ArXiv*, abs/2310.02992, 2023. URL https://api.semanticscholar.org/CorpusID:263620748.

[48] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *ArXiv*, abs/2406.16855, 2024. URL https://api.semanticscholar.org/CorpusID:270702690.

[49] Pexels. Pexels, 2014. https://www.pexels.com/.

[50] Photon78. Photon-v1. https://civitai.com/models/84728/photon78, 2023. Accessed: 2025-05-06.

[51] Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952, 2023. URL https://api.semanticscholar.org/CorpusID:259341735.

[52] Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, Xiangyang Zhu, Manyuan Zhang, Will Beddow, Erwann Millon, Victor Perez, Wen-Hao Wang, Conghui He, Bo Zhang, Xiaohong Liu, Hongsheng Li, Yu-Hao Qiao, Chang Xu, and Peng Gao. Lumina-image 2.0: A unified and efficient image generative framework. *ArXiv*, abs/2503.21758, 2025. URL https://api.semanticscholar.org/CorpusID:277349538.

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

[54] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022.

[55] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

[56] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models, 2023.

[57] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8543–8552, 2024.

[58] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18310–18319, 2023.

[59] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, He Zhang, Wei Xiong, and Daniel Aliaga. Imprint: Generative object compositing by learning identity-preserving representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8048–8058, 2024.

[60] Yizhi Song, Liu He, Zhifei Zhang, Soo Ye Kim, He Zhang, Wei Xiong, Zhe Lin, Brian L Price, Scott Cohen, Jianming Zhang, et al. Refine-by-align: Reference-guided artifacts refinement through semantic alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

[61] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024.

[62] Yunlong Tang, Junjia Guo, Hang Hua, Susan Liang, Mingqian Feng, Xinyang Li, Rui Mao, Chao Huang, Jing Bi, Zeliang Zhang, et al. Vidcomposition: Can mllms analyze compositions in compiled videos? *arXiv preprint arXiv:2411.10979*, 2024.

[63] HiDream-AI Team. Hidream-i1: A 17b parameter open chinese text-to-image generation model. `https://github.com/HiDream-ai/HiDream-I1`, 2024. Accessed: 2025-05-14.

[64] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[65] Jiarui Wang, Huiyu Duan, Yu Zhao, Juntong Wang, Guangtao Zhai, and Xiongkuo Min. Lmm4lmm: Benchmarking and evaluating large-multimodal image generation with lmms. 2025. URL `https://api.semanticscholar.org/CorpusID:277741112`.

[66] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023.

[67] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

[68] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024.

[69] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Lian zi Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. *ArXiv*, abs/2409.18869, 2024. URL `https://api.semanticscholar.org/CorpusID:272968818`.

[70] Yuxiang Wei, Yiheng Zheng, Yabo Zhang, Ming Liu, Zhilong Ji, Lei Zhang, and Wangmeng Zuo. Personalized image generation with deep generative models: A decade survey. *arXiv preprint arXiv:2502.13081*, 2025.

[71] Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Chris Knutsen, Cyrus Rashtchian, Jordi Pont-Tuset, and Aida Nematzadeh. Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings. *arXiv preprint arXiv:2404.16820*, 2024.

[72] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025.

[73] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025.

[74] Zhexiao Xiong, Wei Xiong, Jing Shi, He Zhang, Yizhi Song, and Nathan Jacobs. Groundingbooth: Grounding text-to-image customization. *arXiv preprint arXiv:2409.08520*, 2024.

[75] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran. O. Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. *ArXiv*, abs/2305.10400, 2023. URL `https://api.semanticscholar.org/CorpusID:258740893`.

[76] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[77] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Trans. Mach. Learn. Res.*, 2022, 2022. URL `https://api.semanticscholar.org/CorpusID:249926846`.

[78] Yongsheng Yu, Ziyun Zeng, Hang Hua, Jianlong Fu, and Jiebo Luo. Promptfix: You prompt and we fix the photo. *arXiv preprint arXiv:2405.16785*, 2024.

[79] Yongsheng Yu, Ziyun Zeng, Haitian Zheng, and Jiebo Luo. Omnipaint: Mastering object-oriented editing via disentangled insertion-removal inpainting. *arXiv preprint arXiv:2503.08677*, 2025.

[80] Yongsheng Yu, Haitian Zheng, Zhifei Zhang, Jianming Zhang, Yuqian Zhou, Connelly Barnes, Yuchen Liu, Wei Xiong, Zhe Lin, and Jiebo Luo. Zipir: Latent pyramid diffusion transformer for high-resolution image restoration. *arXiv preprint arXiv:2504.08591*, 2025.

