

490 A Appendix

491 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).

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

496 A.2 Quantitative and Qualitative Results of AMS

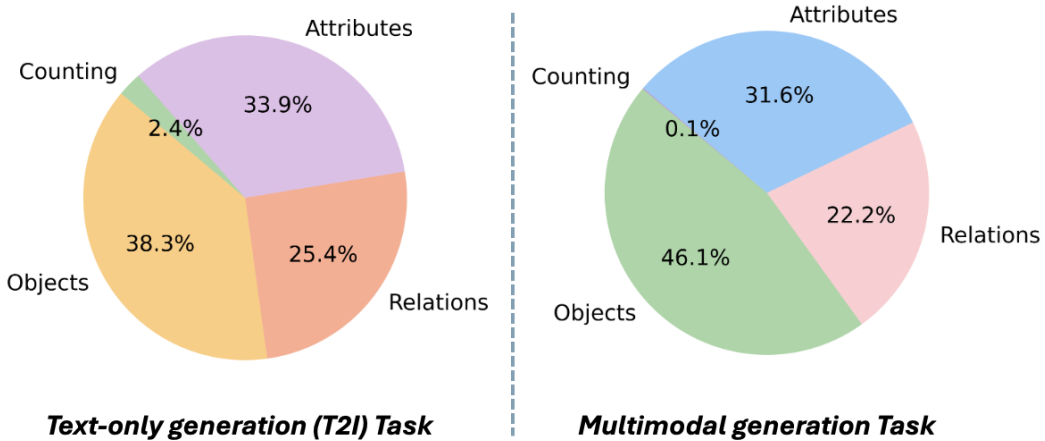


Figure 6: Aspect Distribution of the QA pairs of AMS.

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

Aspect	Objects $\uparrow$	Relations $\uparrow$	Attributes $\uparrow$	Counting $\uparrow$	Overall $\uparrow$
Spearman ρ	0.469	0.909	0.601	0.839	0.699

497 As depicted in Figure 6, the distribution of aspect types differs notably between the text-only
 498 generation (T2I) and multi-modal generation tasks. In the T2I setting, “Objects” dominate with
 499 38.3%, while “Attributes” and “Relations” also constitute substantial proportions (33.9% and 25.4%,
 500 respectively). In multi-modal generation, “Objects” and “Attributes” remain prominent (46.1% and
 501 31.6%, respectively), but the relative proportion of “Relations” decreases significantly (22.2%). The
 502 presence of “Counting” (0.1%) questions suggests this aspect is less frequent in the customized T2I
 503 generation task.

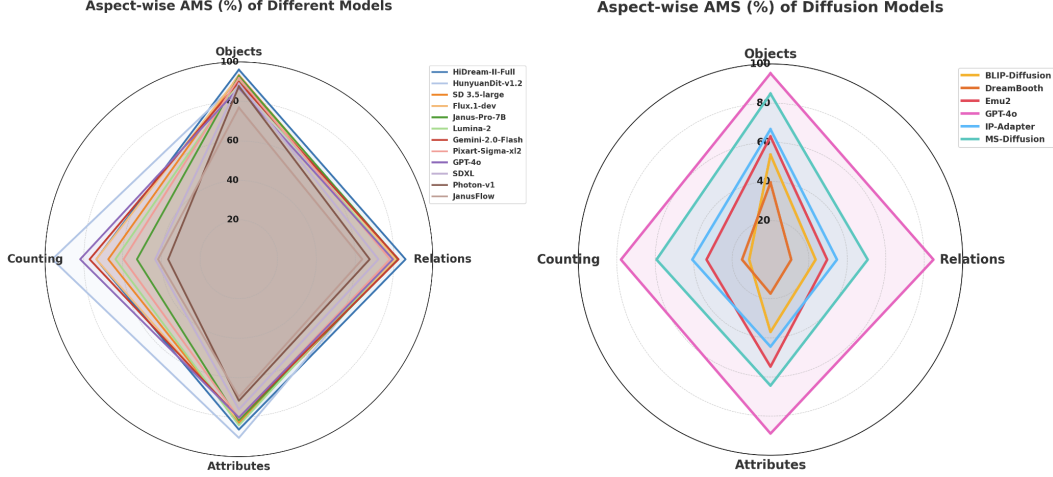


Figure 7: The AMS of different models on the text-only generation (T2I) task (left) and the multimodal generation task (right).

