A Appendix

A.1 Qualitative Results of MMGen-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 MMGen-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

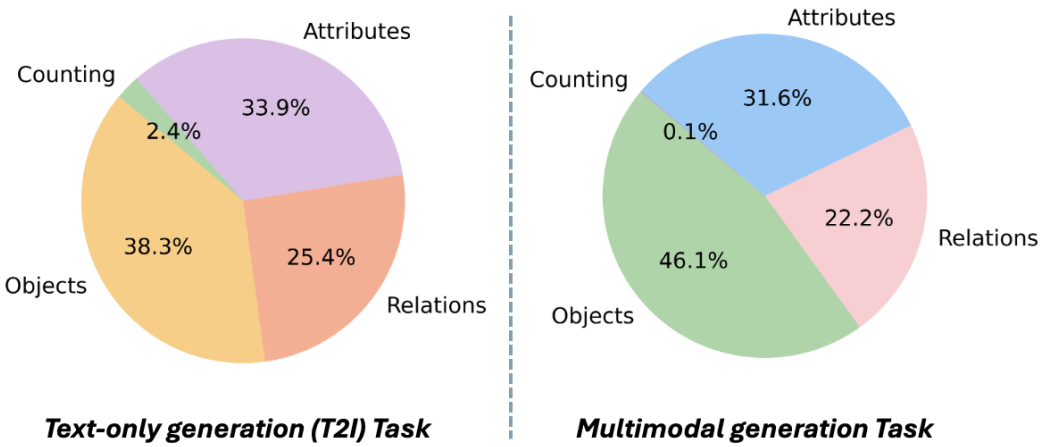

Figure 6: Aspect Distribution of the QA pairs of **AMS**.

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

Aspect	Objects ↑	Relations ↑	Attributes ↑	Counting ↑	Overall ↑
Spearman ρ	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.

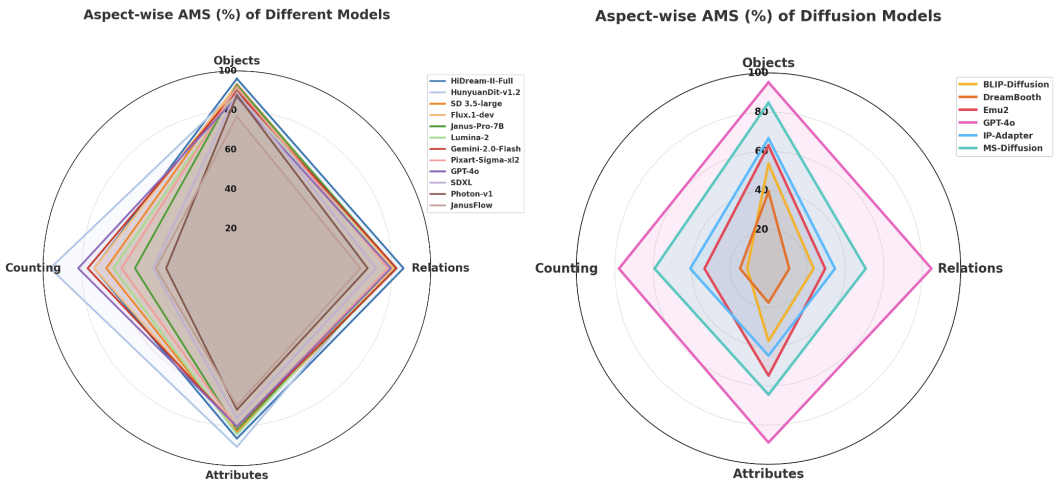

Figure 7: The AMS of different models on the text-only generation (T2I) task (left) and the multimodal generation task (right).

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. MMGen-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 MMGen-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 MMGen-Bench.

