Mastering Text-to-Image Diffusion: Recaptioning, Planning, and Generating with Multimodal LLMs

Ling Yang [*1] **Zhaochen Yu** [*1] **Chenlin Meng** [2 3] **Minkai Xu** [2] **Stefano Ermon** [2] **Bin Cui** [1]

Abstract

Diffusion models have exhibit exceptional performance in text-to-image generation and editing. However, existing methods often face challenges when handling complex text prompts that involve multiple objects with multiple attributes and relationships. In this paper, we propose a brand new *training-free* text-to-image generation/editing framework, namely ***Recaption, Plan and Generate (RPG)***, harnessing the powerful chain-of-thought reasoning ability of multimodal LLMs to enhance the compositionality of text-to-image diffusion models. Our approach employs the MLLM as a global planner to decompose the process of generating complex images into multiple simpler generation tasks within subregions. We propose *complementary regional diffusion* to enable region-wise compositional generation. Furthermore, we integrate text-guided image generation and editing within the proposed RPG in a closed-loop fashion, thereby enhancing generalization ability. Extensive experiments demonstrate our RPG outperforms state-of-the-art text-to-image models, including DALL-E 3 and SDXL, particularly in multi-category object composition and text-image semantic alignment. Notably, our RPG framework exhibits wide compatibility with various MLLM architectures and diffusion backbones. Our code is available at https://github.com/YangLing0818/RPG-DiffusionMaster.

1. Introduction

Recent advancements in diffusion models (Sohl-Dickstein et al., 2015; Dhariwal & Nichol, 2021; Song et al., 2020; Yang et al., 2023c) have significantly improve the synthesis results of text-to-image models, such as Imagen (Saharia et al., 2022), DALL-E 2/3 (Ramesh et al., 2022; Betker et al., 2023) and SDXL (Podell et al., 2023). However, despite their remarkable capabilities in synthesizing realistic images consistent with text prompts, most diffusion models usually struggle to accurately follow some complex prompts (Feng et al., 2022; Lian et al., 2023; Liu et al., 2022; Bar-Tal et al., 2023), which require the model to compose objects with different attributes and relationships into a single image.

Some works begin to solve this problem by introducing additional layouts/boxes (Li et al., 2023b; Xie et al., 2023; Yang et al., 2023e; Qu et al., 2023; Chen et al., 2024; Wu et al., 2023b; Lian et al., 2023) as conditions or leveraging prompt-aware attention guidance (Feng et al., 2022; Chefer et al., 2023; Wang et al., 2023) to improve compositional text-to-image synthesis. For example, RealCompo (Zhang et al., 2024) balances the realism and compositionality of diffusion models by reweighting the textual and layout guidance in denosing process. GLIGEN (Li et al., 2023b) designs trainable gated self-attention layers to incorporate spatial inputs, such as bounding boxes, while freezing the weights of original diffusion model.

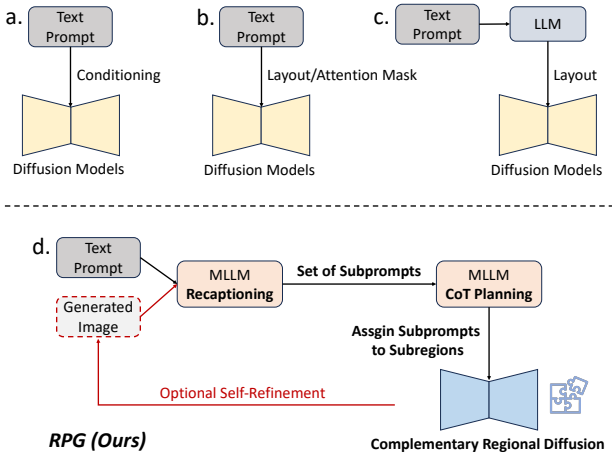


Figure 1. Illustrative architecture comparison between (a) text-conditional diffusion models (Ramesh et al., 2022; Betker et al., 2023), (b) layout/attention-based diffusion models (Feng et al., 2022; Cao et al., 2023), (c) LLM-grounded diffusion models (Lian et al., 2023) and (d) our RPG framework.

[*]Equal contribution [1]Peking University, China [2]Stanford University, USA [3]Pika Labs, USA. Correspondence to: Ling Yang <yangling0818@163.com>, Bin Cui <bin.cui@pku.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Prompt: A beautiful landscape with a river in the middle, the left of the river is **in the evening** and **in the winter** with a big iceberg and a small village while some people are skiing on the river and some people are skating, the right of the river is **in the summer** with a **volcano in the morning** and a small village while some people are playing.

Prompt: A **green twintail** girl in **orange dress** is sitting on the sofa while **a messy desk** in under a big window on the left, while a **lively aquarium is on the top right of the sofa**, realistic style.

Left Prompt: A Chinese general wearing a crown, with whiskers and **golden Chinese** style armor, standing with a **majestic dragon head on his chest**, symbolizing his strength, wearing **black and gold boots**. His appearance exudes a sense of authority, wisdom, and an unyielding spirit , embodying the ideal ancient Chinese hero.

Right Prompt: This painting is a quintessential example of ancient Chinese ink art , At the top of the painting , towering mountains shrouded in **mist rise majestically**. The mountains' craggy peaks are sketched with fine , precise lines , typical of traditional Chinese ink art. **A slender swirling mists**, meandering waterfall begins its descent here , its water appearing almost ethereal amidst the soft. In the middle section, **the waterfall cascades energetically** , creating a dynamic contrast with the serene mountains above. **Lush pine trees** , rendered with graceful , flowing brush strokes , flank the waterfall. These trees appear to dance with the rhythm of the water , adding a vibrant life to the scene. At the bottom , the waterfall concludes its journey in a tranquil pool. The water's surface is calm , reflecting the surrounding nature and the sky above. Here , **delicate flowers** and **small shrubs** are **depicted along the water's edge** , symbolizing peace and harmony with nature.

Figure 2. Compared to SDXL (Podell et al., 2023) and DALL-E 3 (Betker et al., 2023), our proposed RPG exhibits a superior ability to convey intricate and compositional text prompts within generated images (**colored text denotes critical part**).

Another potential solution is to leverage image understanding feedback (Huang et al., 2023a; Xu et al., 2023; Sun et al., 2023a; Fang et al., 2023) for refining diffusion generation. For instance, GORS (Huang et al., 2023a) finetunes a pretrained text-to-image model with generated images that highly align with the compositional prompts, where the fine-tuning loss is weighted by the text-image alignment reward. Inspired by the reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Stiennon et al., 2020) in natural language processing, ImageReward (Xu et al., 2023) builds a general-purpose reward model to improve text-to-image models in aligning with human preference. More specific related works are in Appendix A. Despite some improvements achieved by these methods, there are still two main limitations in the context of compositional/complex image generation: (i) existing layout-based or attention-based methods can only provide rough and suboptimal spatial guidance, and struggle to deal with overlapped objects (Cao et al., 2023; Hertz et al., 2022; Lian et al., 2023) ; (ii) feedback-based methods require to collect high-quality feedback and incur additional training costs.