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-II, 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.

534 A.5 Instruction Templates for Prompt Generation

535 We carefully design eight instruction templates to generate prompts that encompass composi-
536 tionality, common sense, and diverse stylistic variations. For example, the first template fol-
537 lows a fixed structure: [scene description] + [attribute][entity1] + [interaction
538 (spatial or action)] + [attribute][entity2], which guides GPT-4o to produce prompts
539 that include background context, objects, attributes, and relations. In later templates, we provide GPT-
540 4o with detailed instructions and examples to encourage the generation of prompts that are natural,
541 imaginative, professionally written, or that incorporate elements such as negation, comparison, and
542 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.

543 A.6 Text-Image-Conditioned Dataset Overview

544 An overview of our comprehensive MMGen-Bench is shown in Fig. 8. Based on the 207 common
545 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 redundant 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'],  
  'Spatial 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)'],  
  'Spatial 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.

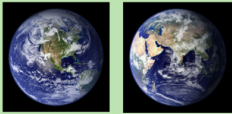
Text Prompts	Reference Images	Text Prompts	Reference Images
"A baseball sits behind a tall wooden bookshelf in a quiet library." "A baseball rests on a dusty shelf." "A baseball drifting gently in a cosmic sea of swirling purple stars."		"A bench next to a rusty bike stands on an old cobblestone street." "A bench with chipped green paint in a quiet park at sunset." "A bench floating on a drifting cloud, joined by tiny birds soaring close by."	
"A basketball bounces beside a tall lamp in a quiet living room." "A basketball rolls across the gym floor toward a rusty hoop." "A basketball sailing across moonlit waves."		"A butterfly flutters next to a shiny lamp on an old wooden desk." "A butterfly drifts over a silent windowsill." "A butterfly and a graceful phoenix dancing together in a midnight orchard."	
"A book rests on a small table in the library." "A book with a worn leather cover sits on a dusty shelf, illuminated by a single ray of sunlight." "A book with floating pages."		"A map lies behind the tall wooden chair near the fireplace." "A map with faint markings of hidden trails." "A map drifting quietly near ancient ruins."	
"A cookie rests beside a tall lamp on a tidy windowsill." "A cookie resting on a quiet windowsill next to a small potted plant." "A cookie with tiny sparkles perched on a giant teapot."		"A planet on the left of a metallic lamp glows softly in a cozy living room." "A planet shrouded in swirling violet clouds." "A planet cradles a tiny dog in its orbit above a suburban backyard."	
"A sandwich on a small ceramic plate sits in a cozy kitchen." "A sandwich with melted cheese and crispy lettuce sits on a wooden plate." "A sandwich with glowing neon cheese."		"A snowflake landed on a rusty mailbox by an old wooden fence." "A snowflake drifts gracefully in the cold breeze, dancing around tall icicles." "A snowflake gently perched on a candle's flickering flame."	
"A soccer ball stands alone next to a rusty metal chair by the dusty roadside." "A soccer ball with peeling paint rolls slowly across the dusty ground." "A soccer ball gliding through a neon-lit galaxy."		"A towel rests on a sturdy wooden chair in the sunny backyard." "A towel rests on a plain wooden chair." "A towel gently drapes over a silent statue, glimmering in the twilight of an abandoned courtyard."	

Figure 8: Overview of MMGen-Bench.