[81] Yu Yuan, Xijun Wang, Yichen Sheng, Prateek Chennuri, Xingguang Zhang, and Stanley Chan. Generative photography: Scene-consistent camera control for realistic text-to-image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7920–7930, 2025.

[82] Jiaxing Zhang, Jiayi Liu, Dongsheng Luo, Jennifer Neville, and Hua Wei. Llmexplainer: Large language model based bayesian inference for graph explanation generation. *arXiv preprint arXiv:2407.15351*, 2024.

[83] Lingzhi Zhang, Zhengjie Xu, Connelly Barnes, Yuqian Zhou, Qing Liu, He Zhang, Sohrab Amirghodsi, Zhe Lin, Eli Shechtman, and Jianbo Shi. Perceptual artifacts localization for image synthesis tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7579–7590, October 2023.

[84] Zicheng Zhang, Tengchuan Kou, Shushi Wang, Chunyi Li, Wei Sun, Wei Wang, Xiaoyu Li, Zongyu Wang, Xuezhi Cao, Xiongkuo Min, Xiaohong Liu, and Guangtao Zhai. Q-eval-100k: Evaluating visual quality and alignment level for text-to-vision content. *ArXiv*, abs/2503.02357, 2025. URL `https://api.semanticscholar.org/CorpusID:276775486`.

[85] Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihan Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and Jie Tang. Cogview3: Finer and faster text-to-image generation via relay diffusion. *ArXiv*, abs/2403.05121, 2024. URL `https://api.semanticscholar.org/CorpusID:268297194`.

[86] Zhuofan Zong, Dongzhi Jiang, Bingqi Ma, Guanglu Song, Hao Shao, Dazhong Shen, Yu Liu, and Hongsheng Li. Easyref: Omni-generalized group image reference for diffusion models via multimodal llm. *arXiv preprint arXiv:2412.09618*, 2024.

## NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The empirical results and theory analysis support our claims in the abstract and introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We provide limitation discussion in conclsion.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [No]

Justification: No theory assumptions and proofs.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide the code and data.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We provide the code and data.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: This paper specifies all the training and test details.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We provide the statistical significance of the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We provide the information in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We provide the information in Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We claim that the dataset is for research purpose only.

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: We provide a README file and the dataset structure is carefully organized.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: [NA]

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: [NA]

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [NA]

Justification: [NA]

Guidelines:

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

# A Appendix

## A.1 Qualitative Results of MMIG-Bench Data


Figure 5: Word clouds of text prompts for the text-only generation (T2I) task (left) and the multimodal generation task (right).

Figure 5 visually summarizes the prominent semantic elements in the benchmark prompts for text-only (T2I) and multimodal generation tasks. The differentiation of the word clouds reflects task-specific features of MMIG-Bench, emphasizing spatial and descriptive details in T2I tasks, while multimodal tasks more frequently involve social and interactive scenarios.

## A.2 Quantitative and Qualitative Results of AMS

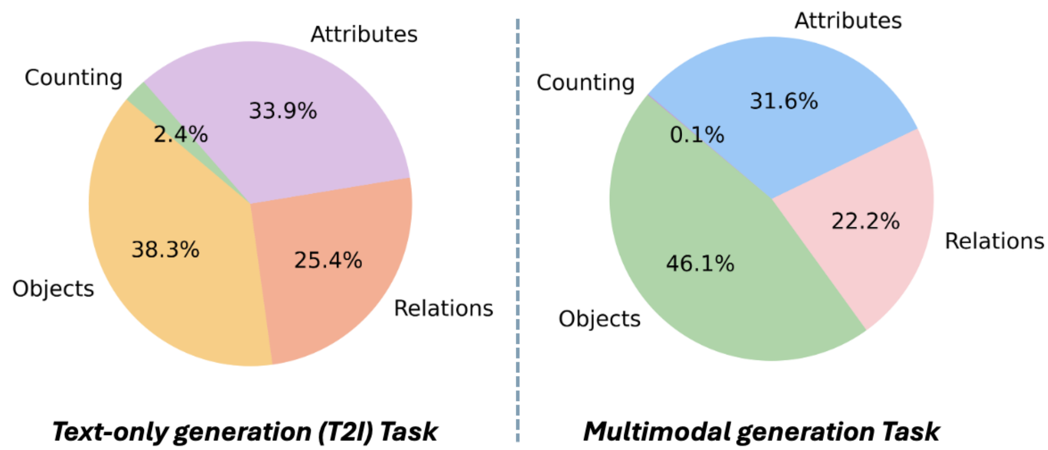


Figure 6: Aspect Distribution of the QA pairs of **AMS**.