To address these limitations, we introduce a new *training-free* text-to-image generation framework, namely ***Recaption, Plan and Generate (RPG)***, unleashing the impressive reasoning ability of multimodal LLMs to enhance the compositionality and controllability of diffusion models. We propose three core strategies in RPG:

Multimodal Recaptioning and Feedback. We use MLLMs to decompose the text prompt into distinct subprompts, and recaption them with more detailed descriptions, offering informative augmented prompt comprehension in diffusion models. we also utilize the MLLMs to directly provide multimodal feedback that identifies cross-modal semantic discrepancies between the generated image and the target prompt, enabling multiple rounds of generation optimization with our iterative editing framework (in Section 2.3).

Chain-of-Thought Planning. Reasoning out visual layout from text prompt is difficult for LLM because there are no visual clues for accurate planning (Lian et al., 2023; Feng et al., 2023). In a pioneering approach, we utilize MLLM to partition the image space into complementary subregions and assign different subprompts to each subregion, breaking down compositional generation tasks into multiple simpler subtasks. The MLLM is pretrained on large-scale vision-language datasets, which sufficiently models the rich vision-language relational prior, thus our regional planning is more powerful than previous planning method. Thoughtfully crafting task instructions and in-context examples, we harness the powerful chain-of-thought reasoning capabilities of MLLMs (Zhang et al., 2023d) for efficient region division.

Complementary Regional Diffusion. Based on the planned non-overlapping subregions, we propose a new complementary regional diffusion, which is fundamentally different from layout-based diffusion. Compared to previous methods that restricts each object within its layout, our initial positions and sizes of objects can be adaptively adjusted during the generation process, and we do not restrict each object to follow its planned subregion. Such region-based division is more flexible and has a richer expressive capacity. We utilize a new layer-wise resize-and-concatenate process, which integrates and optimizes the connectivity between different objects and the background, fusing the base prompt latent with the subprompt latent. Such local-global information fusion ensures that the final generated images are more harmonious and visually appealing. Furthermore, we extend this framework to accommodate editing tasks by employing contour-based regional diffusion, enabling precise manipulation of inconsistent regions targeted for modification. Our RPG can unify both text-guided image generation and editing tasks in a closed-loop fashion (Section 2.3). We compare our RPG framework with previous work in Figure 1 and Figure 2, and summarize main contributions as:

- We propose a new training-free text-to-image generation framework, namely *Recaption, Plan and Generate (RPG)*, to improve the composibility and controllability of diffusion models to the fullest extent.
- RPG is the first to utilize MLLMs as both *multimodal recaptioner and CoT planner* to reason out more informative instructions for steering diffusion models.
- We propose *complementary regional diffusion* to enable extreme collaboration with MLLMs for compositional image generation and precise image editing.
- Our RPG framework is user-friendly, and can be easily generalized to various MLLM architectures and diffusion backbones, as demonstrated in Appendix B.
- Extensive qualitative and quantitative comparisons with SOTA methods, such as SDXL, DALL-E 3 and InstructPix2Pix (Brooks et al., 2023), demonstrate our superior text-guided image generation/editing ability.

2. Method

2.1. Overview of Proposed RPG

In this section, we introduce our novel training-free framework - **R**ecaption, **P**lan and **G**enerate (**RPG**). We delineate three fundamental strategies of our RPG in text-to-image generation (Section 2.2), as depicted in Figure 3. Specifically, given a complex text prompt that includes multiple entities and relationships, we leverage (multimodal) LLMs to *recaption* the prompt by decomposing it into a base prompt and highly descriptive subprompts. Subsequently, we utilize multimodal CoT planning to allocate the split (sub)prompts

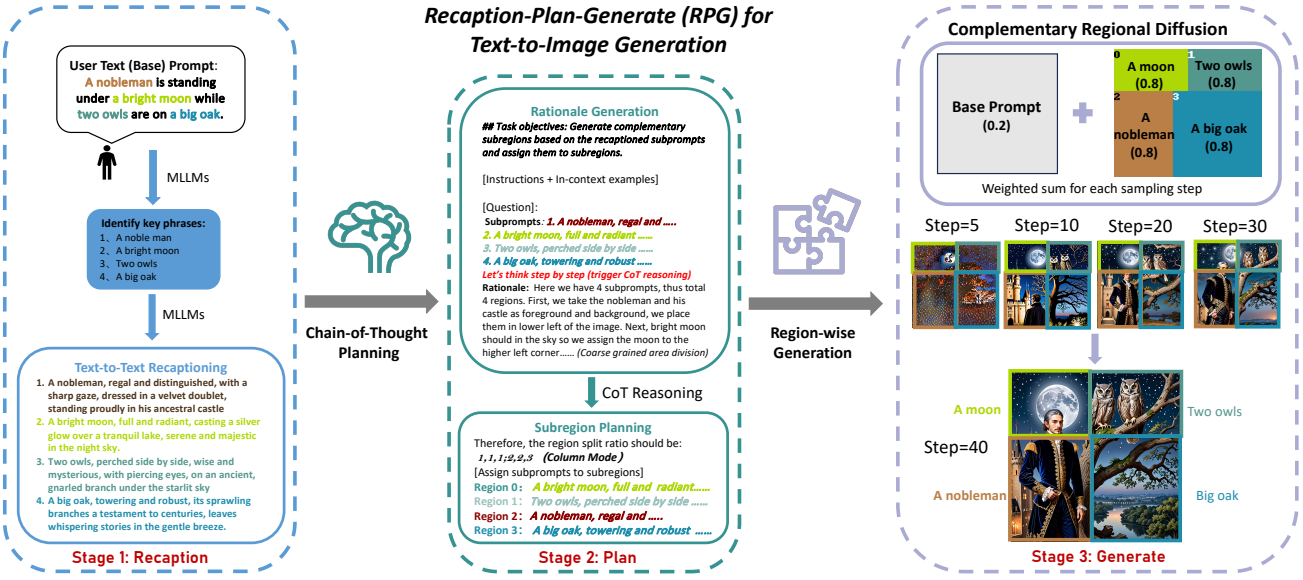


Figure 3. Overview of RPG framework for text-to-image generation. Our editing approach is illustrated in Section 2.3.

to complementary regions along the spatial axes. Building upon these assignments, we introduce *complementary regional diffusion* to independently generate image latents and aggregate them in each sampling step.

