

# PromptEnhancer: Taming Your Rewriter for Text-to-Image Generation via Fine-Grained Reward

## Supplementary Material

### A. Data Curation

A high-quality dataset is critical for prompt rewriter training. We design a multi-stage data curation pipeline for SFT (rewriter initialization) and policy alignment (Figure 4). The SFT dataset is built via four stages (simulating user prompts, generating CoT/re-prompt candidates, human-in-the-loop selection, automated filtering) and contains **485,119** (user prompt, CoT, re-prompt) triplets. The pipeline shown in sec 3.4. Based on fig. 8, the dataset exhibits a thematic distribution across several creative domains. Design constitutes the largest portion at 27%, followed by Art at 23%, and Film & Story at 22%. Illustration accounts for 18%, and Creative tasks make up 10% of the dataset.

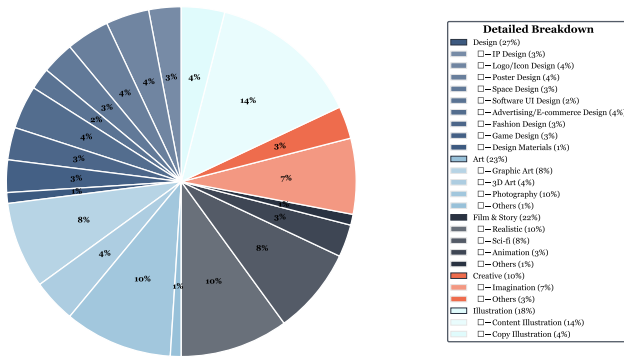


Figure 8. **Distribution of Categories in the Dataset.** The chart provides a detailed breakdown into 20 sub-categories..

As shown in Fig. 9, the re-prompt expands a simple input sentence into a structured, detailed description suitable for image generation, following a sophisticated framework. This framework comprises three core components: **I. Sentence Structures** (defining a four-level descriptive hierarchy from macro to micro), **II. Grammatical Rules** (specifying seven conventions to ensure objectivity and precision), and **III. 9 Key Constraints**, which collectively guarantee the quality and consistency of the generated content. Through this rewriting process, we obtain one-to-one corresponding (user prompt, target re-prompt) pairs.

Additionally, as shown in Fig. 10, we can also leverage Gemini to construct the CoT process: input the user prompt and the final re-prompt, and instruct the model to elaborate on the rewriting rationale. The System Prompt outlines the core requirements: the primary task is not to output a final answer, but to generate a reasoning process that explains how the system derives the optimized new re-prompt ([1]) from the user's initial input ([0]). It requires the model to focus on a series of analytical dimensions (e.g., core elements, composition, attributes, style, world knowledge) and specifies the required format and depth via a detailed "Example Output".

**System Prompt for Reprompt Generation**

You are an expert in writing prompts for image generation. I will give you a sentence, and you are to expand this sentence into a detailed caption for generating an image. And the captions must follow the rules listed below.

### \*\*I. Sentence Structures\*\*

The captions follow a consistent, hierarchical structure that moves from a general overview to specific details.

1. \*\*The Opening Statement: General Overview\*\*
2. \*\*The Body: Systematic and Spatially Organized Description\*\*
3. \*\*Hierarchical Object Description: From Whole to Parts\*\*
4. \*\*The Concluding Statement: Stylistic Identification\*\*

---

### \*\*II. Grammatical Rules\*\*

The grammar is precise, descriptive, and maintains an objective tone.