Table 4: Aspect-level correlation ($\rho$) between **AMS** and human scores across four aspects.

| Aspect | Objects ↑ | Relations ↑ | Attributes ↑ | Counting ↑ | Overall ↑ |
|---|---|---|---|---|---|
| **Spearman** $\rho$ | 0.469 | 0.909 | 0.601 | 0.839 | 0.699 |

As depicted in Figure 6, the distribution of aspect types differs notably between the text-only generation (T2I) and multi-modal generation tasks. In the T2I setting, "Objects" dominate with 38.3%, while "Attributes" and "Relations" also constitute substantial proportions (33.9% and 25.4%, respectively). In multi-modal generation, "Objects" and "Attributes" remain prominent (46.1% and 31.6%, respectively), but the relative proportion of "Relations" decreases significantly (22.2%). The presence of "Counting" (0.1%) questions suggests this aspect is less frequent in the customized T2I generation task.

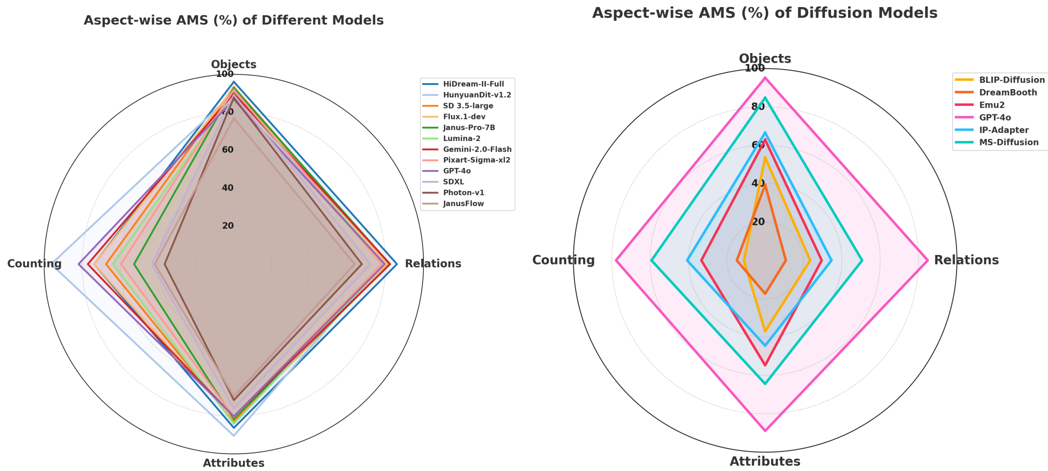


Figure 7: The AMS of different models on the text-only generation (T2I) task (left) and the multimodal generation task (right).

Figure 7 presents a comparative analysis of aspect-wise **AMS** across different models on the text-only generation (T2I) task and the multimodal generation task, highlighting their performance on four key compositional dimensions: Objects, Relations, Attributes, and Counting. On the T2I task, large-scale foundation models such as HiDream-I1, HunyuanDit-v1.2, and SD 3.5-large consistently achieve high AMS scores across aspects, particularly excelling in Objects and Attributes. Specifically, HunyuanDit-v1.2 demonstrates superior Counting performance, underscoring strong numerical understanding in text-driven scenarios. In contrast, for the multimodal generation task, GPT-4o significantly outperforms other diffusion-based models, particularly in complex compositional aspects such as Relations and Counting, highlighting its robust capability in interpreting and synthesizing multimodal inputs. Models like DreamBooth and BLIP-Diffusion show markedly weaker performances, especially in Relations and Counting. These AMS-based comparisons effectively illustrate clear distinctions in compositional understanding capabilities between text-only and multimodal generation settings, emphasizing the metric's sensitivity in capturing fine-grained model differences.

Table 4 further provides quantitative evidence of AMS's effectiveness: AMS achieves high Spearman correlation with human judgment, particularly in the "Relations" (0.909) and "Counting" (0.839) aspects. This indicates AMS reliably captures complex compositional semantics and aligns closely with human evaluative standards, emphasizing its robustness as a metric for fine-grained image-text alignment evaluation.

### A.3 Experiments Compute Resources

We conduct our experiments on 8 Nvidia A100 GPUs.

### A.4 Broader Impact

Multi-modal image generation has wide-ranging applications in areas such as creative design, virtual reality, advertisement, and human-computer interaction. However, the powerful capabilities of these models also pose potential risks, particularly in generating toxic, biased, or harmful visual content. For instance, the human-centric images in our benchmark could be misused to produce misleading or inappropriate material. MMIG-Bench aims to support fair and responsible research by providing a diverse and high-quality dataset while actively mitigating these risks. To this end, we apply thorough filtering to remove toxic, sensitive, or low-quality content from our benchmark. Nevertheless, we encourage the community to consider ethical implications when developing and deploying such models and benchmarks.

