Final Report: Improving Attribute Binding in Text-to-Image Generation

第九組組員: 314581006 林鼎翔、313581037 許哲瑜

Based on:

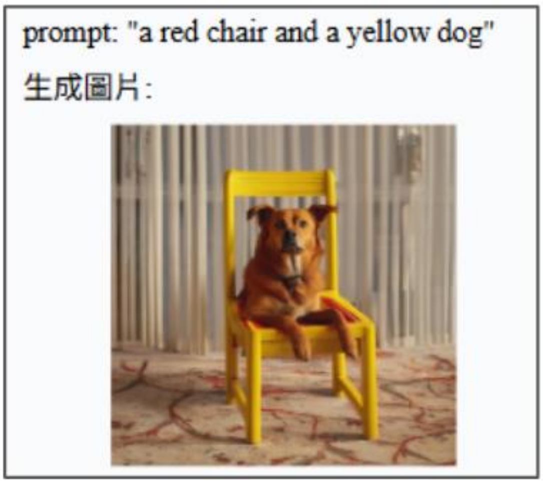
Title: Mastering Text-to-Image Diffusion: Recaptioning, Planning, and Generating with Multimodal LLMs (ICML 2024) [1]

Authors: Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, Bin Cui

Our source code is available at https://github.com/Din1225/NYCU-GAI-for-Image-Synthesis

1 Introduction

Recent advances in diffusion models—such as Imagen, DALL·E 2/3, and SDXL—have greatly improved the quality of text-to-image (T2I) generation. Despite these improvements, current models still struggle to reliably interpret complex prompts involving multiple objects, attributes, and relational structures. This limitation is particularly apparent in fine-grained compositional control, where the model must bind attributes to the correct objects or maintain multi-entity relationships within a single generated image. As shown in Figure 1, even state-of-the-art diffusion models such as SDXL often fail to produce images that satisfy all specified compositional requirements.



prompt: "a red chair and a yellow dog"

生成圖片:

Figure 1. Example of Incorrect Image Generation by SDXL

The RPG (Recaption, Plan, Generate) framework represents a strong solution to this challenge. Prior work demonstrates that by inserting a planning stage before image generation, RPG significantly improves compositional consistency, achieving state-of-the-art performance on T2I-CompBench—especially in Attribute Binding, Object Relationship, and complex compositional tasks. [2] However, the original RPG depends heavily on GPT-4 as its planner. Although the framework claims compatibility with other LLMs, substituting GPT-4 with smaller open-source models leads to substantial performance degradation. These smaller models frequently omit crucial objects, attributes, or attribute–object bindings during the regional planning stage, ultimately undermining overall generation quality.

Motivated by these limitations, our work aims to enhance the Attribute Binding capability of smaller LLM planners. Specifically, we replace GPT-4 with the open-source Llama 2-13B model and introduce a Rule-Based Self-Correction mechanism designed to compensate for the weaknesses of smaller models. By automatically detecting missing objects, adjectives, and mismatched attribute–object pairs, the system can revise and rewrite regional prompts before image generation. Our objective is to narrow the performance gap between small open-source models and GPT-4, enabling a more accessible and scalable RPG framework.