1. \*\*Tense: Consistent Present Tense\*\*
2. \*\*Voice: Mix of Active and Passive\*\*
3. \*\*Prepositional Phrases for Precision\*\*
4. \*\*Participial Phrases for Efficient Detail\*\*
5. \*\*Rich and Specific Adjectives\*\*
6. \*\*Precision and Hedging Language\*\*
7. \*\*Complex and Compound Sentences\*\*

Key constraints:

1. Only provide the final captions, do not use markdown format.
2. The expanded captions must follow the rules listed above.
3. The expanded captions should adhere to the original sentence, especially the subject and the subject's attributes, including color, size, spatial relationships, etc.
4. You can use your world knowledge to expand some professional terminology to proper explanations that suitable for image generation models.
5. If the style of original sentence is not mentioned, you should assume it is a photography style. And you can infer the style from the content of the sentence if the photography style is not suitable.
6. Describe the scene or subject directly, do not use "The image", "The composition", "The scene" and similar words in the beginning of the captions.
7. Unless the original sentence specifies that is a photo, do not assume that the given sentence is a photo, just describe the scene or subject directly.
8. If the original sentence has a IP subject, you should keep the IP subject in the expanded captions, and describe the background of the IP in the expanded captions.
9. If the original sentence has a text that need to be rendered, you should keep the text in the expanded captions, and format text as "rendered text".

Next, I will give you my sentence. Please provide the expanded captions:

Figure 9. The system prompt designed to guide Gemini-2.5-Pro for "Reprompt Generation".

**System Prompt for CoT Generation**

You have the following information:

1. The user's input prompt for text-to-image generation: [0].
2. Based on the user's input prompt, refer to the new prompt: [1].

Your task is not to output the final answer or image. Instead, output the thought process or reasoning chain explaining how you derive the new prompt from the user's input prompt. You must:

- Generate a "thinking" or reasoning chain process to explain how you arrive at the new prompt based on the user's input prompt.
- The new prompt guides the entire thought direction, but no information from the new prompt should be leaked in the thought process.
- Avoid introducing any content/elements/props/text/watermarks unrelated to the "input prompt" (e.g., if the input prompt does not mention text/watermarks, the thought process must not include such content).
- Do not provide excessive explanations; the output length must be less than \*\*384 tokens\*\*.

Below is an example output. Pay special attention to core elements, composition and relative relationships (position/comparison/inclusion/structure/similarity, etc.), attributes (size/quantity/material/expression, etc.), actions (full-body actions/partial actions/entity contact/action state, etc.), grammar (negations/pronouns, etc.), style (sketch/watercolor/game/realistic, etc.), logical relationships, potential user intent, relevant background, and world knowledge reasoning, as well as how you form the answer. The output length must be less than \*\*384 tokens\*\*.

## Example Output

The user wants to generate an image with the following core elements: Person: young woman; Clothing: brown hoodie; Accessories: ski goggles; Props: red snowboard. The action is left hand on hip, style is realistic, background is the capital of China and the national flower of China. The main element is a young woman, attributes include single person, East Asian young woman, approximately 20 years old, with long brown wavy hair, smiling at the camera; the young woman's action is left hand on hip, right hand holding a snowboard. Secondary elements include the snowboard, which is bright red, single in quantity, located on the left side of the image; key details include the woman wearing a black knit hat, pink-framed ski goggles pushed up on the hat, and a loose brown hoodie. Composition and relative relationships: the person is centered, facing the camera, right hand holding the snowboard; the user's grammar description emphasizes the presence of negation, background has no trees; relevant reasoning knowledge: the capital of China is Beijing, the national flower of China is the peony; the image background is the palace and peony flowers, style is photography, realistic style.

Figure 10. The system prompt for generating a Chain-of-Thought.

Through the aforementioned processes, we construct (user prompt, CoT, reprompt) triplets. After completing the first-stage SFT training with this dataset, the model can output the corresponding CoT and reprompt given a user prompt. Via multi-round rollout, multiple candidate outputs are generated; after scoring by the evaluator, GRPO learning is performed to ultimately ensure the Rewriter achieves high-quality rewriting performance.

## B. Evaluation Benchmark Analysis

To address the need for nuanced prompt following evaluation, we develop AlignEvaluator with 24 fine-grained key points (grouped into 6 categories). Trained on large-scale reprompt-image pairs, it scores via metrics like text-image consistency, delivering robust and explicit reward signals for the CoT Rewriter. Details are provided in Section 3.2: Table 1 outlines the key dimensions, while Table 6 offers detailed breakdowns (including subcategories, key point explanations, and example prompts) to enable objective scoring and fair evaluation of rewritten prompts—critical for subsequent reward learning.