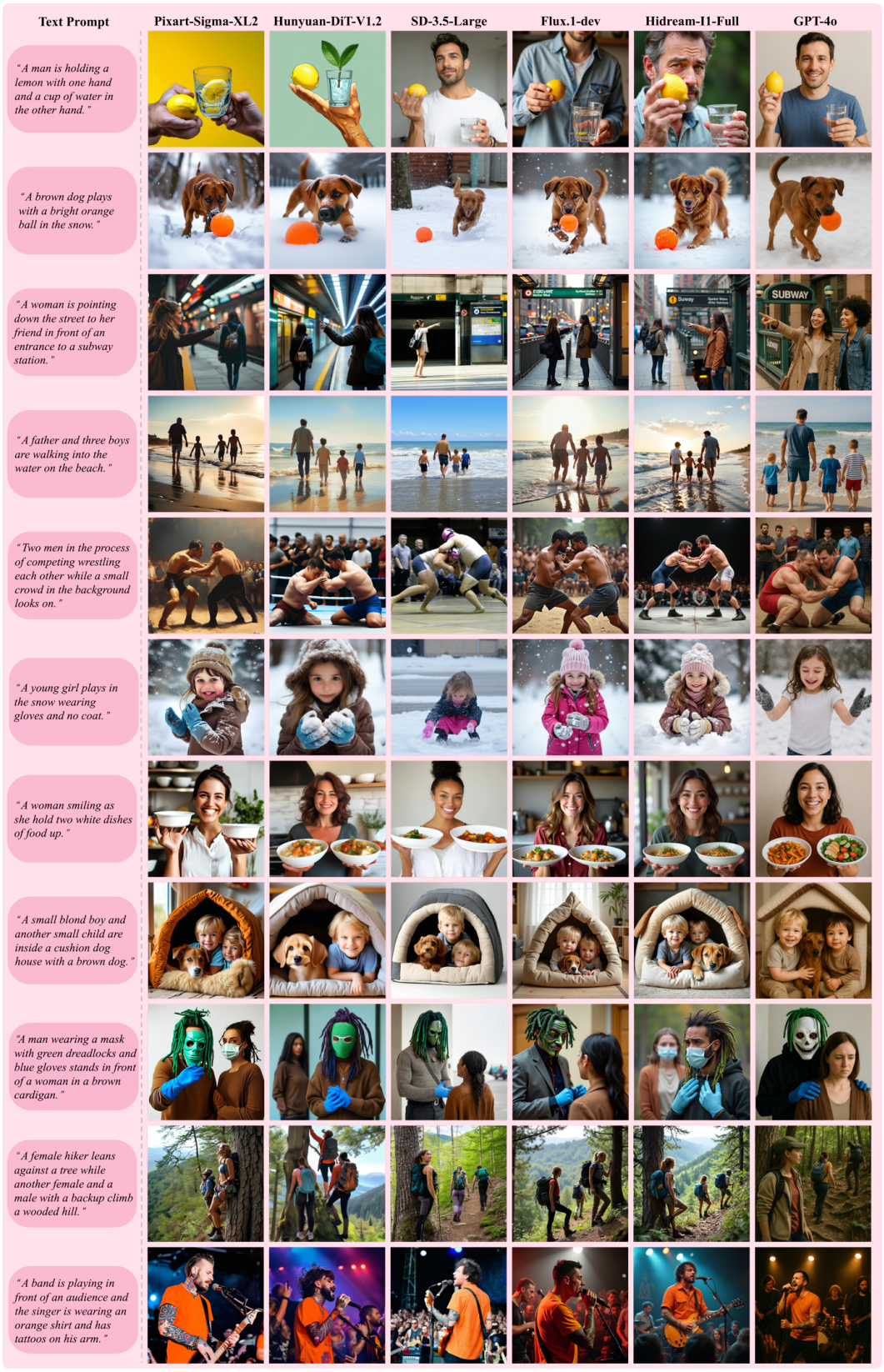

Figure 9: More qualitative results of text-only generation methods on MMGen-Bench.