### A.5 Instruction Templates for Prompt Generation

We carefully design eight instruction templates to generate prompts that encompass compositionality, common sense, and diverse stylistic variations. For example, the first template follows a fixed structure: `[scene description] + [attribute][entity1] + [interaction (spatial or action)] + [attribute][entity2]`, which guides GPT-4o to produce prompts that include background context, objects, attributes, and relations. In later templates, we provide GPT-4o with detailed instructions and examples to encourage the generation of prompts that are natural, imaginative, professionally written, or that incorporate elements such as negation, comparison, and numeracy.

**Instruction Template for T2I Prompts Generation (fixed pattern)**

Please generate natural sentences following a format of "[scene description] + [attribute][entity1] + [interaction (spatial or action)] + [attribute][entity2]"; follow the rules below:

1. "entity" should be common objects; e.g., chair, dog, car, lamp, etc. "entity2" is optional. Use "{entity}" as entity1 here.
2. "attribute" should be an adjective that describes "shape / color / material / size / condition / etc."
3. "interaction" should describe the relationship between "entity1" and "entity2". "spatial interaction" can be "on the left of / on the right of / on / on top of / on the bottom of / beneath / on the side of / neighboring / next to / touching / in front of / behind / with / etc."; "action interaction" can be any action happening between "entity1" and "entity2", such as "play with, eat, sit, place, hold, etc."
4. "scene description" is the background where the entities appear. It can contain other objects. It is optional.
5. The "interaction action" can be either in active or passive voice.
6. The order of these terms should not be fixed, as long as the sentence still looks natural. E.g., "scene description" can be put at the end.

**Instruction Template for T2I Prompts Generation (natural)**

Please generate prompts in a NATURAL format. It should contain one or more "entities / nouns", (optional) "attributes / adjective" that describes the entities, (optional) "spatial or action interactions" between entities, and (optional) "background description". Randomly ignore one or more items from [attributes, interactions, background]. One of the entities should be "{entity}".

**Instruction Template for T2I Prompts Generation (unreal)**

Please generate prompts in a NATURAL format. It should contain one or more "entities / nouns", (optional) "attributes / adjective" that describes the entities, (optional) "spatial or action interactions" between entities, and (optional) "background description". Note that:

1. Randomly ignore one or more items from [attributes, interactions, background].
2. The description should be imaginative. If imaginative, an example: "A robot and a dolphin dancing under the ocean, surrounded by swirling schools of fish".
3. Avoid repeating sentences you've already generated.

### A.6 Text-Image-Conditioned Dataset Overview

An overview of our comprehensive MMIG-Bench is shown in Fig. 8. Based on the 207 common entities we curated, we collect 386 reference image groups, each containing 3–5 multi-view, object-

**Instruction Template for T2I Prompts Generation (professional)**

Imagine that you are a professional designer, please write prompt for testing text-to-image diffusion models. The prompts should look like natural sentences. Please do not include descriptions about styles, such as "minimalism meets hygge vibes / editorial photoshoot style / baroque detail / etc.". One of the entities/nouns should be "{entity}".

**Instruction Template for T2I Prompts Generation (negation)**

Please generate prompts in a NATURAL format. It should contain one or more "entities / nouns", (optional) "attributes / adjective" that describes the entities, (optional) "spatial or action interactions" between entities, and (optional) "background description". Note that:

1. Randomly ignore one or more items from [attributes, interactions, background].
2. It should include the logic of "negation", such as the examples below:
"The girl with glasses is drawing, and the girl without glasses is singing.",
"In the supermarket, a man with glasses pays a man without glasses.",
"The larger person wears a yellow hat and the smaller person does not.",
"Adjacent houses stand side by side; the left one sports a chimney, while the right one has none.",
"A tailless, not black, cat is sitting.",
"A smiling girl with short hair and no glasses.",
"A bookshelf with no books, only a single red vase.".
One of the entities/nouns should be "{entity}".

**Instruction Template for T2I Prompts Generation (comparison)**

Please generate prompts in a NATURAL format. It should contain one or more "entities / nouns", (optional) "attributes / adjective" that describes the entities, (optional) "spatial or action interactions" between entities, and (optional) "background description". Note that:

1. Randomly ignore one or more items from [attributes, interactions, background].
2. It should have the logic of "comparison", such as the examples below:
"In a magnificent castle, a red dragon sits and a green dragon flies.",
"A magician holds two books; the left one is open, the right one is closed.",
"One cat is sleeping on the table and the other is playing under the table.".
"A green pumpkin is smiling happily, while a red pumpkin is sitting sadly.",
One of the entities/nouns should be "{entity}".