As we mentioned in Section 3.3, we propose the T2I-Keypoints-Align benchmark for comprehensive, fine-grained T2I model evaluation—decomposing complex prompts into semantic keypoints. It includes 6,687 balanced English (3,000) and Chinese (3,687) prompts, each annotated with multiple keypoint categories. We further analyze the distribution of prompts/keypoints, revealing structural differences between the two languages.

To better understand the characteristics of our benchmark, we conducted a statistical analysis of the prompts and their associated keypoints, with key findings visualized in Figure 11 and Figure 12. Co-occurrence heatmaps of the top 24 most frequent keypoints, comparing the Chinese and English portions of our benchmark. The color intensity reflects the co-occurrence frequency. The Chinese data shows strong correlations between ‘Style’ and ‘World Knowledge’, while the English data displays more varied pairings.

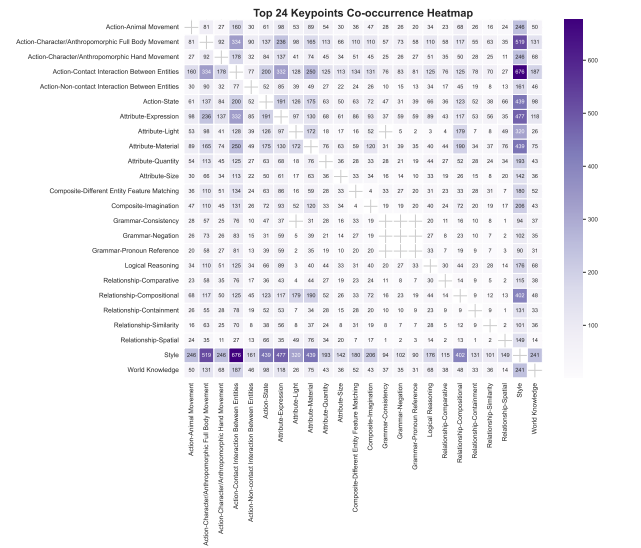


Figure 11. The heatmap of Chinese data in evaluation benchmark.

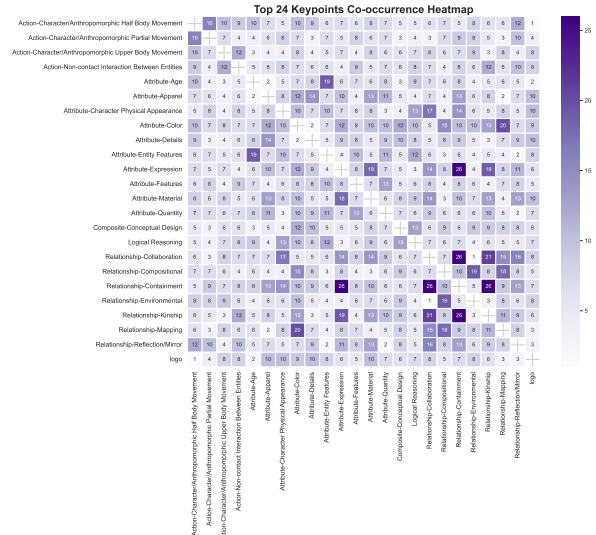


Figure 12. The heatmap of English data in evaluation benchmark.

We also statistically analyze the distributions of prompt characteristics by comparing the Chinese (Fig. 13) and English (Fig. 14) datasets, which reveals significant structural differences between the prompts of the two languages. Each figure presents the distribution of two key metrics: prompt length (counted in characters) and keypoint density (defined as the number of keypoints per prompt).

