

APPENDIX

A MORE DETAILS ON PROMPT COLLECTION PROCESS

Idiom Interpretation. In idiom collection, we leverage a book titled “The Exhaustive List of American Idioms” [idi \(2023\)](#), which systematically documents over 11k idioms. These idioms were collected from diverse sources, including TV shows, movies, and everyday conversations. Each idiom in the book is accompanied by its actual meaning in context. In addition to this resource, we also refer to idioms available on the Internet. From this extensive pool, we manually select 200 idioms that are commonly used in daily life and challenging for T2I models due to their figurative meanings. We input the selected idioms and their actual meanings into an LLM and prompt it to generate new sentences. These sentences are designed to describe visible scenes involving the idioms, providing contextual clues for reasoning while avoiding directly revealing the idiom’s meaning.

Textual Image Design. For textual image design, we collect 6 types of text-rich images from 4 distinct sources.

(1) LLAVar-2 Dataset [Zhou et al. \(2024\)](#): This dataset contains 42k text-rich images sourced from LAION [Schuhmann et al. \(2022\)](#), representing various categories such as quotes, memes, book covers, posters, and product packaging. However, images in this dataset are of various quality and formats, so we filter out 80 aesthetically pleasing images that have a resolution greater than 384x384 and exhibit clear design intentions.

(2) InfographicVQA Dataset [Mathew et al. \(2022\)](#): This dataset comprises 5k high-quality infographics. We select 40 with normal height-width ratio that exemplify well-crafted layouts to convey structured information.

(3) POSTA Dataset [Chen et al. \(2025a\)](#): This dataset includes over 300 posters with professional background, layout, and text formats designed by experts. We select 40 posters that demonstrate a balance between text and visual design elements.

(4) CoSyn-400k Dataset [Yang et al. \(2025\)](#): This dataset consists of 400k synthetic text-rich images, generated by LLM-drive codes. These images cover diverse formats, such as charts, diagrams, tables, documents (e.g., menus or business cards), math examples, and musical scores. From this dataset, we select 40 samples that exemplify structured text-visual integration, including 10 tables, 10 diagrams, and 20 documents.

The design intentions of the collected images are summarized using Qwen2.5-VL [Bai et al. \(2025\)](#), yielding 200 prompts in this dimension.



Figure 4: Word cloud to visualize the word distribution of each dimension in our prompt suite.

B EVALUATION FRAMEWORK

We adopt Qwen2.5-VL [Bai et al. \(2025\)](#) as our evaluation tool due to its state-of-the-art performance in visual-textual grounding and fine-grained object recognition. The evaluation of T2I models on our benchmark focuses on two key aspects: reasoning accuracy and image quality. To assess this, we generate specific pairs of question and criterion for each prompt with DeepSeek-R1 [Guo et al. \(2025\)](#)

Table 4, 5, 6, and 7 present the templates used to generate the prompt-specific question-criterion pairs for Idiom Interpretation, Textual Image Design, Entity-Reasoning and Scientific-Reasoning, respectively. Each template is tailored to focus on the unique aspects of its corresponding dimension.

Table 8 presents the template used to evaluate the generated images for all four dimensions. Only the evaluation question-criterion pairs need to be replaced for each prompt.

C QUALITATIVE EXAMPLES

	GPT-Image-1	Gemini 2.0	playground-v2.5	SD-3.5-large	Emu3
He told a funny joke to <u>break the ice</u> at the start of the meeting.					
At parties, she felt like a <u>wallflower</u> , too shy to join in on the dancing.					
Create an <u>infographic</u> summarizing American coffee consumption trends in 2020.					
Design a vibrant <u>promotional poster</u> for a tropical theme park...					
The <u>US president</u> giving a speech at Brandenburg Gate in 1987					
The first <u>mammal</u> successfully cloned from an adult somatic cell in 1996					
A <u>trampoline</u> with a bowling ball, a basketball, and a ping pong ball placed on it					
A <u>seesaw</u> with a 1-kilogram cotton ball on one end and a 1-kilogram iron cube on the other.					

Figure 5: Qualitative examples.

<USER>: I have a text-to-image generation model that can generate images based on given prompts. However, the model is not perfect and may fail to accurately capture the meaning of the prompt or depict it correctly. Your task is to evaluate the generated image based on a specific prompt that contains an idiom. Given the prompt: {‘id’: {prompt_id}, ‘prompt’: {prompt}, ‘idiom’: {idiom}, ‘idiom_meaning’: {idiom_meaning}}, you need to:

1. identify what should be depicted in the image or the meaning the image should convey.
2. analyze the prompt and create a list of questions based on the key elements that the image should be checked against.
3. consider factors that could impact the aesthetics or visual quality of the image and list relevant questions.

Please also design a scoring criterion for each question, where a score of 1 means “yes (to the question)”, 0 means “no”, and 0.5 means “partially yes”.

Provide your answer in json format: {{‘id’: [prompt id], ‘prompt’: [the prompt], ‘image_content’: [what the image should convey], ‘reason_evaluation’: (here should be a dictionary with 3-5 pairs of question and criterion: ‘q1’: [question 1], ‘c1’: [criterion 1], ‘q2’: [question 2], ‘c2’: [criterion 2]...), ‘quality_evaluation’: (same format as ‘reason_evaluation’ with 1-3 pairs of question and criterion)}}.