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

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

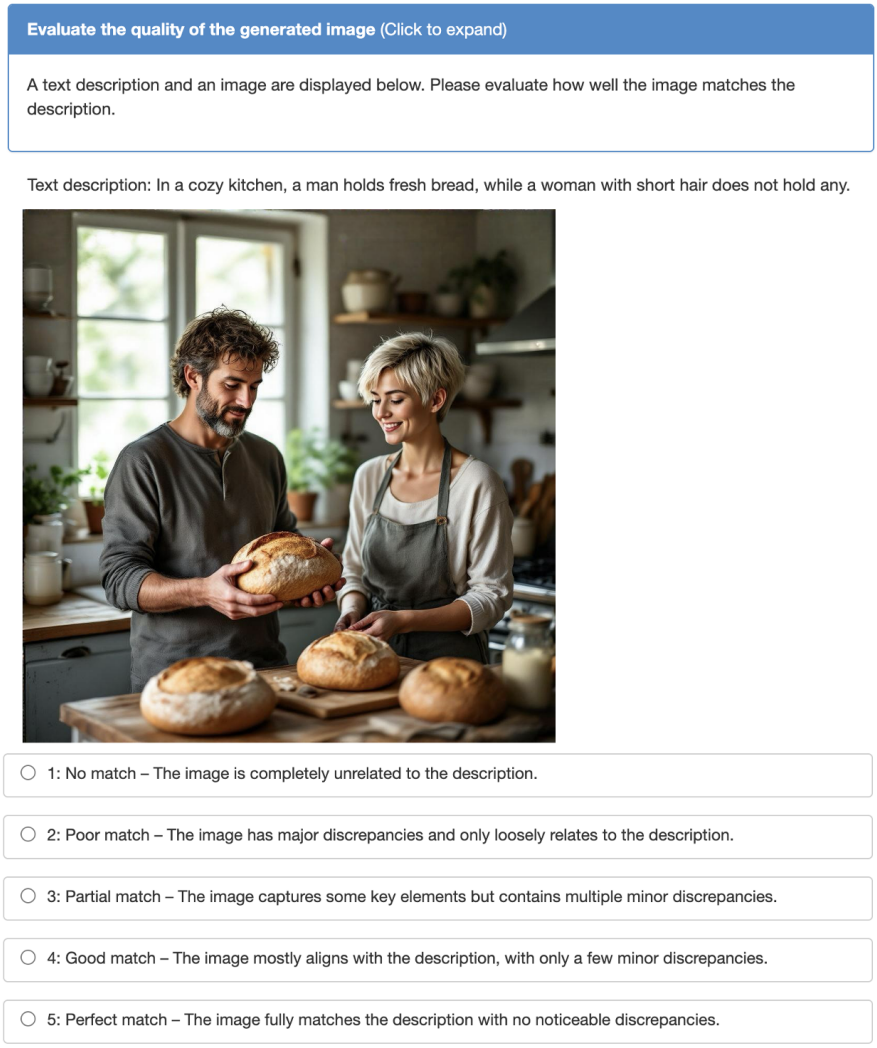

Figure 12: The interface of user study for general prompt following.