A notable discrepancy is observed in prompt length: English prompts are considerably longer and more descriptive, while Chinese prompts are concise with lengths tightly clustered around a mean of approximately 100 characters. For keypoint density, Chinese prompts peak at 4, with most containing 3 to 5 keypoints. In contrast, English prompts are markedly more verbose, with lengths centered around 500 characters. Their keypoint density exhibits a more uniform distribution, where a high and consistent number of prompts contain 3, 4, 5, or 6 keypoints.

This distinction underscores that the English subset evaluates a model’s ability to parse longer narratives with high compositional complexity, whereas the Chinese subset focuses on efficiently capturing the core of more succinct requests. Together, they constitute a robust benchmark for assessing T2I models across diverse prompt styles.

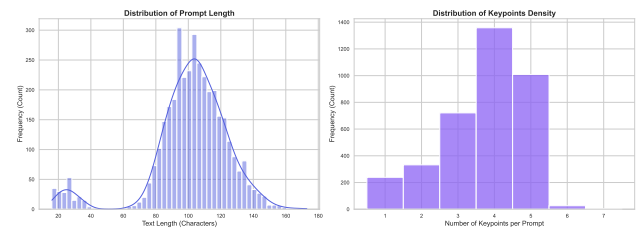


Figure 13. The distribution of prompt characteristics in Chinese.

Table 6. A Multi-dimensional Evaluation Framework for Text-to-Image Generation in detail. TIC = Text-Image Consistency, SI = Structural Integrity, TIC&SI = both.

Super-Category	Category	Key Point	Example	Criteria
Linguistic Comprehension	Logical Ops	<b>Negation</b> – interpret negatives	Prompt: A bowl of beef noodles, no scallions. (No scallions)	TIC
	Logical Ops	<b>Attribute Consistency</b> – one attribute bound to many	Prompt: Five people all wearing red clothes. (All red)	TIC
	Co-reference	<b>Pronoun Resolution</b> – resolve ambiguity	Prompt: The large ball broke the table because it was made of metal. (“it” = ball)	TIC
Visual Attributes	Obj-level	<b>Counting</b> – numeracy ( $n \geq 3$ )	Prompt: A picture with four dogs. (Four dogs)	TIC
	Obj-level	<b>Size</b> – relative comparison	Prompt: Two large spheres. (Large spheres)	TIC
	Obj-level	<b>Material</b> – render different materials	Prompt: An ice sculpture of an eagle. (Ice sculpture)	TIC
	Obj-level	<b>Expression</b> – capture facial emotions	Prompt: A strong man, low-angle shot, with a contemptuous expression.	TIC
	Global Style	<b>Artistic Style</b> – adhere to style	Prompt: Eight galloping horses in Chinese ink wash.	TIC
Action & Interaction	Individual Action	<b>Full-body Action</b> – complex movement	Prompt: A girl performing a Thomas flare.	TIC&SI
	Individual Action	<b>Hand Action</b> – detailed hand/finger structure	Prompt: A hand using chopsticks to pick up food.	TIC&SI
	Individual Action	<b>Animal Action</b> – actions performed by animals	Prompt: A puppy happily running.	TIC&SI
	Interaction	<b>Contact Interaction</b> – physical interaction	Prompt: A boxer lands a punch on a punching bag.	TIC&SI
	Interaction State	<b>Interaction w/o Contact</b> – non-physical interaction <b>State</b> – continuous state of being or action	Prompt: Einstein looking at Hawking. Prompt: A gust of wind blows, cherry blossoms dance in the air.	TIC TIC
Relations & Structure	Semantic Rel.	<b>Comparative Relation</b> – attribute comparison	Prompt: Woman in red dress taller than woman in yellow.	TIC
	Semantic Rel.	<b>Compositional Relation</b> – entity composed of others	Prompt: A cat made of orange slices.	TIC
	Semantic Rel.	<b>Containment Relation</b> – container holds an entity	Prompt: A cup full of soda water.	TIC
	Semantic Rel. Spatial Layout	<b>Similarity Relation</b> – resemblance in shape <b>Cross-Entity Binding</b> – distinct attributes to entities	Prompt: A lake shaped like a guitar. Prompt: Man (buzz cut, blue shirt) and woman (long hair, yellow shirt).	TIC TIC
	Spatial Layout	<b>Entity Layout</b> – specific arrangement of entities	Prompt: A race car on a city track, with a mini-map in the top-left corner.	TIC
World Knowledge & Reasoning	World Knowledge	<b>Knowledge Application</b> – famous entities	Prompt: The Great Wall of China / Marie Curie.	TIC
	Abstract Reasoning	<b>Counterfactual</b> – surreal impossible scenes	Prompt: A girl held onto the stem of a huge dandelion with both hands, suspended above the clouds.	TIC
Scene Text & Typography	In-Image Text	<b>Text Rendering</b> – render text content accurately	Prompt: Poster with text “Game of Thrones” at the bottom.	TIC
	In-Image Text	<b>Text Layout</b> – position text as instructed	Prompt: Poster of a woman on a throne of waves, text “Game of Thrones” at the bottom.	TIC