**Instruction Template for T2I Prompts Generation (counting)**

Please generate prompts in a NATURAL format. It should contain one or more "entities / nouns", and "numeracy" that describes the number of the entity.
Follow the six examples below:

1. four dogs played with two toys.
2. two chickens, four pens and one lemon.
3. Five cylindrical mugs beside two rectangular napkins.
4. three helicopters buzzed over two pillows.
5. Three cookies on a plate.
6. A group of sheep being led by two shepherds across a green field.

Avoid repeating sentences you've already generated.

#### Instruction Template for T2I Prompts Generation (numeracy in fixed structure)

Please generate natural sentences following a format of "[scene description (optional)] + [number][attribute][entity1] + [interaction (spatial or action)] + [number (optional)][attribute][entity2]"; follow the rules below:

1. "entity" should be common objects; e.g., chair, dog, car, lamp, etc. "entity2" is optional. Use "entity" as entity1 here.
2. "attribute" should be an adjective that describes "shape / color / material / size / condition / etc."
3. "number" should be "two/three/four/..." before the attribute, indicating the number of entities. It is optional for entity2.
4. "interaction" should describes the relationship between "entity1" and "entity2". "spatial interaction" can be "on the left of / on the right of / on / on top of / on the bottom of / beneath / on the side of / neighboring / next to / touching / in front of / behind / with / and / etc."; "action interaction" can be any action happening between "entity1" and "entity2", such as "play with, eat, sit, place, hold, etc."
5. "scene description" is the background where the entities appear. It can contain other objects. It is optional.
6. The "interaction action" can be either in active or passive voice.
7. The order of these terms should not be fixed, as long as the sentence still looks natural. E.g., "scene description" can be put at the end.

#### Prompt Template for Text Prompts Aspect Extraction

You need to analyze the query to a aspect graph that matches all the objects, relations (e.g.,spatial relations, action, complex relation), attributes, and counting (number of objects). Please ignore all the redundant phrases that are irrelevant to the contents of the image in the query, for example, 'a photo/picture of something, 'something in the background' etc., should not appear in the parsed graph.
Please also remove all the redundent aspects in the parsed graph. Here are some examples, if there are no such aspect, you can use an empty list to represent:
For the counting information, please ignore the object numbers that less than 2 (<2).

Context:

A group of women is playing the piano in the room.
{'Objects':['woman','room'],
'Other Relations':['play piano'],
'Spatila Relations':['in, (the room)'],
'Attributes':[],
'Counting':['a group of, (Non-specific quantity of woman)']}

Two Chihuahuas run after a child on a bicycle.
{'Objects':['Chihuahua','child','bicycle'],
'Other Relations':['runs after, (Chihuahua runs after child)','ride, (ride by the child)'],
'Spatila Relations':['on, (child on bicycle)']
'Attributes':['Chihuahua, (Chihuahua is a breed of dog)'],
'Counting':[Two (number of Chihuahua)]} }

A Delta Boeing 777 taxiing on the runway.
{'Objects':['Delta Boeing 777','runway'],
'Other Relations':['taxiing on, ( the runway)'],
'Spatial Relations':['on (plane on the runway)'],
'Attributes':['None'], 'Counting':[]}

Please extrace all the aspects precisely!

Prompt Template for AMS QA Pair Generation

Given an image and its corresponding caption, generate Visual Question Answering pairs that assess the presence of specific objects, attributes, relations, and counting information in the image.
The questions should be phrased naturally, appropriate, and reasonable. Input:
Caption: Two dogs are fighting over a red Frisbee that is bent in half. Target Elements:
{{"Objects": ["dog", "Frisbee"], "Relations": ["fighting over, (dogs fighting over Frisbee)"], "Attributes": ["red, (color of Frisbee)", "bent in half, (condition of Frisbee)"], "Counting": ["two, (number of dogs)"]}}

Example Output (JSON):
{{"question": "Is there a dog in the image?", "answer": "Yes", "Aspect":'Objects'}},
{{"question": "Is there a Frisbee in the image?", "answer": "Yes","Aspect":'Objects'}},
{{"question": "Are the dogs fighting over a Frisbee?", "answer": "Yes","Aspect":'Relations'}},
{{"question": "Is the Frisbee red?", "answer": "Yes","Aspect":'Attributes'}},
{{"question": "Is the Frisbee bent in half?", "answer": "Yes","Aspect":'Relations'}},
{{"question": "Are there two dogs in the image?", "answer": "Yes","Aspect":'Counting'}}
If the counting aspect is related to 'one, (number of something)', please ignore it!
Please reduce the redundancy of the questions, don't repeat !
If the question includes relational references—such as friend, mother, daughter, etc.—please specify the associated referent (for example, the woman's friend).
If the aspect entity has no practical significance, please ignore it.
Input:
Caption:
Target Elements:
Output (JSON):

centric images, and generate 4,850 text prompts that include these entities. The prompts are densely labeled and exhibit rich, detailed semantics, covering compositionality, common sense, and styles.