<ASSISTANT>:

```
{
  "id": 1,
  "prompt": "He told a funny joke to break the ice at the start of the meeting.",
  "image_content": "The image should depict a person telling a joke in a meeting setting, with the audience showing visible signs of relaxation and social easing (e.g., smiling, laughing, or beginning to converse) as a result. It must convey the metaphorical meaning of ‘breaking the ice’—reducing tension and initiating interaction—without literal depictions of ice or breaking.",
  "reason_evaluation": {
    "q1": "Does the image show a meeting environment (e.g., conference table, office setting, or group in professional attire)?",
    "c1": "1: Clear meeting setting; 0.5: Ambiguous setting (e.g., generic group); 0: No meeting context",
    "q2": "Is a person visibly acting as the joke-teller (e.g., speaking, gesturing, or commanding group attention)?",
    "c2": "1: Clear speaker with active role; 0.5: Speaker present but passive; 0: No identifiable speaker",
    "q3": "Do audience members display relaxed/post-icebreaking reactions (e.g., genuine smiles, laughter, or eased body language)?",
    "c3": "1: Clear reactions showing tension relief; 0.5: Mild reactions (e.g., slight smiles); 0: No reactions or negative expressions",
    "q4": "Is the idiom depicted metaphorically (avoiding literal ice/breaking)?",
    "c4": "1: Entirely metaphorical; 0.5: Partial literal elements (e.g., ice visuals but not dominant); 0: Literal depiction (e.g., shattered ice)"
  },
  "quality_evaluation": {
    "q1": "Is the image visually clear with well-defined subjects (not blurry or distorted)?",
    "c1": "1: High clarity; 0.5: Minor blur/distortion; 0: Major quality issues",
    "q2": "Is the composition balanced (e.g., speaker and audience arranged cohesively, no awkward cropping)?",
    "c2": "1: Harmonious composition; 0.5: Minor imbalances; 0: Chaotic or disjointed layout",
    "q3": "Are colors/lighting appropriate and enhance the mood (e.g., warm tones for positivity, no visual clashes)?",
    "c3": "1: Lighting/colors support the scene; 0.5: Partially fitting; 0: Inconsistent or jarring"
  }
}
```

Table 4: **Template used by DeepSeek-R1 to generate evaluation question-criterion pairs for the dimension of Idiom Interpretation and example output.** The text highlighted in cyan should be replaced with details from the specific prompt.

<USER>: I have a text-to-image generation model that can generate images based on given prompts. However, the model is not perfect and may fail to accurately reflect the prompt or depict the details correctly. Given a prompt which is a design intention for a text-rich image like infographic or poster, your task is to evaluate whether the generated image correctly fulfill the design intention.

Here is the prompt: `{{'id': {prompt_id}, 'prompt': {prompt}}}`, you need to:

1. identify what should be depicted in the image and its functional purposes.
2. analyze the design intention and create a list of questions based on the key elements that the image should be checked against, including presence of required text elements.
3. consider factors that could impact the aesthetics or visual quality of the image and list relevant questions.

Please also design a scoring criterion for each question, where a score of 1 means "yes (to the question)", 0 means "no", and 0.5 means "partially yes".

Provide your answer in json format: `{{'id': [prompt id], 'prompt': [the prompt], 'image_content': [what the image should convey], 'reason_evaluation': (here should be a dictionary with 3-5 pairs of question and criterion: 'q1': [question 1], 'c1': [criterion 1], 'q2': [question 2], 'c2': [criterion 2]...), 'quality_evaluation': (same format as 'reason_evaluation' with 1-3 pairs of question and criterion)}}`.

Table 5: **Template used by DeepSeek-R1 to generate evaluation question-criterion pairs for the dimension of Textual Image Design.** The text highlighted in cyan should be replaced with details from the specific prompt.

<USER>: I have a text-to-image generation model that can generate images based on given prompts. However, the prompts given to the model may contain implicit meanings or entities that are not directly stated. Your task is to evaluate whether the generated image accurately represents the intended meaning of the prompt. Given the prompt: `{{'id': {prompt_id}, 'prompt': {prompt}, 'explicit_meaning': {explicit_meaning}}}`, you need to:

1. identify what should be depicted in the image in order to fully and accurately reflect the explicit meaning of the prompt.
2. identify the entity that the model needs to infer from the prompt, and create a list of questions that check whether the image has correctly identified and depicted this entity.
3. Consider other elements or details in the prompt (apart from the implicit entity), create a list of questions that check if the image accurately reflects these additional key elements.
4. consider factors that could impact the aesthetics or visual quality of the image and list relevant questions.

Please also design a scoring criterion for each question, where a score of 1 means "yes (to the question)", 0 means "no", and 0.5 means "partially yes".

