PSG-Adapter: Controllable Planning Scene Graph for Improving Text-to-Image Diffusion (Supplementary Material)

Yi Gao[0009−0004−9168−1792]

Nanjing University of Science and Technology, Nanjing, CN
gaoyi@njust.edu.cn

1 Prompt Templates

Prompts to Recaption and Split the Original Text Prompt.

1. **Task Instruction**
   You are a master of composition, adept at identifying key objects and their attributes from input text. You specialize in expanding sentences and dividing them into subprompts, guided by principles of aesthetics and photographic visual appeal. By enriching the original text with detailed imagination, you create layouts that align with human aesthetic sensibilities.
2. **In-context examples**
   Caption: The garden was beautiful.

   - Identify Key Objects and Attributes
     Key Objects: garden
     Attributes: beautiful
   - Expand the Sentence with Detailed Imagination
     The garden was filled with vibrant, colorful flowers, lush green plants, and a serene pond reflecting the clear blue sky.

   - Divide into Subprompts for Aesthetic Appeal
     Subprompt 1: Lush green plants with leaves swaying gently in the breeze.
     Subprompt 2: Vibrant, colorful flowers in full bloom.
     Subprompt 3: A serene pond reflecting the clear blue sky.
     Subprompt 4: A decorative fountain adding a soothing sound of flowing water.
3. **Trigger CoT Reasoning**
   Caption: The beach was relaxing.
   Reasoning: Let's think step by step. First, identify the key objects and attributes in the caption. In this case, the key object is "beach" and the attribute is "relaxing." Second, expand the sentence with detailed imagination. Visualize the elements that make a beach relaxing: soft, golden sand; gentle waves lapping at the shore; the sound of seagulls in the distance; and a

In a quaint village square, a young artist with striking violet hair meticulously paints a mural on the side of an old brick building. She wears a vintage denim jacket adorned with patches of colorful abstract designs and faded jeans splattered with paint. Nearby, elderly locals sit on benches, engaged in lively conversations under the shade of towering oak trees. Above, sparrows flit about, their chirps blending with the distant sounds of a babbling brook. In the distance, a historic stone church with ivy-covered walls adds a timeless charm to the serene village square.

In a bustling city market, a skilled artisan with a warm smile meticulously carves intricate designs into wooden figurines. He wears a traditional apron adorned with colorful patterns and a worn leather tool belt. Nearby, a group of tourists admires his craftsmanship, drawn to the vibrant display of carved creations. Above, colorful banners flutter in the breeze, casting playful shadows on the cobblestone paths. In the distance, a row of quaint cafes and bistros offers a variety of tantalizing aromas, enticing passersby to pause and indulge.

In a cozy coastal town, a young fisherman with sandy blond hair patiently untangles a net on a weathered dock. He wears a faded blue fishing sweater adorned with intricate nautical motifs and sturdy rubber boots speckled with dried seaweed. Nearby, seagulls circle overhead, their calls mingling with the rhythmic crashing of ocean waves against rugged cliffs. Above, a lighthouse stands sentinel, its beacon casting a comforting glow over the rocky shoreline. In the distance, colorful fishing boats bob gently in the harbor, their hulls painted with whimsical maritime designs.

Fig. 1: More qualitative results on Attribute Binding. Colored text denotes critical part.

> *sunset casting a warm glow over the water. Next, divide the expanded sentence into subprompts that highlight the aesthetic appeal. Each subprompt should focus on a specific aspect of the relaxing beach scene to create a vivid and appealing composition. Finally, organize these subprompts into a coherent layout that enhances the visual and sensory experience, ensuring each element contributes to the overall relaxing atmosphere of the beach.*

The results of each caption are data like:
Enriched prompt:
Subprompts:
Now, I will provide you with a text prompt. Please generate corresponding Enriched prompt and Subprompts based on the tutorial mentioned above.
The text prompts are:

Prompts to Parse Text Prompt into Scene Graphs. We modified the prompt templates provided in the SG-Adapter [1]:
Here I have a group of captions and please help me to parse each one. Each

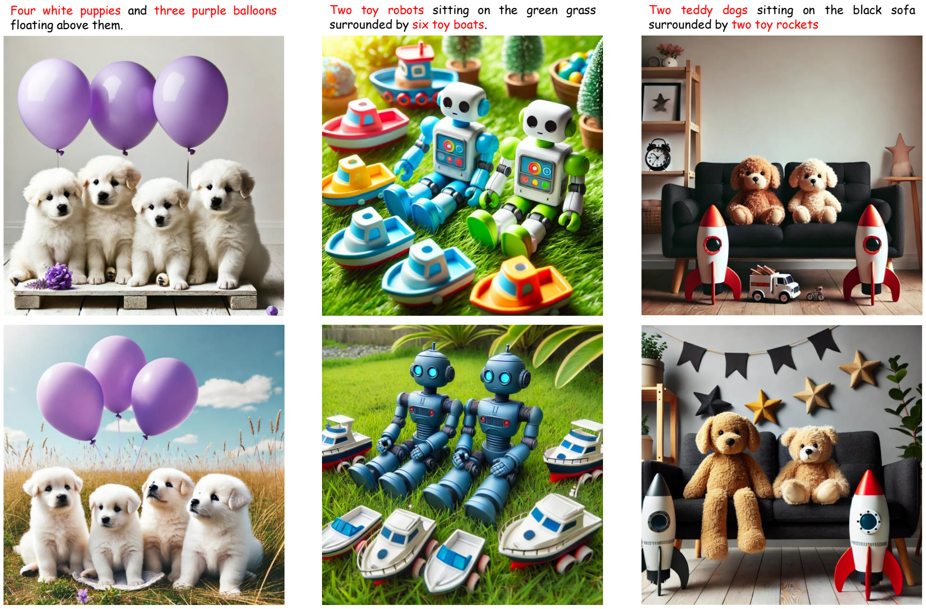

Four white puppies and three purple balloons floating above them.
Two toy robots sitting on the green grass surrounded by six toy boats.
Two teddy dogs sitting on the black sofa surrounded by two toy rockets