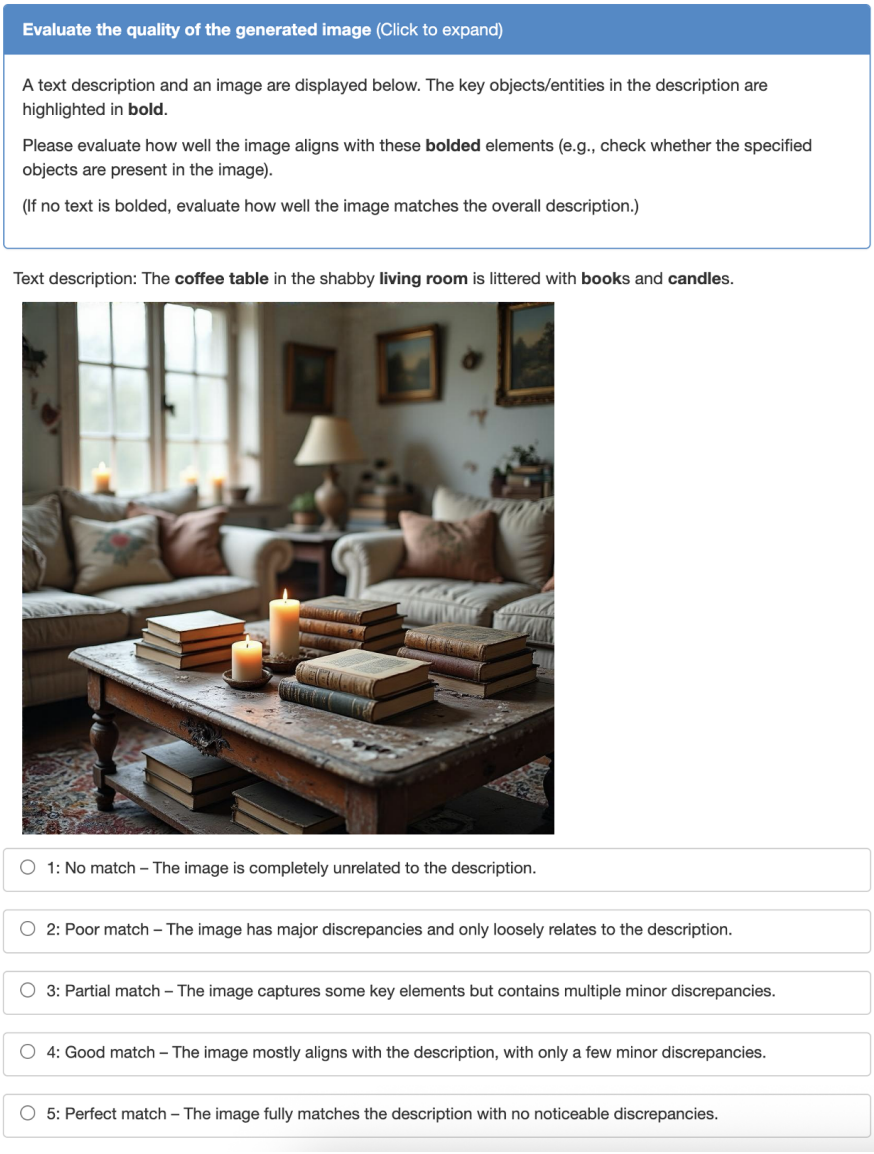

Figure 13: The interface of user study for prompt following on *Object*.