Our RPG framework exhibits versatility by extending its application to text-guided image editing with minimal adjustments, as introduced in Section 2.3. For instance, in the recaptioning phase, we utilize MLLMs to analyze the paired target prompt and source image, which results in informative multimodal feedback that captures their cross-modal semantic discrepancies. In multimodal CoT planning, we generate a step-by-step edit plan and produce *precise contours* for our regional diffusion. Furthermore, we demonstrate the ability to execute our RPG workflow in a closed-loop manner for progressive self-refinement, as showcased in Figure 6.

2.2. Text-to-image Generation with RPG

Prompt Recaptioning Let y^c be a complex user prompt which includes multiple entities with different attributes and relationships. We use MLLMs to identify the key phrases in y^c to obtain subpormpts denoted as:

$$\{y^i\}_{i=0}^n = \{y^0, y^1, ..., y^n\} \subseteq y^c, \tag{1}$$

where n denotes the number of key phrases. Inspired by DALL-E 3 (Betker et al., 2023), which uses pre-trained **image-to-text** (I2T) caption models to generate descriptive prompts for images, and construct new datasets with high-quality image-text pairs. In contrast, we leverage the impressive language understanding and reasoning abilities of LLMs and use the LLM as the **text-to-text** (T2T) captioner to further *recaption* each subprompt with more informative detailed descriptions:

$$\{\hat{y}^0, \hat{y}^1, ..., \hat{y}^n\} = \text{Recaption}(\{y^i\}_{i=0}^n). \tag{2}$$

In this way, we can produce denser fine-grained details for each subprompt in order to effectively improve the fidelity of generated image, and reduce the semantic discrepancy between prompt and image.

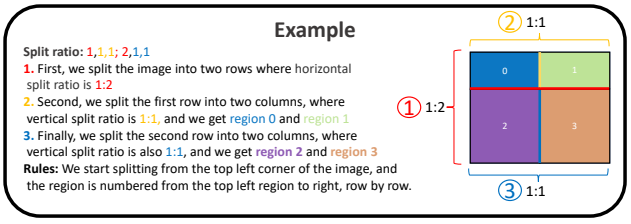


Figure 4. An illustrative example for region division.

CoT Planning for Region Division Based on the recaptioned subprompts, we leverage the powerful multimodal chain-of-thought (CoT) reasoning ability of LLMs (Zhang et al., 2023d) to plan the compositions of final image content for diffusion models. Concretely, we divide image space $H \times W$ into several *complementary regions*, and assign each augmented subprompt $\hat{y}^i$ to specific region R^i:

$$\{R^i\}_{i=0}^n = \{R^0, R^1, ..., R^n\} \subseteq H \times W, \tag{3}$$

In order to produce meaningful and accurate subregions, we need to carefully specify two components for planning region divisions: (i) region parameters: we define that rows are separated by ";" and each column is denoted by a series of numbers separated by commas (e.g., "1,1,1"). To be specific , we first use ";" to split an image into different rows, then within each row, we use commas to split a row

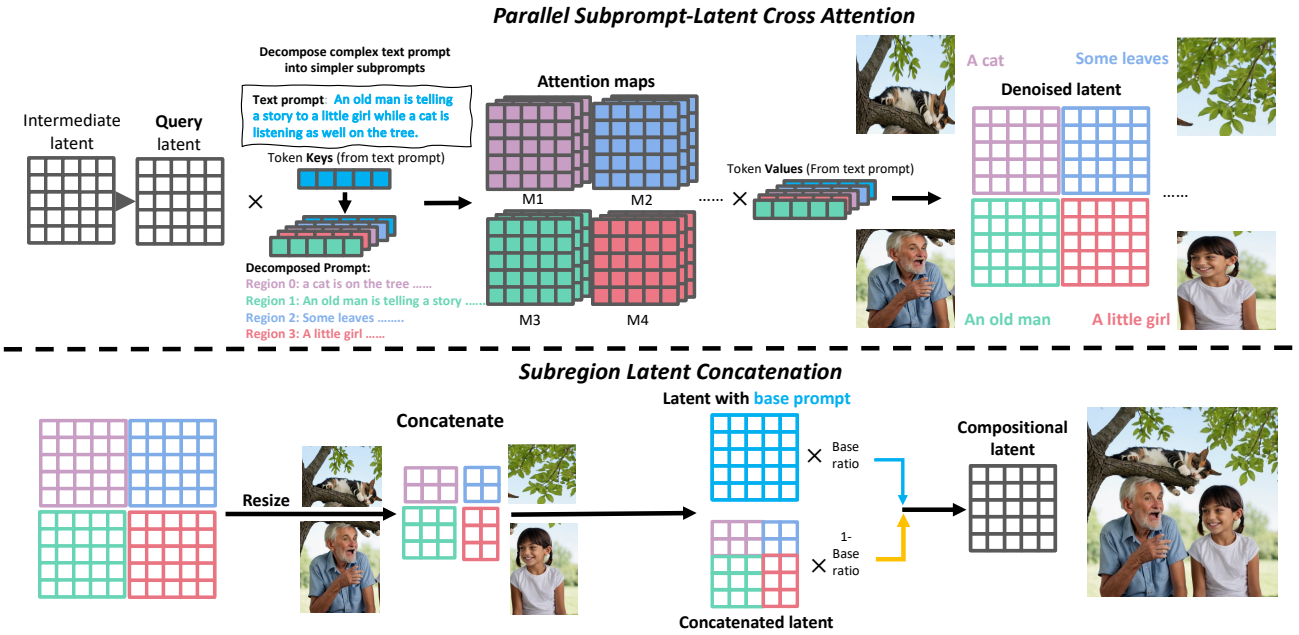


Figure 5. Demonstration of each sampling step in our **complementary regional diffusion** for text-to-image generation.

into different regions, see Figure 4 for better comprehension; (ii) region-wise task specifications to instruct MLLMs: we utilize the CoT reasoning of MLLMs with some designed in-context examples to reason out the plan of region division. We here provide a simplified template of our instructions and in-context examples:

> ***1.Task instruction***
> *You are an smart region planner for image. You should use split ratio to specify the split method of the image, and then recaption each subregion prompts with more descriptive prompts while maintaining the original meaning.*
> ***2.Multi-modal split tutorial***
> .....
> ***3. In-context examples***
> ***User Prompt:*** *A girl with white ponytail and black dress are chatting with a blonde curly hair girl in a white dress in a cafe.*
> *# Key pharses extraction and Recaption*
> *# Split ratio Planning*
> *# Composition Logic*
> *# Aesthetic Considerations:*
> *# Final output*
> ***4.Trigger CoT reasoning ability of MLLMs***
> ***User Prompt:*** *An old man with his dog is looking at a parrot on the tree.*
> *Reasoning: Let's think step by step.....*

To facilitating inferring the region for each subprompt, we adhere to **two key principles** in designing in-context example and generating informative rationales: (i) the objects with same class name (e.g., five apples) will be separately assign to different regions to ensure the numeric accuracy; (ii) If the prompt focuses more on the appearance of a specific entity, we treat the different parts of this entity as different entities (e.g., A green hair twintail in red blouse , wearing blue skirt. $\Longrightarrow$ green hair twintail, red blouse, blue skirt).