### A.7 More Qualitative Results

We show more visual comparisons of the state-of-the-art models in Fig. 9, 10 and 11.

### A.8 Human Evaluation Interface

The Amazon Mechanical Turk interfaces used in the user studies are shown in Fig. 12-16. The study is divided into five categories to assess the compositionality of prompt-image alignment across different aspects: general prompt following (Fig. 12), object (Fig. 13), attribute (Fig. 14), relation (Fig. 15) and numeracy (Fig. 16). In each session, a randomly selected prompt-image pair is presented to the user, who is then asked to rate the generation quality using a 5-point scale. Each question is independently rated by three different workers to ensure reliability.


Figure 8: Overview of MMIG-Bench.

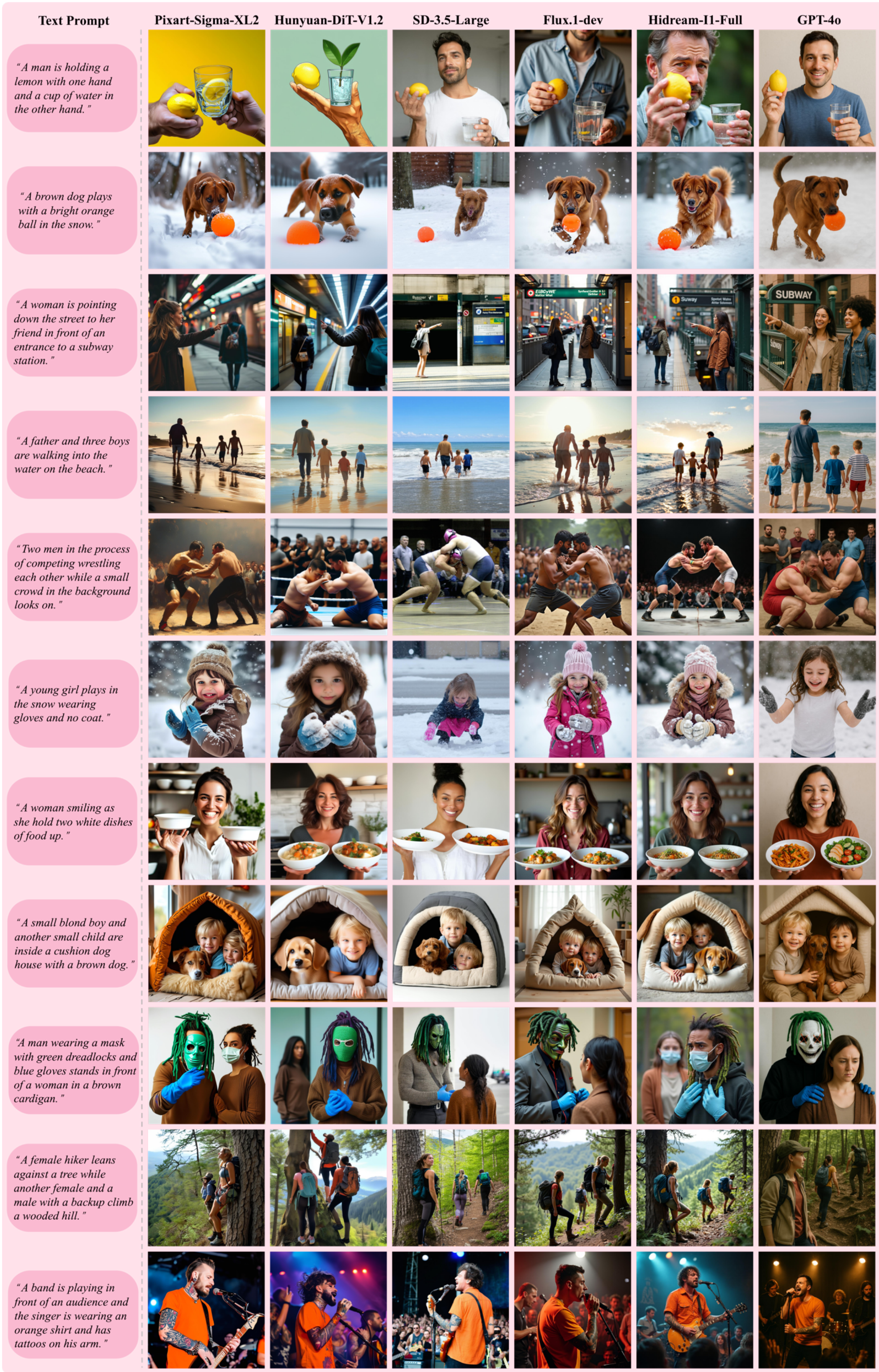


Figure 9: More qualitative results of text-only generation methods on MMIG-Bench.