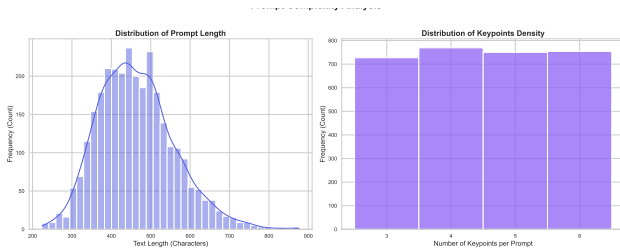


Figure 14. The distribution of prompt characteristics in English.

## C. More Visualization Results

### Visualization of prompt Enhancers results.

We provide a detailed presentation of original user prompts

not illustrated in Fig. 1 and their revised versions generated by PromptEnhancer. It is observed that PromptEnhancer performs refined revisions targeting key elements of prompts, including positional relationships, style, background, and camera language. These revisions not only align with the user’s original intent but also significantly enhance the presentation quality of the prompts.

We also present more comparison examples of revisions in Fig. 16. PromptEnhancer significantly improves the detailed rendering of images and comprehensively enhances prompt adherence in terms of style, layout, and other aspects, thereby effectively bridging the gap between user intent and the final rendering effect.

<p>A futuristic pool under a starlit sky reflects vibrant nebulae and galaxies. Neon cyan, magenta, and violet lights create a cyberpunk ambience. Above, glowing "Prompt Enhancer" text floats, casting a glow on the water. Distant planets with rings and comets add to the surreal, dreamlike atmosphere.</p>		<p>A futuristic swimming pool under a starlit sky, reflecting the deep cosmos. The pool's surface glows with vibrant nebulae, swirling galaxies, and twinkling stars. Neon lights in shades of cyan, magenta, and violet illuminate the area, creating an uncanny, ethereal atmosphere. Above the pool, glowing purple and pink text "Prompt Enhancer" floats in mid-air, surrounded by a soft halo, casting a reflective glow on the water. The background features distant planets with ethereal rings and comets, enhancing the surreal and dreamlike atmosphere. The scene combines sleek, modern architecture with an otherworldly, tech-inspired design, evoking a sense of wonder and futuristic elegance.</p>		<p>系列头像包装设计。头像：一只猴子咧嘴笑，下方写有“Happy”。头像：猴子戴墨镜，下方写有“Cool”。头像：猴子拿着一朵花，下方写有“Shy”。头像：猴子张大嘴巴，下方写有“Surprise”。</p>		<p>一套四个头像包，每个头像包中都有一只猴子的头像。下方文字为“Happy”，头像：猴子咧嘴笑；下方文字为“Cool”，头像：猴子戴墨镜；下方文字为“Shy”，头像：猴子拿着一朵花；下方文字为“Surprise”，头像：猴子张大嘴巴。头像包为白色，文字为黑色。头像包下方有四个小方格，分别写着“Happy”、“Cool”、“Shy”、“Surprise”。</p>	
<p>A vintage travel poster advertising the French Riviera, styled in bright, bold watercolor colors with a retro flair.</p>		<p>A vintage travel poster advertising the French Riviera is presented, characterized by its bright, bold watercolor palette and a distinct retro aesthetic. At the top of the poster, the title "French Riviera" is written in a stylized, cursive font. Below the title, the poster is filled with vibrant, saturated colors, primarily yellows, oranges, and reds, which evoke a sense of warmth and sun-drenched landscapes. The central focus of the poster is a woman in a classic, elegant dress, depicted in a soft, painterly style. She is surrounded by various elements of the Riviera, including palm trees, a beach, and a boat. The overall composition is balanced and visually appealing, capturing the essence of the French Riviera's allure and charm. The poster's design is reminiscent of mid-20th-century travel posters, with its emphasis on vibrant colors and a nostalgic feel.</p>		<p>客厅的墙上挂着一幅油画，画的是一只哈士奇睡在沙发上，背景是蓝色的。</p>		<p>一幅精美的油画悬挂在客厅的墙上，画面为一只哈士奇犬躺在沙发上睡觉。画面色彩柔和，背景为淡蓝色。画面下方有文字“Happy”、“Cool”、“Shy”、“Surprise”。</p>	