Complementary Regional Diffusion Recent works (Liu et al., 2022; Wang et al., 2023; Chefer et al., 2023; Feng et al., 2022) have adjusted cross-attention masks or layouts to facilitate compositional generation. However, these approaches predominantly rely on simply stacking latents, leading to conflicts and ambiguous results in overlapped regions. To address this issue, as depicted in Figure 5, we introduce a novel approach called complementary regional diffusion for region-wise generation and image composition. We extract non-overlapping complementary rectangular regions and apply a **resize-and-concatenate** post-processing step to achieve high-quality compositional generation. Additionally, we enhance coherence by combining the base prompt with recaptioned subprompts to reinforce the conjunction of each generated region and maintain overall image coherence (detailed ablation study in Appendix B). The whole process can be summarized as:

$$\boldsymbol{x}_{t-1} = \mathrm{CRD}(\boldsymbol{x}_t, y^{\text{base}}, \{\hat{y}^i\}_{i=0}^n, \{R^i\}_{i=0}^n, t, s), \quad (4)$$

where s is a fixed random seed, CRD is the abbreviation for complementary regional diffusion.

More concretely, we construct a prompt batch with base

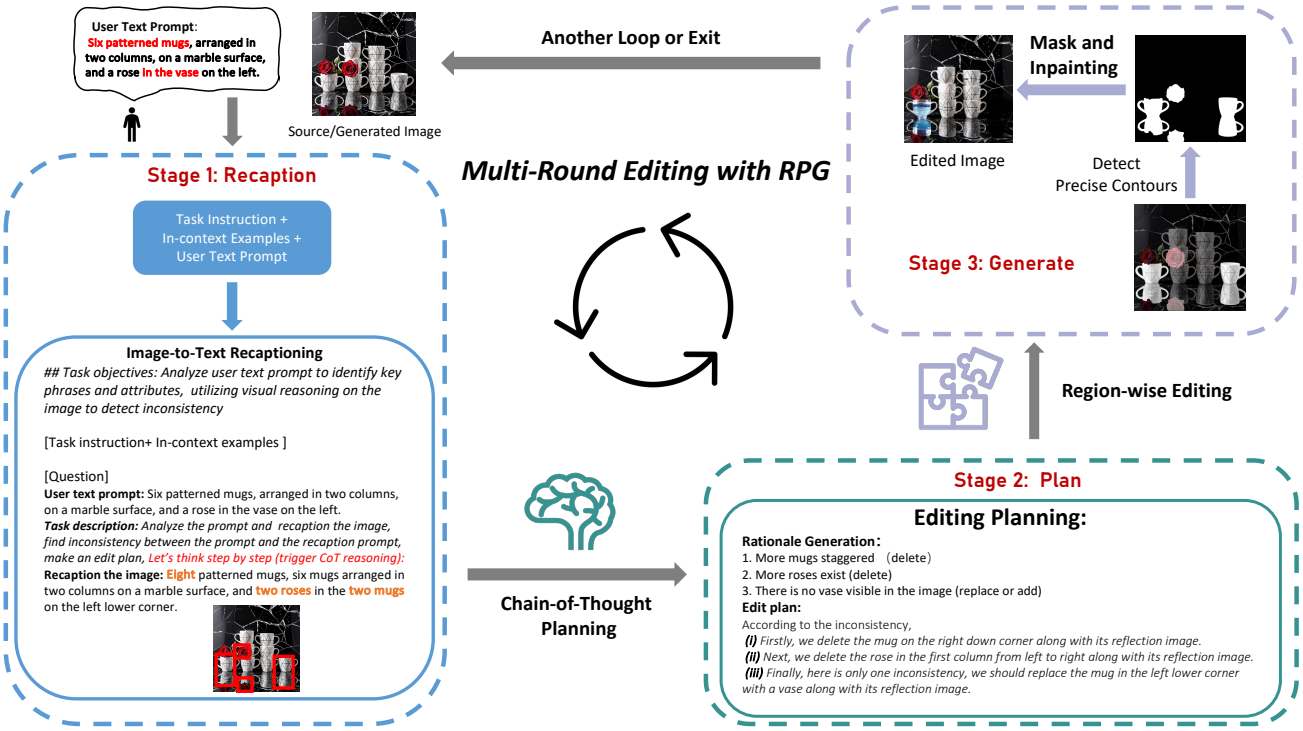


Figure 6. RPG unifies text-guided image generation and editing in a closed-loop approach.

prompt $y^{\text{base}} = y^c$ and the recaptioned subprompts:

$$\text{Prompt Batch:} \quad \{y^{\text{base}}, \{\hat{y}^i\}_{i=0}^n\}. \tag{5}$$

In each timestep, we deliver the prompt batch into the denoising network and manipulate the cross-attention layers to generate different latents $\{\boldsymbol{z}_{t-1}^i\}_{i=0}^n$ and $\boldsymbol{z}_{t-1}^{\text{base}}$ in *parallel*, as demonstrated in Figure 5. We formulate this process as:

$$\boldsymbol{z}_{t-1}^i = \text{Softmax}(\frac{(W_Q \cdot \phi(\boldsymbol{z}_t))(W_K \cdot \psi(\hat{y}^i))^T}{\sqrt{d}})W_V \cdot \psi(\hat{y}^i), \tag{6}$$

where image latent $\boldsymbol{z}_t$ is the query and each subprompt $\hat{y}^i$ works as a key and value. W_Q, W_K, W_V are linear projections and d is the latent projection dimension of the keys and queries. Then, we shall proceed with resizing and concatenating the generated latents $\{\boldsymbol{z}_{t-1}^i\}_{i=0}^n$, according to their assigned region numbers (from 0 to n) and respective proportions. Here we denote each resized latent as:

$$\boldsymbol{z}_{t-1}^i(h, w) = \text{Resize}(\boldsymbol{z}_{t-1}^i, R^i), \tag{7}$$

where h, w are the height and the width of its assigned region R^i. We directly concatenate them along the spatial axes:

$$\boldsymbol{z}_{t-1}^{\text{cat}} = \text{Concatenate}(\{\boldsymbol{z}_{t-1}^i(h, w)\}_{i=0}^n). \tag{8}$$