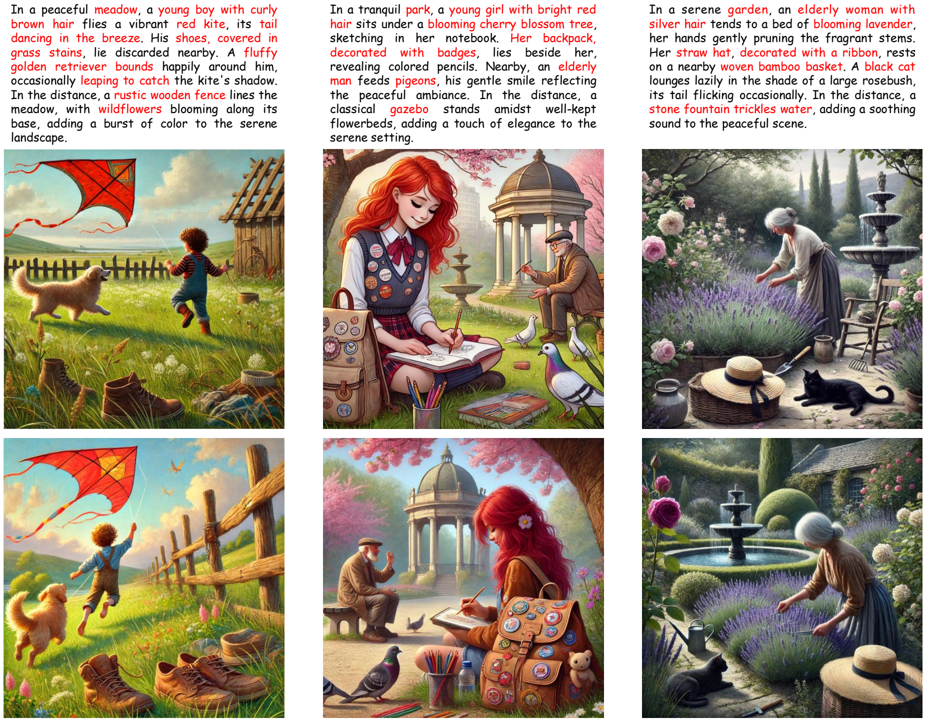
Fig. 2: More qualitative results on Numeric Accuracy. Colored text denotes critical part.

caption should be transformed to a Scene Graph reasonably which is composed of some relations. A relation is a triplet like [subject, predicate, object] and please replace the original pronouns with reasonable nouns. Both subject and object should only have one object or person. "and" relation should not be included in the Scene Graph. For example, the caption is: a boy holding a bottle shakes hands with a girl sitting on a bench.
The corresponding Scene Graph should be: [[a boy, holding, a bottle], [a boy, shakes hands with, a girl], [a girl, sitting on, a bench]]. The results of each caption are data like:
scene graph:
Now, I will provide you with a text prompt. Please parse the prompt into scene graphs based on the tutorial mentioned above.
The text prompts are:

2 Additional Results

We extended our experiments to three challenging scenarios: (i) Attribute Binding, (ii) Numeric Accuracy, and (iii) Complex Relationship.

Attribute Binding. In this scenario, as illustrated in Fig. 1, each text prompt includes multiple attributes assigned to different entities.



In a peaceful meadow, a young boy with curly brown hair flies a vibrant red kite, its tail dancing in the breeze. His shoes, covered in grass stains, lie discarded nearby. A fluffy golden retriever bounds happily around him, occasionally leaping to catch the kite's shadow. In the distance, a rustic wooden fence lines the meadow, with wildflowers blooming along its base, adding a burst of color to the serene landscape.

In a tranquil park, a young girl with bright red hair sits under a blooming cherry blossom tree, sketching in her notebook. Her backpack, decorated with badges, lies beside her, revealing colored pencils. Nearby, an elderly man feeds pigeons, his gentle smile reflecting the peaceful ambiance. In the distance, a classical gazebo stands amidst well-kept flowerbeds, adding a touch of elegance to the serene setting.

In a serene garden, an elderly woman with silver hair tends to a bed of blooming lavender, her hands gently pruning the fragrant stems. Her straw hat, decorated with a ribbon, rests on a nearby woven bamboo basket. A black cat lounges lazily in the shade of a large rosebush, its tail flicking occasionally. In the distance, a stone fountain trickles water, adding a soothing sound to the peaceful scene.

Fig. 3: More qualitative results on Complex Relationship. Colored text denotes critical part.

Numeric Accuracy. As shown in Fig. 2, each text prompt in this scenario includes multiple entities with the same class name, with the number of each entity being two or more.

Complex Relationship. As depicted in Fig. 3, each text prompt in this context features multiple entities characterized by diverse attributes and relationships, encompassing both spatial and non-spatial aspects.

References

1. Shen, G., Wang, L., Lin, J., Ge, W., Zhang, C., Tao, X., Zhang, Y., Wan, P., Wang, Z., Chen, G., et al.: Sg-adapter: Enhancing text-to-image generation with scene graph guidance. arXiv preprint arXiv:2405.15321 (2024)