2 Review of previous work

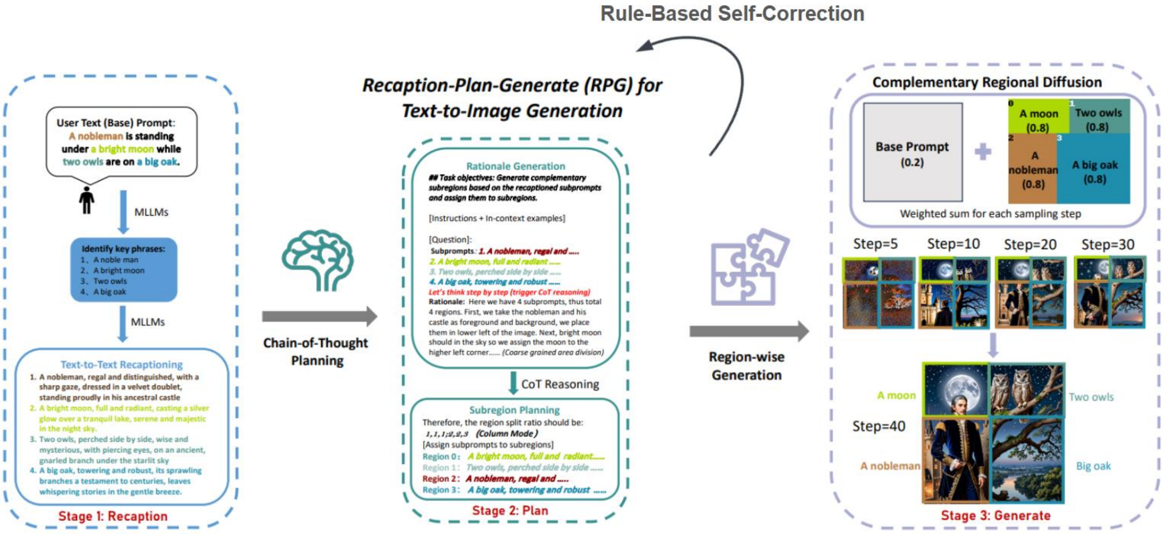
The advancement of text-to-image (T2I) generation has been significantly driven by large-scale diffusion models, including state-of-the-art architectures such as Imagen, DALL·E 2/3, and SDXL. These models have demonstrated remarkable abilities in generating high-fidelity images from natural language prompts. However, a persistent challenge remains in maintaining fine-grained compositional control, particularly when prompts involve multiple objects, complex relationships, or the critical task of Attribute Binding (i.e., correctly associating a specific attribute, like color or texture, with its corresponding object). As evidenced by the results in T2I-CompBench, even leading models like SDXL frequently fail to satisfy all specified compositional requirements.

To address these limitations, the RPG (Recaption, Plan, Generate) framework was introduced, offering a robust solution by integrating a planning stage before image synthesis. This framework leverages the strong reasoning capabilities of Large Language Models (LLMs) to decompose a complex user prompt into a structured set of simpler, region-specific prompts. Prior work has demonstrated that this planning-based approach substantially improves compositional consistency, achieving state-of-the-art performance on benchmarks like T2I-CompBench, especially across Attribute Binding, Object Relationship, and complex compositional tasks.

Despite its effectiveness, the original RPG framework is highly dependent on powerful, closed-source LLMs, specifically GPT-4, as its core planner. This reliance presents a major barrier to accessibility and scalability. Crucially, substituting GPT-4 with smaller, open-source models—such as Llama 2-13B—results in a significant performance degradation. These smaller models frequently generate incomplete or inaccurate regional plans, leading to the omission of vital objects or attributes and, critically, the failure to preserve correct attribute–object bindings. This performance gap between proprietary and open-source LLM planners motivates the current work to enhance the Attribute Binding capability of smaller LLMs through an effective self-correction mechanism.

3 Summary of the technical solution

To address the difficulty smaller LLMs face in producing accurate regional plans, we integrate a Rule-Based Self-Correction module into the RPG pipeline. In the original framework, the planner directly generates a set of regional prompts from the user's input, which are then sent to the diffusion model for image synthesis. As illustrated in Figure 2, our approach enhances this process by jointly examining both the user prompt and the planner-generated prompts to detect missing entities, omitted attributes, or incorrect attribute–object bindings.



User Text (Base) Prompt: A nobleman is standing under a bright moon while two owls are on a big oak.