To ensure a coherent transition in the boundaries of different regions and a harmonious fusion between the background and the entities within each region, we use the weighted sum of the *base latents* $\boldsymbol{z}_{t-1}^{\text{base}}$ and the *concatenated latent* $\boldsymbol{z}_{t-1}^{\text{cat}}$ to produce the final denoising output:

$$\boldsymbol{x}_{t-1} = \beta * \boldsymbol{z}_{t-1}^{\text{base}} + (1 - \beta) * \boldsymbol{z}_{t-1}^{\text{cat}}. \tag{9}$$

Here β is used to achieve a suitable balance between human aesthetic perception and alignment with the complex text prompt of the generated image. It is worth noting that complementary regional diffusion can generalize to arbitrary diffusion backbones including SDXL (Podell et al., 2023), ConPreDiff (Yang et al., 2023b) (in Appendix B) and ControlNet (Zhang et al., 2023a) (in Figure 13).

2.3. Text-Guided Image Editing with RPG

Image Recaptioning Our RPG can also generalize to text-guided image editing tasks as illustrated in Figure 6. In recaptioning stage, RPG adopts MLLMs as a captioner to recaption the source image, and leverage its powerful reasoning ability to identify the fine-grained semantic discrepancies between the image and target prompt. We directly analyze how the input image $\boldsymbol{x}$ aligns with the target prompt y^{tar}. Specifically, we identify the key entities in $\boldsymbol{x}$ and y^{tar}:

$$\begin{aligned} \{y^i\}_{i=0}^n &= \{y^0, y^1, ..., y^n\} \subseteq y^{\text{tar}}, \\ \{e^i\}_{i=0}^m &= \{e^0, e^1, ..., e^m\} \subseteq \text{Recaption}(\boldsymbol{x}), \end{aligned} \tag{10}$$

Then we utilize MLLMs (e.g., GPT4 (OpenAI, 2023)) to check the differences between $\{y^i\}_{i=0}^n$ and $\{e^i\}_{i=0}^m$ regarding numeric accuracy, attribute binding and object relationships. The resulting multimodal understanding feedback would be delivered to MLLMs for reason out editing plans.

Figure 7. Qualitative comparison between our RPG and SOTA text-to-image models (SDXL (Podell et al., 2023) and DALL-E 3 (Betker et al., 2023)), and LLM-grounded diffusion model LMD+ (Lian et al., 2023).

CoT Planning for Editing Based on the captured semantic discrepancies between prompt and image, RPG triggers the CoT reasoning ability of MLLMs with high-quality filtered in-context examples, which involves manually designed step-by-step editing cases such as entity missing/redundancy, attribute mismatch, ambiguous relationships. Here, in our RPG, we introduce three main edit operations for dealing with these issues: addition Add(), deletion Del(), modification Mod(). Take the multimodal feedback as the grounding context, RPG plans out a series of editing instructions. An example Plan$(y^{\text{tar}}, \boldsymbol{x})$ can be denoted as a composed operation list:

$$\{\text{Del}(y^i, \boldsymbol{x}), \cdots, \text{Add}(y^j, \boldsymbol{x}), \cdots, \text{Mod}(y^k, \boldsymbol{x})\}, \quad (11)$$

where $i, j, k <= n$, length$(\text{Plan}(y^{\text{tar}}, x^0)) = L$. In this way, we are able to decompose original complex editing task into simpler editing tasks for more accurate results.

Contour-based Regional Diffusion To collaborate more effectively with CoT-planned editing instructions, we generalize our complementary regional diffusion to text-guided editing. We locate and mask the target contour associated with the editing instruction (Kirillov et al., 2023), and apply diffusion-based inpainting (Rombach et al., 2022) to edit the target contour region according to the planned operation list Plan$(y^{\text{tar}}, \boldsymbol{x})$. Compared to traditional methods that manipulate cross-attention map (Hertz et al., 2022; Cao et al., 2023) for editing, our **mask-and-inpainting** method powered by CoT planning enables more accurate and complex editing operations (i.e., addition, deletion and modification).

Multi-Round Editing for Closed-Loop Refinement Our text-guided image editing workflow is adaptable for a closed-loop self-refined text-to-image generation, which combines the contour-based editing with complementary regional diffusion generation. We could conduct multi-round closed-loop RPG workflow controlled by MLLMs to progressively refine the generated image for aligning closely with the target text prompt. Considering the time efficiency, we set a maximum number of rounds to avoid being trapped in the closed-loop procedure. Based on this closed-loop paradigm, we can unify text-guided generation and editing in our RPG, providing more practical framework for the community.

3. Experiments

3.1. Text-to-Image Generation

Implementation Details Our RPG is general and extensible, we can incorporate **arbitrary** MLLM architectures and diffusion backbones [1][2][3] into the framework. In our experi-

[1]https://github.com/CompVis/stable-diffusion
[2]https://github.com/huggingface/diffusers
[3]https://github.com/hako-mikan/sd-webui-regional-prompter

Table 1. Evaluation results on T2I-CompBench. RPG consistently demonstrates best performance regarding attribute binding, object relationships, and complex compositions. We denote the best score in blue, and the second-best score in green. The baseline data is quoted from Chen et al. (2023a).

Model	Attribute Binding			Object Relationship		Complex↑
	Color ↑	Shape↑	Texture↑	Spatial↑	Non-Spatial↑	
Stable Diffusion v1.4 (Rombach et al., 2022)	0.3765	0.3576	0.4156	0.1246	0.3079	0.3080
Stable Diffusion v2 (Rombach et al., 2022)	0.5065	0.4221	0.4922	0.1342	0.3096	0.3386
Composable Diffusion (Liu et al., 2022)	0.4063	0.3299	0.3645	0.0800	0.2980	0.2898
Structured Diffusion (Feng et al., 2022)	0.4990	0.4218	0.4900	0.1386	0.3111	0.3355
Attn-Exct v2 (Chefer et al., 2023)	0.6400	0.4517	0.5963	0.1455	0.3109	0.3401
GORS (Huang et al., 2023a)	0.6603	0.4785	0.6287	0.1815	0.3193	0.3328
DALL-E 2 (Ramesh et al., 2022)	0.5750	0.5464	0.6374	0.1283	0.3043	0.3696
SDXL (Betker et al., 2023)	0.6369	0.5408	0.5637	0.2032	0.3110	0.4091
PixArt-α (Chen et al., 2023a)	0.6886	0.5582	0.7044	0.2082	0.3179	0.4117
ConPreDiff (Yang et al., 2023b)	0.7019	0.5637	0.7021	0.2362	0.3195	0.4184
RPG (Ours)	0.8743	0.6927	0.8506	0.4781	0.3697	0.5775