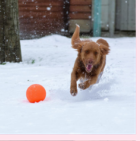
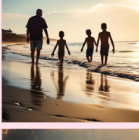
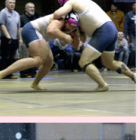
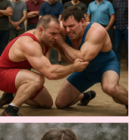
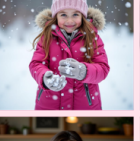
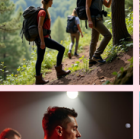
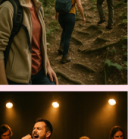
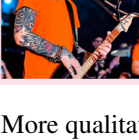
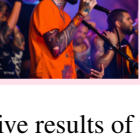
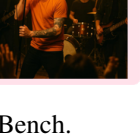
Text Prompt	Pixart-Sigma-XL2	Hunyuan-DiT-V1.2	SD-3.5-Large	Flux.1-dev	Hidream-I1-Full	GPT-4o
"A man is holding a lemon with one hand and a cup of water in the other hand."						
"A brown dog plays with a bright orange ball in the snow."						
"A woman is pointing down the street to her friend in front of an entrance to a subway station."						
"A father and three boys are walking into the water on the beach."						
"Two men in the process of competing wrestling each other while a small crowd in the background looks on."						
"A young girl plays in the snow wearing gloves and no coat."						
"A woman smiling as she hold two white dishes of food up."						
"A small blond boy and another small child are inside a cushion dog house with a brown dog."						
"A man wearing a mask with green dreadlocks and blue gloves stands in front of a woman in a brown cardigan."						
"A female hiker leans against a tree while another female and a male with a backpack climb a wooded hill."						
"A band is playing in front of an audience and the singer is wearing an orange shirt and has tattoos on his arm."						

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

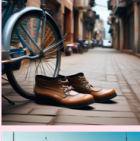
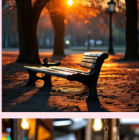
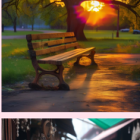
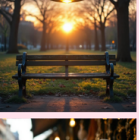
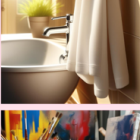
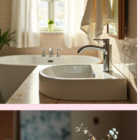
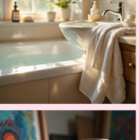
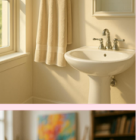
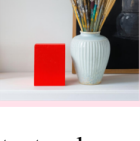
Text Prompt	Pixart-Sigma-XL2	Hunyuan-DiT-V1.2	SD-3.5-Large	Flux.1-dev	Hidream-I1-Full	GPT-4o
"In the quiet study, a slender blue bottle rests on top of a weathered oak desk."						
"In a quiet office, a sleek computer sits next to a vintage chair."						
"In the cozy living room, a sleek television sits next to a petite wooden chair."						
"In the bright kitchen, a handwoven basket sits next to a shiny lamp."						
"In a busy downtown alley, the scuffed shoes rested beside a shiny bicycle."						
"Amid the bustling marina, a sleek boat sailed past a bright orange lifebuoy."						
"At a bustling outdoor fair, a freshly baked cake sits atop a rustic wooden table."						
"In a quiet park at sunset, a vibrant light hovers above a weathered bench."						
"In a bustling artisan market, the delicate jewelry sparkles next to a worn-out clock."						
"In a sunlit bathroom, the soft towel hangs next to a gleaming sink."						
"In the busy art studio, a vibrant red block sits beside an elegant porcelain vase."						