Identify key phrases:
1、A noble man
2、A bright moon
3、Two owls
4、A big oak

Text-to-Text Recaptioning
1. A nobleman, regal and distinguished, with a sharp gaze, dressed in a velvet doublet, standing proudly in his ancestral castle
2. A bright moon, full and radiant, casting a silver glow over a tranquil lake, serene and majestic in the night sky.
3. Two owls, perched side by side, wise and mysterious, with piercing eyes, on an ancient, gnarled branch under the starlit sky
4. A big oak, towering and robust, its sprawling branches a testament to centuries, leaves whispering stories in the gentle breeze.

Rationale Generation
Task objectives: Generate complementary subregions based on the recaptioned subprompts and assign them to subregions.
[Instructions + In-context examples]
[Question]:
Subprompts: 1. A nobleman, regal and .....
2. A bright moon, full and radiant ......
3. Two owls, perched side by side ......
4. A big oak, towering and robust ......
Let's think step by step (trigger CoT reasoning)
Rationale: Here we have 4 subprompts, thus total 4 regions. First, we take the nobleman and his castle as foreground and background, we place them in lower left of the image. Next, bright moon should in the sky so we assign the moon to the higher left corner...... (Coarse grained area division)

Subregion Planning
Therefore, the region split ratio should be: 1,1,1;2,2,3 (Column Mode)
[Assign subprompts to subregions]
Region 0: A bright moon, full and radiant......
Region 1: Two owls, perched side by side ......
Region 2: A nobleman, regal and .....
Region 3: A big oak, towering and robust ......

Figure 2. Overview of Our Framework

The system first performs keyword verification by extracting all nouns and adjectives from the user prompt using spaCy and comparing them with those found in each BREAK-segmented regional prompt. Any term present in the user prompt but absent from all regional prompts is treated as missing. It then checks attribute–object consistency by identifying noun phrases such as "a red chair," forming their corresponding adjective–noun pairs, and verifying that these pairs appear together—either adjacent or in close syntactic proximity—within each regional prompt segment. Whenever the regional prompts fail either completeness or pairing consistency, the system rewrites them by reintroducing the missing elements and instructing the LLM to regenerate the plan. This refinement may repeat up to three times, and only regional prompts that satisfy both conditions proceed to the final image generation stage.

4 Experiments

4.1 Experimental Setup

We evaluate our approach on the Attribute Binding subset of T2I-CompBench, a benchmark specifically designed to assess whether a model can correctly associate attributes—such as color, shape, and texture—with their corresponding objects. This subset comprises 860 prompts for color, 1000 for shape, and 1000 for texture. Our experiments compare three system configurations: (1) the SDXL model used directly as a baseline, (2) the RPG framework in which GPT-4 is replaced by Llama 2-13B as the planner, and (3) our proposed system, which augments this modified RPG pipeline with the Rule-Based Self-Correction mechanism. To ensure a fair comparison, we first reproduced the original RPG results under the new planner setting before conducting further evaluations.

4.2 Evaluation Metrics

We adopt the Disentangled BLIP-VQA Score, which decomposes each text prompt into independent attribute–object queries to assess attribute binding fidelity. This metric mitigates the well-known limitations of captioning-based evaluations, which frequently misassign attributes in multi-object scenes—for example, confusing *“blue curtain and yellow chair”* with *“yellow curtain and blue chair.”*

As illustrated in Figure 3, each prompt is converted into a set of attribute–object questions (e.g., “A blue backpack?”, “A red bench?”), and BLIP-VQA is used to compute the probability that each queried binding is correct. The final score is obtained by taking the product of the individual VQA probabilities, thereby imposing strict penalties for incorrect or swapped attribute–object bindings.