ment, we choose GPT-4 (OpenAI, 2023) as the recaptioner and CoT planner, and use SDXL (Podell et al., 2023) as the base diffusion backbone to build our RPG framework. Concretely, in order to trigger the CoT planning ability of MLLMs, we carefully design task-aware template and high-quality in-context examples to conduct few-shot prompting. ***Base prompt*** and its weighted hyperparameter ***base ratio*** are critical in our regional diffusion, we have provide further analysis in Figure 16. When the user prompt includes the entities with same class (e.g., two women, four boys), we need to set higher base ratio to highlight these distinct identities. On the contrary, when user prompt includes the the entities with different class name (e.g., ceramic vase and glass vase), we need lower base ratio to avoid the confusion between the base prompt and subprompts.

Main Results We compare with previous SOTA text-to-image models DALL-E 3 (Betker et al., 2023), SDXL and LMD+ (Lian et al., 2023) in three main compositional scenarios: **(i) Attribute Binding**. Each text prompt in this scenario has multiple attributes that bind to different entities. **(ii) Numeric Accuracy**. Each text prompt in this scenario has multiple entities sharing the same class name, the number of each entity should be greater than or equal to two. **(iii) Complex Relationship**. Each text prompt in this scenario has multiple entities with different attributes and relationships (e.g., spatial and non-spational). As demonstrated in Figure 7, our RPG is significantly superior to previous models in all three scenarios, and achieves remarkable level of both fidelity and precision in aligning with text prompt. We observe that SDXL and DALL-E 3 have poor generation performance regarding numeric accuracy and complex relationship. In contrast, our RPG can effectively plan out precise number of subregions, and utilize proposed complementary regional diffusion to accomplish compositional generation. Compared to LMD+ (Lian et al., 2023), a LLM-grounded layout-based text-to-image diffusion model, our RPG demonstrates both enhanced semantic expression

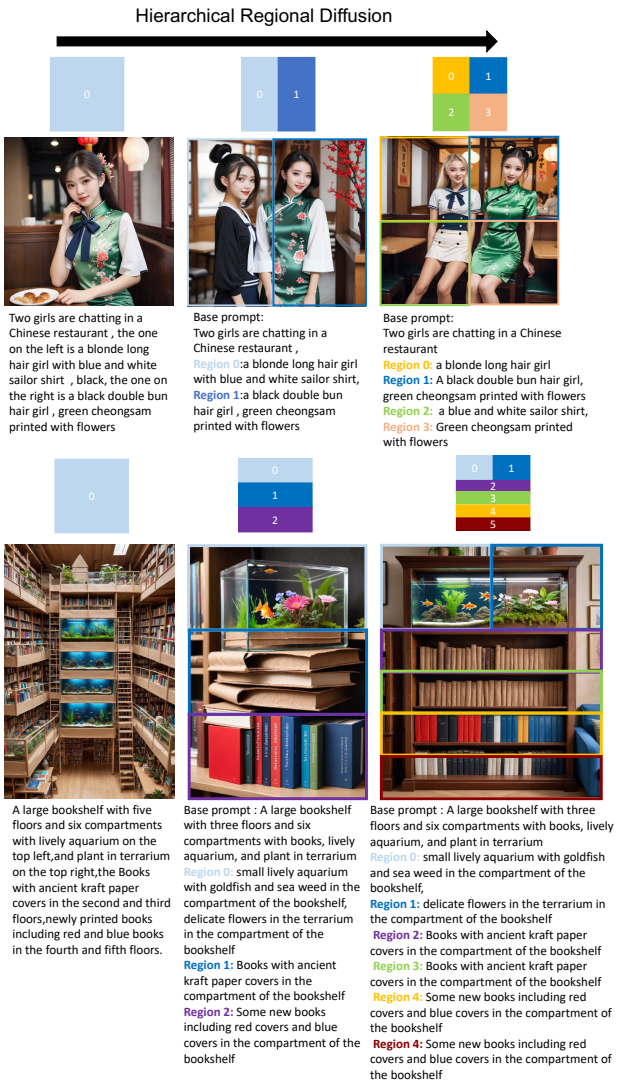


Figure 8. Demonstration of our hierarchical regional diffusion. Diffusion with more hierarchies can produce more satisfying results.

Figure 9. Qualitative comparison in text-guided image editing. We outperform previous all previous powerful methods.

capabilities and image fidelity. We attribute this to our CoT planning and complementary regional diffusion. For quantitative results, we assess the text-image alignment of our method in a comprehensive benchmark, T2I-Compbench (Huang et al., 2023a), which is utilized to evaluate the compositional text-to-image generation capability. In Table 1, we consistently achieve best performance among all methods including both general text-to-image generation and compositional generation. We conduct comprehensive ablation study (in Appendix B) in RPG.

Hierarchical Regional Diffusion We can extend our regional diffusion to a hierarchical format by splitting certain subregion to smaller subregions. As illustrated in Figure 8, when we increase the hierarchies of our region split, RPG can achieve a significant improvement in text-to-image generation. This promising result reveals that our complementary regional diffusion provides a new perspective for handling complex generation tasks and has the potential to generate arbitrarily compositional images.

3.2. Experiment on Text-Guided Image Editing

Qualitative Results In the qualitative comparison of text-guided image editing (Figure 9), we choose some strong

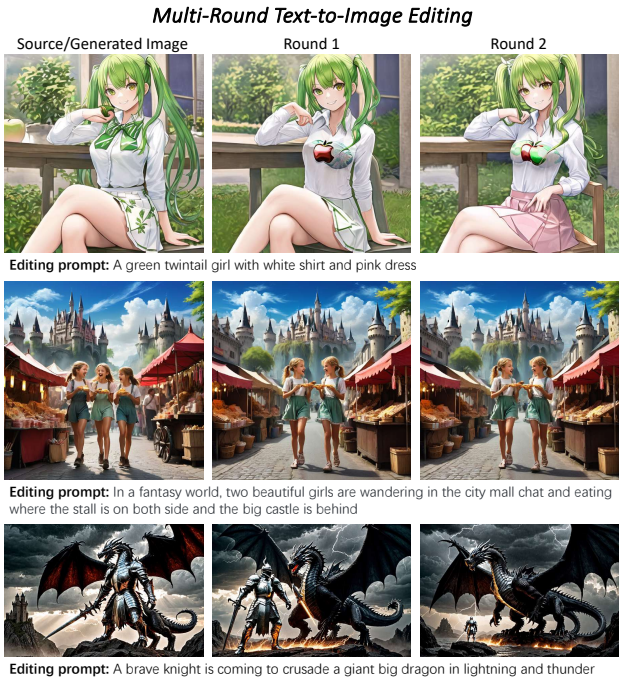


Figure 10. Multi-round text-guided image editing with RPG.

baseline methods, including Prompt2Prompt (Hertz et al., 2022), InstructPix2Pix (Brooks et al., 2023) and MasaCtrl (Cao et al., 2023). Prompt2Prompt and MasaCtrl perform editing mainly through text-grounded cross-attention swap or replacement, InstructPix2Pix aims to learn a model that can follow human instructions. RPG not only produces more precise editing results than previous methods, but also perfectly preserves the semantic structure of source image.