Figure 10: More qualitative results of text-only generation methods on MMIG-Bench.


Figure 11: More qualitative results of text-image-conditioned generation methods on MMIG-Bench.

**Evaluate the quality of the generated image** (Click to expand)

A text description and an image are displayed below. Please evaluate how well the image matches the description.

Text description: In a cozy kitchen, a man holds fresh bread, while a woman with short hair does not hold any.

○ 1: No match – The image is completely unrelated to the description.

○ 2: Poor match – The image has major discrepancies and only loosely relates to the description.

○ 3: Partial match – The image captures some key elements but contains multiple minor discrepancies.

○ 4: Good match – The image mostly aligns with the description, with only a few minor discrepancies.

○ 5: Perfect match – The image fully matches the description with no noticeable discrepancies.

Figure 12: The interface of user study for general prompt following.

**Evaluate the quality of the generated image** (Click to expand)

A text description and an image are displayed below. The key objects/entities in the description are highlighted in **bold**.

Please evaluate how well the image aligns with these **bolded** elements (e.g., check whether the specified objects are present in the image).

(If no text is bolded, evaluate how well the image matches the overall description.)

Text description: The **coffee table** in the shabby **living room** is littered with **book**s and **candle**s.

- ○ 1: No match – The image is completely unrelated to the description.
- ○ 2: Poor match – The image has major discrepancies and only loosely relates to the description.
- ○ 3: Partial match – The image captures some key elements but contains multiple minor discrepancies.
- ○ 4: Good match – The image mostly aligns with the description, with only a few minor discrepancies.
- ○ 5: Perfect match – The image fully matches the description with no noticeable discrepancies.

Figure 13: The interface of user study for prompt following on *Object*.

**Evaluate the quality of the generated image** (Click to expand)

A text description and an image are displayed below. Key attributes (**color**, **shape**, **condition**, etc.) in the description are highlighted in **bold**.

Please evaluate how well the image aligns with these **bolded** elements (e.g., whether the specified attributes are accurately represented).

(If no text is bolded, evaluate how well the image matches the overall description.)

Text description: beneath a **clear twilight** sky, the **flowing** dress rests next to a **bright**, **metal** lamp.

- 1: No match – The image is completely unrelated to the description.
- 2: Poor match – The image has major discrepancies and only loosely relates to the description.
- 3: Partial match – The image captures some key elements but contains multiple minor discrepancies.
- 4: Good match – The image mostly aligns with the description, with only a few minor discrepancies.
- 5: Perfect match – The image fully matches the description with no noticeable discrepancies.

Figure 14: The interface of user study for prompt following on *Attributes*.