<p>An oil painting of a woman dressed in a simple, well-fitted jacket and scarf, the brushwork capturing the texture of the fabric. She appears content and at ease, the muted background not detracting from her subtle grace.</p>		<p>An oil painting portrays a woman in a simple, well-fitted jacket and scarf, her form captured with thick, visible brushstrokes. The woman is the central focus, depicted with a serene expression and a gentle smile, exuding a sense of quiet confidence. She is dressed in a simple, well-fitted jacket, with the paint applied in bold, expressive strokes that capture the texture of the fabric. A scarf is draped around her neck, rendered with delicate, layered brushwork. The background is a muted, textured wash of colors, providing a subtle contrast to the woman's figure. The overall composition is balanced and harmonious, with a focus on the woman's subtle grace and the tactile quality of the oil paint.</p>		<p>中国水墨画风格，五个男人穿着传统服饰，背着一个大包裹，走在路上。</p>		<p>一幅中国水墨画，描绘了五个男人正在赶路。他们穿着传统的中国服饰，背着大包小包，走在一条蜿蜒的山路上。画面充满了动感，线条流畅，墨色浓淡相宜，展现了中国山水画的独特魅力。</p>	
<p>Vintage national park poster for Yellowstone. Painted style depicting Old Faithful geyser erupting. Rich, warm colors, "Yellowstone" in rustic, wood-carved font effect. America's First National Park subtitle. Silhouette of bison in corner.</p>		<p>A vintage poster for Yellowstone National Park is presented, featuring a dramatic depiction of the Old Faithful geyser erupting. The central focus is the geyser, shown in a powerful, dynamic pose as it sends a plume of white steam high into the air. The scene is rendered in a classic, painterly style, with rich, warm colors that evoke a sense of natural wonder and grandeur. The background is a mix of earthy tones, from deep reds and oranges to soft yellows and whites, creating a harmonious and visually appealing composition. The text "Yellowstone" is written in a rustic, wood-carved font, adding to the vintage feel of the poster. Below the main title, the subtitle "America's First National Park" is displayed in a smaller, elegant font. In the bottom right corner, a silhouette of a bison is visible, further emphasizing the park's natural heritage. The overall design is both nostalgic and eye-catching, capturing the essence of Yellowstone's beauty and history.</p>		<p>混合媒介拼贴设计，画面主体是一个城市街景，背景是蓝色的。画面下方有文字“Happy”、“Cool”、“Shy”、“Surprise”。</p>		<p>一幅混合媒介拼贴设计，画面主体是一个城市街景，背景为蓝色。画面下方有文字“Happy”、“Cool”、“Shy”、“Surprise”。</p>	

Figure 15. Detailed prompts for Figure 1, including original user prompts and their revised versions generated by PromptEnhancer, with both Chinese and English iterations.