Multi-Round Editing We conduct multi-round editing to evaluate the self-refinement with our RPG framework in Figure 10. We conclude that the self-refinement based on RPG can significantly improve precision, demonstrating the effectiveness of our recaptioning-based multimodal feedback and CoT planning. We also find that RPG is able to achieve satisfying editing results within 3 rounds.

4. Conclusion

In this paper, to address the challenges of complex and compositional text-to-image generation, we propose a SOTA training-free framework RPG, harnessing MLLMs to master diffusion models. In RPG, we propose complementary regional diffusion models to collaborate with our designed MLLM-based recaptioner and planner. Furthermore, RPG unifies text-guided imgae generation and editing in a closed-loop approach, and is capable of generalizing to different MLLM architectures and diffusion backbones. Extensive qualitative and quantitative experiments demonstrate the effectiveness of our RPG. For future work, we will extend it to more challenging applications.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.U23B2048 and U22B2037).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., and Yin, X. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18370–18380, 2023.

Bar-Tal, O., Yariv, L., Lipman, Y., and Dekel, T. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2023.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., and Zheng, Y. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*, 2023.

ChatGPT, I. Introducing chatgpt, 2022.

Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., and Cohen-Or, D. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023a.

Chen, M., Laina, I., and Vedaldi, A. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5343–5353, 2024.

Chen, W.-G., Spiridonova, I., Yang, J., Gao, J., and Li, C. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. *arXiv preprint arXiv:2311.00571*, 2023b.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.

Fang, G., Jiang, Z., Han, J., Lu, G., Xu, H., and Liang, X. Boosting text-to-image diffusion models with fine-grained semantic rewards. *arXiv preprint arXiv:2305.19599*, 2023.

Feng, W., He, X., Fu, T.-J., Jampani, V., Akula, A. R., Narayana, P., Basu, S., Wang, X. E., and Wang, W. Y. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2022.

Feng, W., Zhu, W., Fu, T.-j., Jampani, V., Akula, A., He, X., Basu, S., Wang, X. E., and Wang, W. Y. Layoutgpt: Compositional visual planning and generation with large language models. *arXiv preprint arXiv:2305.15393*, 2023.

Fu, T.-J., Hu, W., Du, X., Wang, W. Y., Yang, Y., and Gan, Z. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023.

Gupta, T. and Kembhavi, A. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Huang, K., Sun, K., Xie, E., Li, Z., and Liu, X. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*, 2023a.

Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., and Zhou, J. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023b.

Iyer, S., Lin, X. V., Pasunuru, R., Mihaylov, T., Simig, D., Yu, P., Shuster, K., Wang, T., Liu, Q., Koura, P. S., et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Koh, J. Y., Fried, D., and Salakhutdinov, R. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*, 2023.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a.

Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., and Lee, Y. J. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22511–22521, 2023b.

Li, Y., Zhang, C., Yu, G., Wang, Z., Fu, B., Lin, G., Shen, C., Chen, L., and Wei, Y. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *arXiv preprint arXiv:2308.10253*, 2023c.

Lian, L., Li, B., Yala, A., and Darrell, T. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Liu, N., Li, S., Du, Y., Torralba, A., and Tenenbaum, J. B. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pp. 423–439. Springer, 2022.

Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., and Qie, X. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

OpenAI, R. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2:3, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Pan, X., Dong, L., Huang, S., Peng, Z., Chen, W., and Wei, F. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*, 2023.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Qu, L., Wu, S., Fei, H., Nie, L., and Chua, T.-S. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 643–654, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text

transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., and Chechik, G. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *arXiv preprint arXiv:2306.08877*, 2023.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. Generative adversarial text to image synthesis. In *International conference on machine learning*, pp. 1060–1069. PMLR, 2016.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Sun, J., Fu, D., Hu, Y., Wang, S., Rassin, R., Juan, D.-C., Alon, D., Herrmann, C., van Steenkiste, S., Krishna, R., et al. Dreamsync: Aligning text-to-image generation with image understanding feedback. *arXiv preprint arXiv:2311.17946*, 2023a.

Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., and Wang, X. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023b.

Surís, D., Menon, S., and Vondrick, C. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Wang, R., Chen, Z., Chen, C., Ma, J., Lu, H., and Lin, X. Compositional text-to-image synthesis with attention map control of diffusion models. *arXiv preprint arXiv:2305.13921*, 2023.

Workshop, B., Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., and Duan, N. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023a.

Wu, T.-H., Lian, L., Gonzalez, J. E., Li, B., and Darrell, T. Self-correcting llm-controlled diffusion models. *arXiv preprint arXiv:2311.16090*, 2023b.

Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., and Shou, M. Z. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pp. 7452–7461, 2023.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023.

Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023a.

Yang, L., Liu, J., Hong, S., Zhang, Z., Huang, Z., Cai, Z., Zhang, W., and Bin, C. Improving diffusion-based image synthesis with context prediction. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., and Yang, M.-H. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023c.

Yang, L., Qian, H., Zhang, Z., Liu, J., and Cui, B. Structure-guided adversarial training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Yang, L., Zhang, Z., Yu, Z., Liu, J., Xu, M., Ermon, S., and CUI, B. Cross-modal contextualized diffusion models for text-guided visual generation and editing. In *ICLR*, 2024b.

Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., and Wang, L. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023d.

Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14246–14255, 2023e.

Yu, L., Shi, B., Pasunuru, R., Muller, B., Golovneva, O., Wang, T., Babu, A., Tang, B., Karrer, B., Sheynin, S., et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023.

Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023a.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zhang, T., Zhang, Y., Vineet, V., Joshi, N., and Wang, X. Controllable text-to-image generation with gpt-4. *arXiv preprint arXiv:2305.18583*, 2023b.