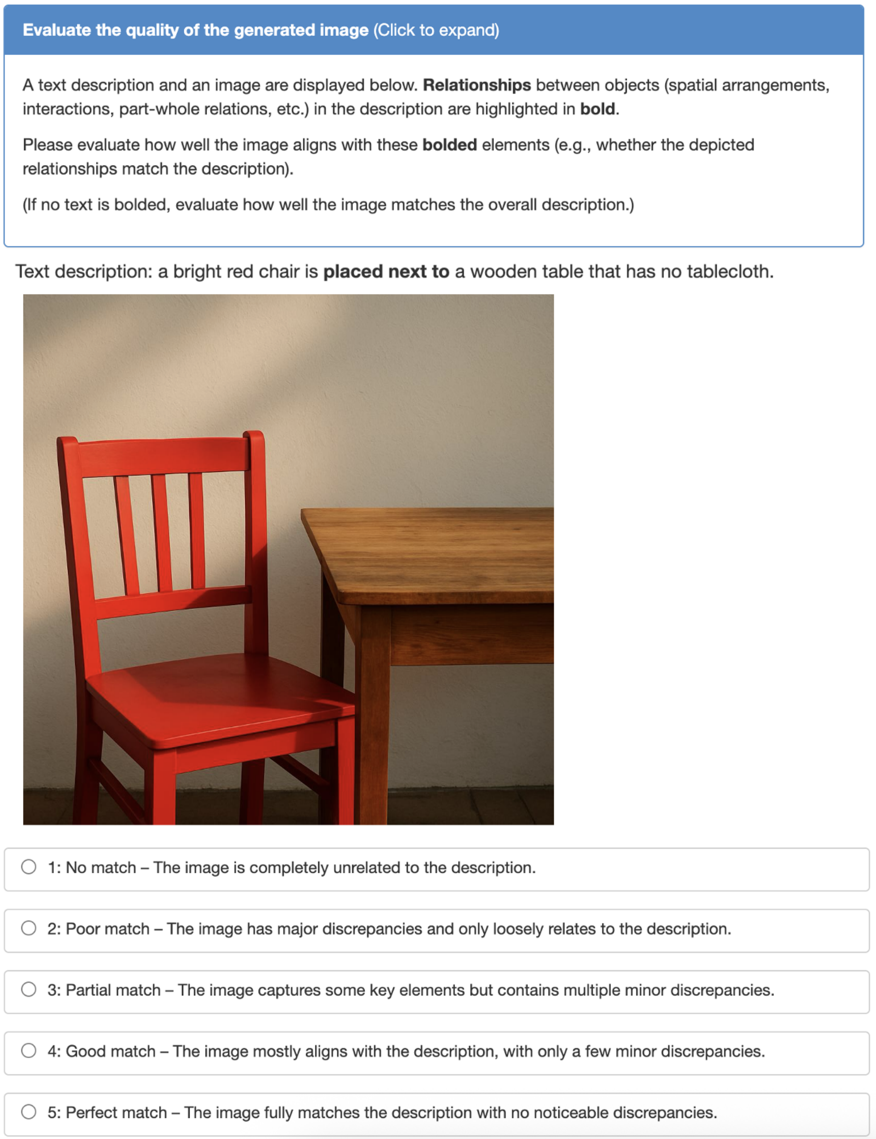


Figure 15: The interface of user study for prompt following on *Relations*.

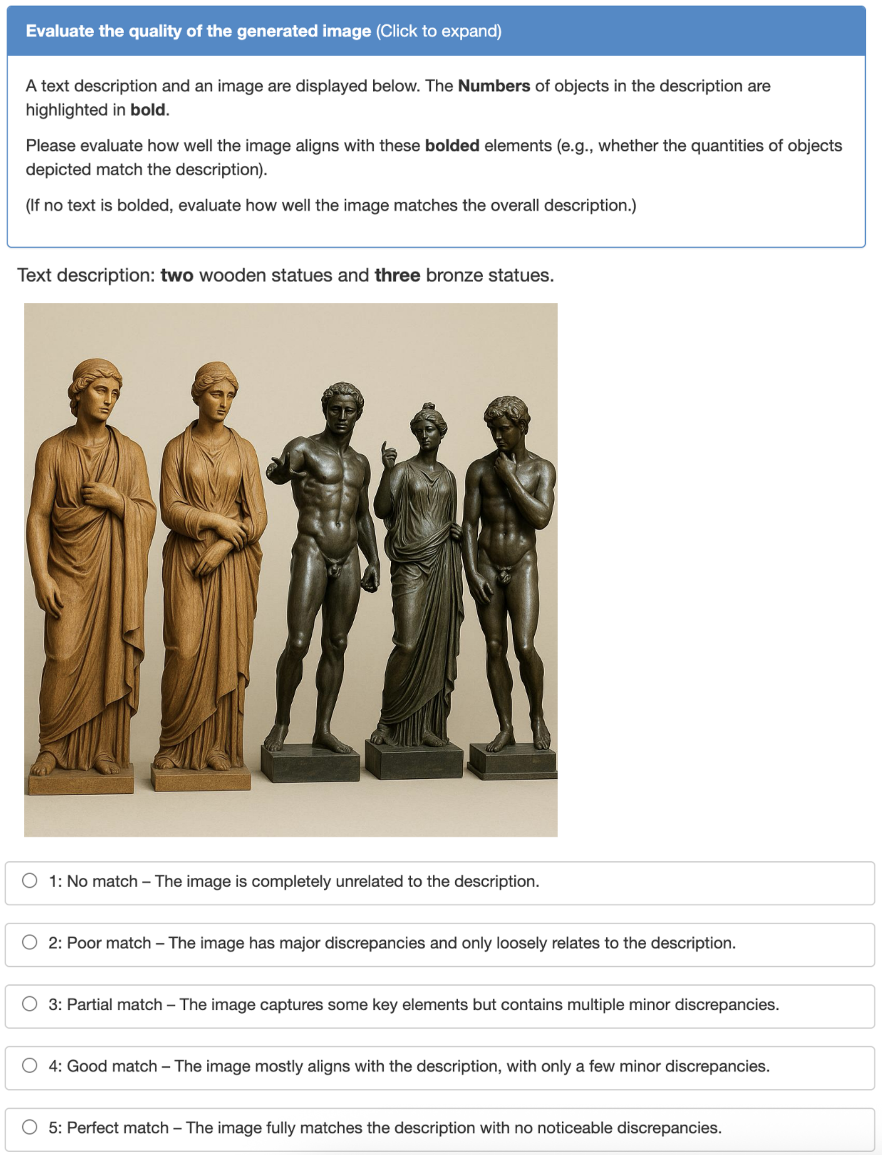


Figure 16: The interface of user study for prompt following on *Numeracy*.