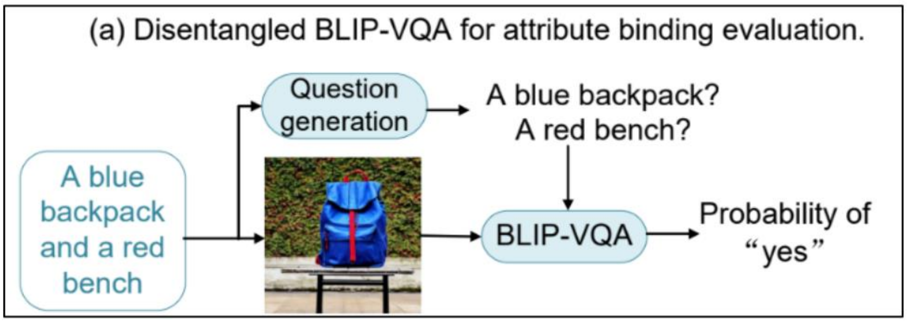

(a) Disentangled BLIP-VQA for attribute binding evaluation.
A blue backpack and a red bench
Question generation
A blue backpack?
A red bench?
BLIP-VQA
Probability of “yes”

Figure 3. Illustration of the Disentangled BLIP-VQA Score

4.3 Quantitative results

Table 1 presents the Disentangled BLIP-VQA Scores across the three Attribute Binding categories. Replacing GPT-4 with Llama 2-13B results in a substantial performance decline, with scores falling below the SDXL baseline. This confirms that smaller LLMs struggle to perform reliable regional planning.

Table 1. Quantitative Results

Model	Attribute Binding		
	Color ↑	Shape↑	Texture↑
SDXL (baseline)	**0.6314**	**0.5832**	**0.6361**
Llama 2-13B + SDXL (RPG)	0.5804	0.525	0.5888
Llama 2-13B + rule-based self-correction + SDXL (Ours)	0.6081	0.5489	0.5973

Introducing our Rule-Based Self-Correction mechanism improves performance across all categories. Although the gains are modest and still below the SDXL-only baseline, they demonstrate that rule-based refinement can partially compensate for the weaknesses of smaller planners..

More specifically, replacing GPT-4 with Llama 2-13B markedly reduces the effectiveness of the RPG framework, as every metric falls below the SDXL baseline. These results clearly indicate that smaller LLMs fail to preserve essential attribute–object bindings during regional planning. Incorporating the Rule-Based Self-Correction mechanism consistently enhances the performance of the Llama 2-13B planner, yielding improvements of 2.77 points in the Color category, 2.39 points in Shape, and 0.85 points in Texture. While these improvements are insufficient to surpass the SDXL baseline—and therefore do not enable the smaller model to replicate GPT-4's planning capabilities—the upward trend confirms that rule-based refinement effectively alleviates issues such as missing attributes and incorrect attribute–object pairings.

4.4 Qualitative Results

We further analyze representative prompts from the Color, Shape, and Texture categories. These examples illustrate how our method addresses specific weaknesses of smaller planners and leads to improved compositional consistency. Color attributes are often omitted by smaller planners; shape attributes require stricter geometric consistency; and texture attributes depend on accurate adjective–noun alignment. Our rule-based correction mechanism targets all three by restoring missing color bindings, reinforcing geometric constraints, and preserving texture-related attribute–object relations.

Figure 4 presents qualitative comparisons across the three attribute categories. In the Color example, we selected a failure case from SDXL, where the model was unable to generate a black sink. The RPG variant similarly failed to satisfy the color requirement. In contrast, our system successfully produced an image featuring a sink whose overall appearance adheres to the specified black color attribute, demonstrating improved attribute realization. In the Shape example, the prompt involved generating a cube and a cylindrical toothpaste tube. SDXL incorrectly produced an additional toothpaste object, while RPG generated the correct quantity and overall shapes. Our method further strengthened shape adherence by producing a more distinctly cylindrical toothpaste tube. Although the resulting tube appears somewhat irregular in form, it still conforms to the intended geometric specification. For Texture, the prompt required generating a leather belt and a glass vase. RPG failed to produce a belt, whereas both SDXL and our system successfully generated both objects. Notably, the image produced by our method exhibits higher visual fidelity, with the belt and vase rendered in a manner that more clearly reflects their respective texture attributes.