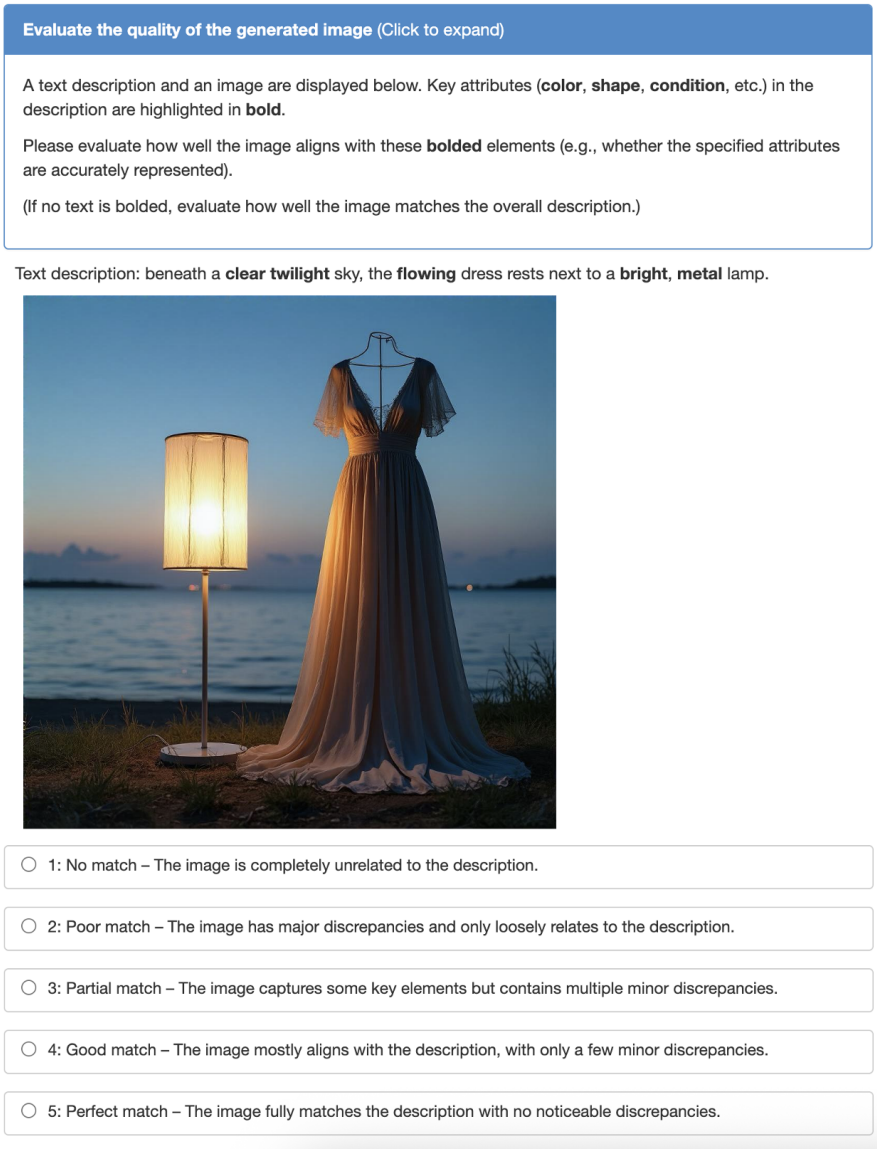

Figure 14: The interface of user study for prompt following on *Attributes*.