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

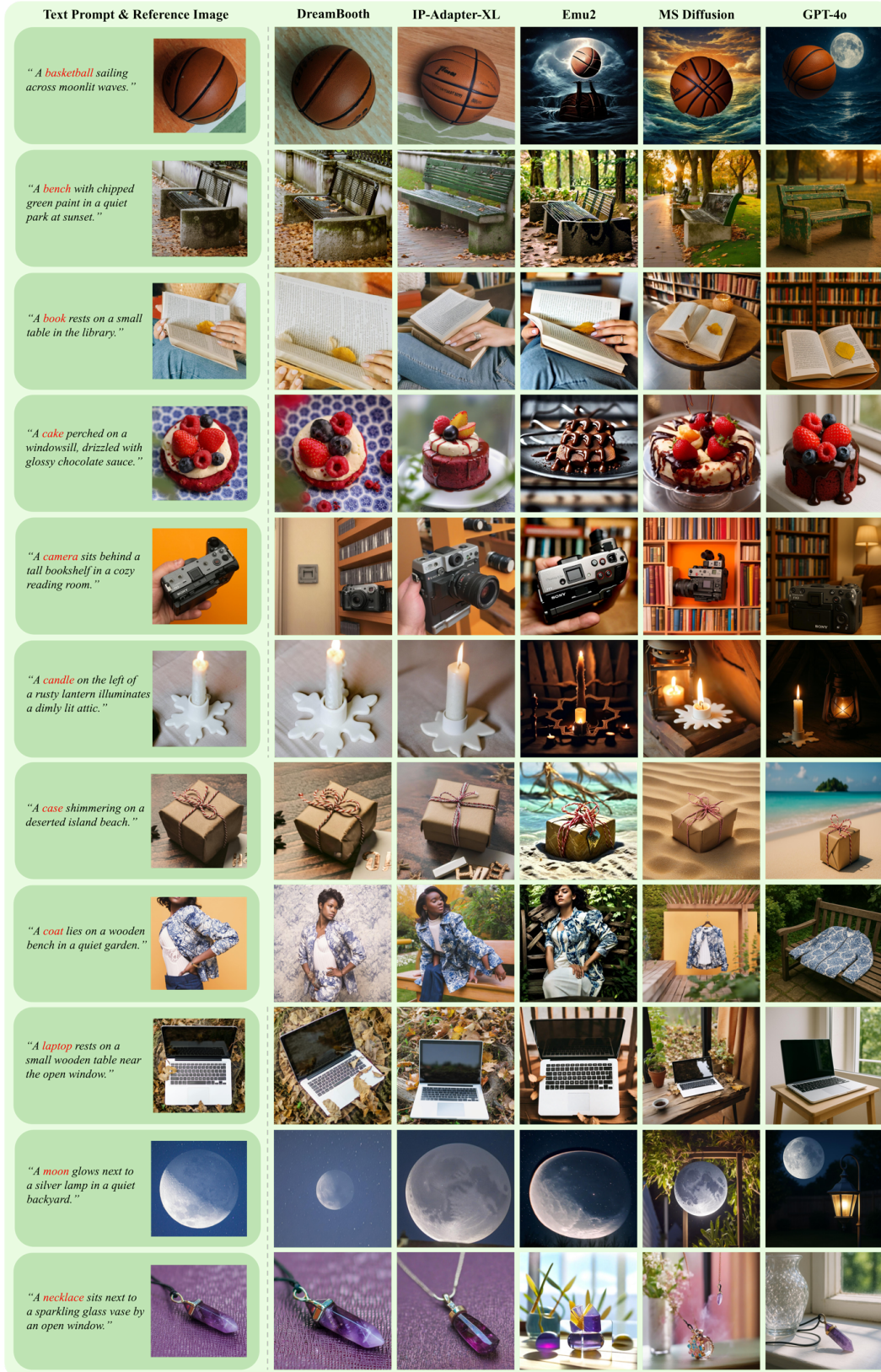


Figure 11: More qualitative results of text-image-conditioned generation methods on MMGen-Bench.

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.



The image shows a man and a woman in a kitchen. The man is holding a large round loaf of bread. The woman is standing next to him, looking at the bread. There are several other loaves of bread on a wooden table in front of them. The kitchen has a window in the background and various kitchen items on shelves.

☐ 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 **books** and **candles**.



- ☐ 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*.

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

A text description and an image are displayed below. **Relationships** between objects (spatial arrangements, interactions, part-whole relations, etc.) in the description are highlighted in **bold**.

Please evaluate how well the image aligns with these **bolded** elements (e.g., whether the depicted relationships match the description).

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

Text description: a bright red chair is **placed next to** a wooden table that has no tablecloth.



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

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

A text description and an image are displayed below. The **Numbers** of objects in the description are highlighted in **bold**.

Please evaluate how well the image aligns with these **bolded** elements (e.g., whether the quantities of objects depicted match the description).

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

Text description: **two** wooden statues and **three** bronze statues.



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