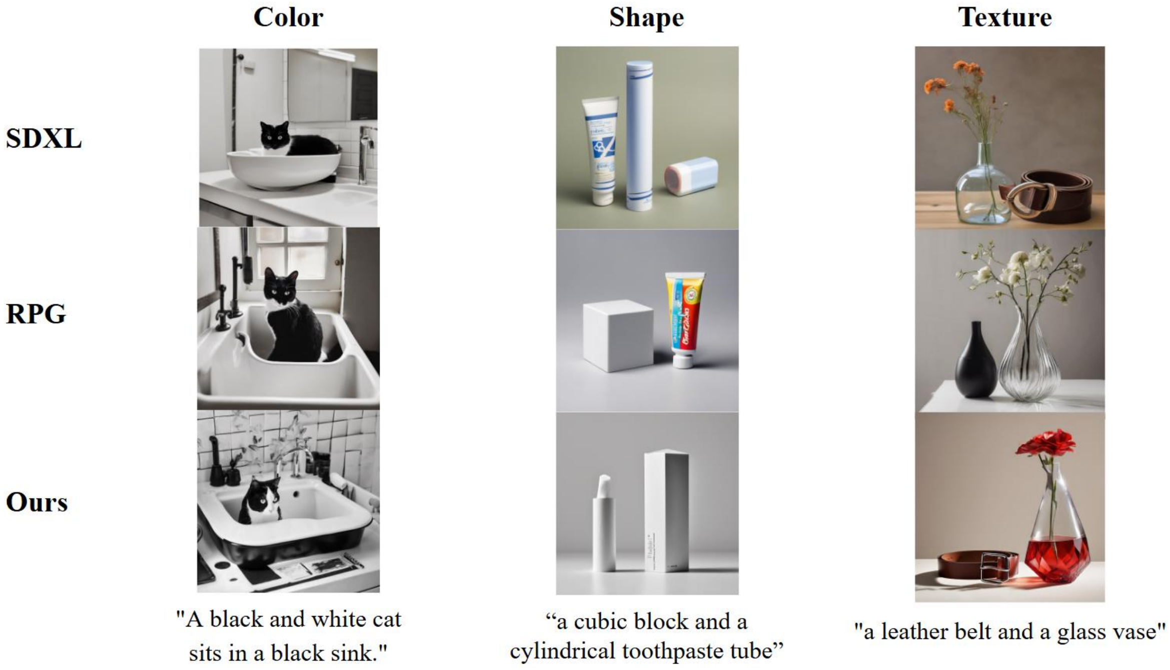

Color
Shape
Texture
SDXL
RPG
Ours
"A black and white cat sits in a black sink."
“a cubic block and a cylindrical toothpaste tube”
"a leather belt and a glass vase"

Figure 4. Qualitative Results

Overall, these qualitative observations reinforce the quantitative trends: although smaller planners struggle to maintain attribute–object integrity, the incorporation of our Rule-Based Self-Correction mechanism consistently improves adherence to color, geometric form, and texture fidelity, even in cases where the base RPG pipeline fails to produce valid compositional structures.

5 Conclusions

Our objective was to improve the performance of the Llama-2-13B Planner and bring it closer to that of the GPT-4 Planner. Although this goal was not fully achieved, the proposed self-refinement mechanism nonetheless led to performance gains compared with the RPG framework without self-refinement. This indicates that the mechanism itself is effective even if it does not surpass the baseline.

References

[1] Yang, L., Yu, Z., Meng, C., Xu, M., Ermon, S., & Cui, B. (2024). Mastering Text-to-Image Diffusion: Recaptioning, Planning, and Generating with Multimodal LLMs. *ArXiv, abs/2401.11708*.

[2] Huang, K., Duan, C., Sun, K., Xie, E., Li, Z., & Liu, X. (2023). T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 47*, 3563-3579.