Provide your answer in json format: `{{'id': [prompt id], 'prompt': [the prompt], 'explicit_meaning': [the explicit meaning], 'image_content': [what the image should depict], 'entity_evaluation': (here should be a dictionary with 1-3 pairs of question and criterion: 'q1': [question 1], 'c1': [criterion 1], 'q2': [question 2], 'c2': [criterion 2]...), 'other_details_evaluation': (same format as 'entity_evaluation' with 1-3 pairs of question and criterion), 'quality_evaluation': (same format as 'entity_evaluation' with 1-3 pairs of question and criterion)}}`.

Table 6: **Template used by DeepSeek-R1 to generate evaluation question-criterion pairs for the dimension of Entity-Reasoning.** The text highlighted in cyan should be replaced with details from the specific prompt.

<USER>: I have a text-to-image generation model that can generate images based on given prompts. However, the prompts given to the model imply scientific laws (e.g., physics, chemistry, biology, or astronomy) that can affect how the scene looks without explicit explanation. Your task is to evaluate whether the generated image accurately reflects the scientific law and correctly portrays the resulting scene. Given the prompt: `{{'id': {prompt_id}, 'prompt': {prompt}, 'explicit_meaning': {explicit_meaning}}}`, you need to:

1. describe what should be depicted in the image in order to fully and accurately reflect the explicit meaning of the prompt.
2. identify any scientific law(s) that the model needs to infer from the prompt, and create a list of questions that check whether the image correctly demonstrates and complies with these scientific laws.
3. consider other elements or details in the prompt that are not directly affected by the scientific law(s), create a list of questions that check if the image accurately represents these additional key elements.
4. consider factors that could impact the aesthetics or visual quality of the image and list relevant questions.

Please also design a scoring criterion for each question, where a score of 1 means “yes (to the question)”, 0 means “no”, and 0.5 means “partially yes”.

Provide your answer in json format: `{{'id': [prompt id], 'prompt': [the prompt], 'explicit_meaning': [the explicit meaning], 'image_content': [what the image should depict], 'scientific_evaluation': (here should be a dictionary with 2-4 pairs of question and criterion: 'q1': [question 1], 'c1': [criterion 1], 'q2': [question 2], 'c2': [criterion 2]...), 'other_details_evaluation': (same format as 'scientific_evaluation' with 1-3 pairs of question and criterion), 'quality_evaluation': (same format as 'scientific_evaluation' with 1-3 pairs of question and criterion)}}`.

Table 7: **Template used by DeepSeek-R1 to generate evaluation question-criterion pairs for the dimension of Scientific-Reasoning.** The text highlighted in cyan should be replaced with details from the specific prompt.

<USER>: Describe this image.

<ASSISTANT>: The image depicts a group of four people in a meeting setting, with one person standing and the others seated around a table. The standing individual, dressed in a dark suit and light blue shirt, is playfully pointing at the nose of one of the seated individuals, who is smiling broadly...

<USER>: Based on the image and your previous description, answer the following questions: q1, q2, ...

For each question, assign a score of 1, 0.5 or 0 according to the corresponding scoring criteria: c1, c2, ...

Here are the questions and criteria: {accuracy_evaluation_qc}.

Carefully consider the image and each question before responding, then provide your answer in json format: {{ "reason": [your detailed reasoning], "score": [s1,s2, ...] }}

<ASSISTANT>: { "reason": ["q1: The image clearly shows a meeting environment with a conference table, office-like setting, and individuals in professional attire. This aligns well with the criteria for a clear meeting setting.",

"q2: One person is visibly acting as the joke-teller by making a playful gesture towards another person's face, which suggests an active role in engaging the group. This meets the criteria for a clear speaker with an active role.",

"q3: The audience members are displaying relaxed and post-icebreaking reactions, such as genuine smiles and laughter, indicating that the tension has been relieved. This fulfills the criteria for clear reactions showing tension relief.",

"q4: The idiom 'break the ice' is depicted entirely metaphorically through the playful interaction and the text overlay, without any literal elements like actual ice. This satisfies the criteria for an entirely metaphorical depiction."], "score": [1, 1, 1, 1] }

<USER>: Based on the image and your previous description, answer the following questions: q1, q2, ...

For each question, assign a score of 1, 0.5 or 0 according to the corresponding scoring criteria: c1, c2, ...

Here are the questions and criteria: {quality_evaluation_qc}

Carefully consider the image and each question before responding, then provide your answer in json format: {{ "reason": [your detailed reasoning], "score": [s1,s2, ...] }}

<ASSISTANT>: { "reason": ["The image is visually clear with well-defined subjects. There is no noticeable blur or distortion, allowing for easy identification of the individuals and their actions. Therefore, it scores high on clarity.",

"The composition is balanced, with the speaker positioned centrally and the audience arranged cohesively around the table. There is no awkward cropping, and all elements contribute to a harmonious arrangement.",

"The colors and lighting are appropriate, enhancing the mood of the scene. Warm tones and even lighting create a positive and inviting atmosphere, supporting the lighthearted interaction depicted."], "score": [1, 1, 1] }

Table 8: **Template used by Qwen2.5-VL to evaluate generated images in T2I-ReasonBench.** The text highlighted in cyan should be replaced with the specific evaluation question and criterion pairs for the given prompt. An example output for Idiom Interpretation is also provided.