User Prompt	T2I	PromptEnhancer	PromptEnhancer + T2I
<p>Design a billboard for a new electric sports car, featuring a sleek, metallic blue car zooming through a neon-lit cityscape with the slogan "Drive the Future."</p>		<p>A large, brightly illuminated billboard is presented, advertising a new electric sports car against the backdrop of a neon-lit cityscape. The central focus is the billboard itself, which displays a dynamic scene. On the billboard, a sleek, electric sports car, finished in a metallic blue paint that reflects the surrounding lights, is depicted zooming from left to right. The car has an aerodynamic, low-profile design, and its movement is emphasized by motion blur and bright light trails trailing behind it. In the background of the billboard, a futuristic cityscape is rendered with towering skyscrapers whose windows glow with neon blues and purples, and vibrant neon signs with abstract symbols float in the air. The entire cityscape has a blurred motion effect, enhancing the sense of speed. Arching above the car in a bold, clean, glowing white font are the words "Drive the Future". The billboard is positioned at a slight angle, making it a prominent feature. The overall presentation has the style of a professional advertisement, combining elements of futuristic concept art and dynamic action photography.</p>	
<p>Instagram Reels cover for PawPerfect: Puppy in shopping cart filled with toys. '20% OFF FIRST PURCHASE' on price tag hanging from cart. 'Swipe Up for Puppy Essentials' at bottom.</p>		<p>An Instagram Reels cover is presented, featuring a heartwarming scene centered on a puppy in a shopping cart against a softly blurred, bright background. At the top of the frame, the brand name "PawPerfect" is displayed in a playful, purple cursive font. The central focus is a small puppy with golden-brown fur and floppy ears, sitting upright inside a red plastic shopping cart; the cart is overflowing with an assortment of dog toys, including colorful plush bones and squeaky chickens. Hanging from the handle of the cart is a white price tag with a brown paper texture, connected by a string, which prominently displays the text "20% OFF FIRST PURCHASE" in a bold, black, sans-serif font. At the bottom of the cover, a banner contains the phrase "Swipe Up for Puppy Essentials" in a clear, blue, uppercase font. The overall presentation is in the style of bright and cheerful photography, designed for social media marketing.</p>	
<p>A group of employees volunteering at a local community center, highlighting corporate social responsibility and the human side of business.</p>		<p>A group of diverse employees is captured volunteering inside a brightly lit local community center, engaged in various activities. In the foreground, a woman with brown hair, wearing a blue company-branded t-shirt under an open vest, kneels down to help a young child with a painting project at a wooden table. To her right, in the middle ground, a man with glasses is organizing stacks of books and supplies on another table, while a third employee laughs while making a paper craft. The background is filled with the bustling atmosphere of the center, featuring poster-covered walls, other children playing, and large windows letting in natural light. All volunteers are dressed in a mix of casual jeans and company apparel, displaying friendly and engaged expressions. This image presents a photography style.</p>	
<p>An oil painting of a sailor adjusting ropes on a vintage ship under a cloudy sky.</p>		<p>An oil painting captures a sailor on the deck of a vintage ship, busily adjusting ropes under a dramatic, cloudy sky. The central figure is the sailor, depicted with thick, expressive brushstrokes, wearing a classic sailor's cap and a heavy coat as he strains to pull a thick rope. The ropes are rendered with bold, impasto brushstrokes, wrapping around the ship's rigging. He stands on a deck made of weathered wood planks, with the mast and other sailing apparatus visible in the background. The sky above is a turbulent composition of thick, layered clouds, with visible brushstrokes in shades of blue, gray, and yellow. The overall artwork is rendered in the style of an oil painting, characterized by thick paint application, visible brushstrokes, and a textured canvas.</p>	

Figure 16. More visualization results of PromptEnhancer.