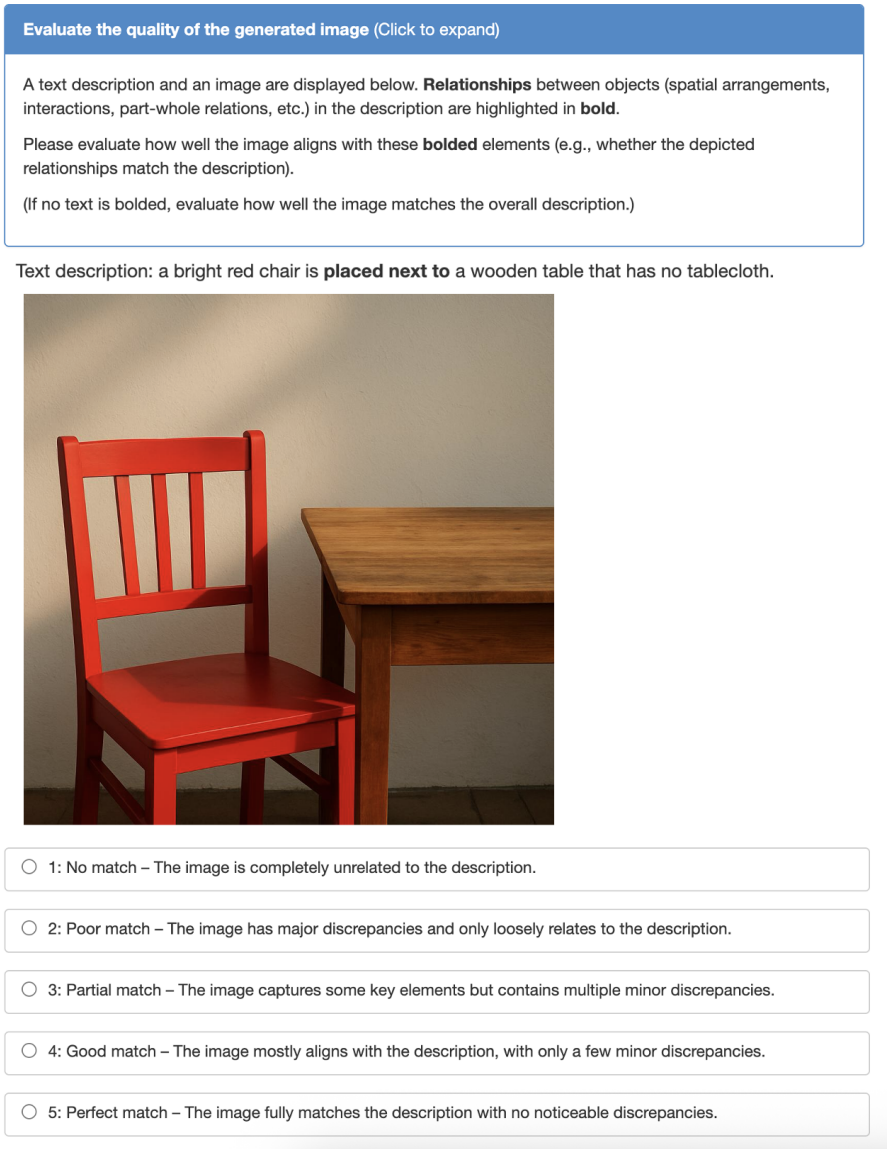

Figure 15: The interface of user study for prompt following on *Relations*.

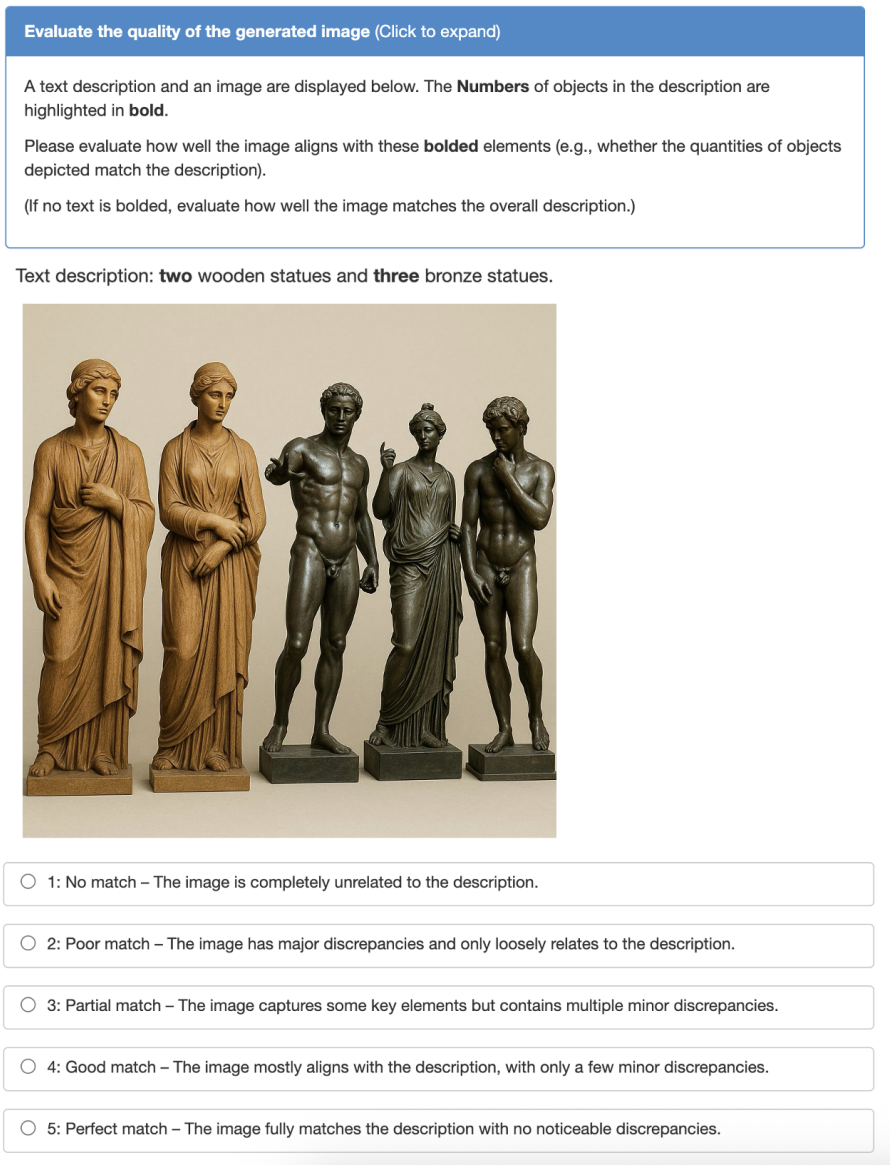

Figure 16: The interface of user study for prompt following on *Numeracy*.