Zhang, X., Yang, L., Cai, Y., Yu, Z., Xie, J., Tian, Y., Xu, M., Tang, Y., Yang, Y., and Cui, B. Realcompo: Dynamic equilibrium between realism and compositionality improves text-to-image diffusion models. *arXiv preprint arXiv:2402.12908*, 2024.

Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., and Sun, T. Enhanced visual instruction tuning for text-rich image understanding. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023c.

Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., and Smola, A. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023d.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Zou, X., Dou, Z.-Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15116–15127, 2023.

A. Related Work

Text-Guided Diffusion Models Diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Song & Ermon, 2020; Song et al., 2020; Yang et al., 2024a) are a promising class of generative models, and Dhariwal & Nichol (2021) have demonstrated the superior image synthesis quality of diffusion model over generative adversarial networks (GANs) (Reed et al., 2016; Creswell et al., 2018). GLIDE (Nichol et al., 2021) and Imagen (Saharia et al., 2022) focus on the text-guided image synthesis, leveraging pre-trained CLIP model (Radford et al., 2021; Raffel et al., 2020) in the image sampling process to improve the semantic alignment between text prompt and generated image. Latent Diffusion Models (LDMs) (Rombach et al., 2022) move the diffusion process from pixel space to latent space for balancing algorithm efficiency and image quality. Recent advancements in text-to-image diffusion models , such as SDXL (Podell et al., 2023) Dreambooth (Ruiz et al., 2023), DALL-E 3 (Betker et al., 2023) and ContextDiff (Yang et al., 2024b), further improve both quality and alignment from different perspectives. Despite their tremendous success, generating high-fidelity images with complex prompt is still challenging (Ramesh et al., 2022; Betker et al., 2023; Huang et al., 2023a). This problem is exacerbated when dealing with compositional descriptions involving spatial relationships, attribute binding and numeric awareness. In this paper, we aim to address this issue by incorporating the powerful CoT reasoning ability of MLLMs into text-to-image diffusion models.

Compositional Diffusion Generation Recent researches aim to improve compositional ability of text-to-image diffusion models. Some approaches mainly introduce additional modules into diffusion models in training (Li et al., 2023b; Avrahami et al., 2023; Zhang et al., 2023a; Mou et al., 2023; Yang et al., 2023e; Huang et al., 2023b;a). For example, GLIGEN (Li et al., 2023b) and ReCo (Yang et al., 2023e) design position-aware adapters on top of the diffusion models for spatially-conditioned image generation. T2I-Adapter and ControlNet (Zhang et al., 2023a; Mou et al., 2023) specify some high-level features of images for controlling semantic structures (Zhang et al., 2023b). These methods, however, result in additional training and inference costs. Training-free methods aim to steer diffusion models through manipulating latent or cross-attention maps according to spatial or semantic constraints during inference stages (Feng et al., 2022; Liu et al., 2022; Hertz et al., 2022; Cao et al., 2023; Chen et al., 2024; Chefer et al., 2023). Composable Diffusion (Liu et al., 2022) decomposes a compositional prompt into smaller sub-prompts to generate distinct latents and combines them with a score function. Chen et al. (2024) and Lian et al. (2023) utilize the bounding boxes (layouts) to propagate gradients back to the latent and enable the model to manipulate the cross-attention maps towards specific regions. Other methods apply Gaussian kernels (Chefer et al., 2023) or incorporate linguistic features (Feng et al., 2022; Rassin et al., 2023) to manipulate the cross-attention maps. Nevertheless, such manipulation-based methods can only make rough controls, and often lead to unsatisfied compositional generation results, especially when dealing with overlapped objects (Lian et al., 2023; Cao et al., 2023). Hence, we introduce an effective *training-free complementary regional diffusion model*, grounded by MLLMs, to progressively refine image compositions with more precise control in the sampling process.

Multimodal LLMs for Image Generation Large Language Models (LLMs) (ChatGPT, 2022; Chung et al., 2022; Zhang et al., 2022; Iyer et al., 2022; Workshop et al., 2022; Muennighoff et al., 2022; Zeng et al., 2022; Taylor et al., 2022; Chowdhery et al., 2023; Chen et al., 2023b; Zhu et al., 2023; Touvron et al., 2023a; Yang et al., 2023a; Li et al., 2023a) have profoundly impacted the AI community. Leading examples like ChatGPT (ChatGPT, 2022) have showcased the advanced language comprehension and reasoning skills through techniques such as instruction tuning (Ouyang et al., 2022; Li et al., 2023c; Zhang et al., 2023c; Liu et al., 2023). Further, Multimodal Large language Models (MLLMs), (Koh et al., 2023; Yu et al., 2023; Sun et al., 2023b; Dong et al., 2023; Fu et al., 2023; Pan et al., 2023; Wu et al., 2023a; Zou et al., 2023; Yang et al., 2023d; Gupta & Kembhavi, 2023; Surís et al., 2023) integrate LLMs with vision models to extend their impressive abilities from language tasks to vision tasks, including image understanding, reasoning and synthesis. The collaboration between LLMs (ChatGPT, 2022; OpenAI, 2023) and diffusion models (Ramesh et al., 2022; Betker et al., 2023) can significantly improve the text-image alignment as well as the quality of generated images (Yu et al., 2023; Chen et al., 2023b; Dong et al., 2023; Wu et al., 2023b; Feng et al., 2023; Pan et al., 2023). For instance, GILL (Koh et al., 2023) can condition on arbitrarily interleaved image and text inputs to synthesize coherent image outputs, and Emu (Sun et al., 2023b) stands out as a generalist multimodal interface for both image-to-text and text-to-image tasks. Recently, LMD (Lian et al., 2023) utilizes LLMs to enhance the compositional generation of diffusion models by generating images grounded on bounding box layouts from the LLM (Li et al., 2023b). However, existing works mainly incorporate the LLM as a simple plug-in component into diffusion models, or simply take the LLM as a layout generator to control image compositions. In contrast, we utilize MLLMs to plan and manipulate the image compositions for diffusion models where MLLMs serves as a global task planner in both *region-based* generation and editing process.

B. Model Analysis

Generalizing to Various LLMs and Diffusion Backbones Our RPG framework is of great generalization ability, and can be easily generalized to various (M)LLM architectures (in Figure 11) and diffusion backbones (in Figure 12). We observe that both LLM and diffusion architectures can influence the generation results. We also generalize RPG to ControlNet (Zhang et al., 2023a) for incorporating more conditional modalities. As demonstrated in Figure 13, our RPG can significantly improve the composibility of original ControlNet in both image fidelity and textual semantic alignment.

Figure 11. Generalizing RPG to different (multimodal) LLM architectures, including Llama 2 (Touvron et al., 2023b), Vicuna (Chiang et al., 2023) and MiniGPT-4 (Zhu et al., 2023).

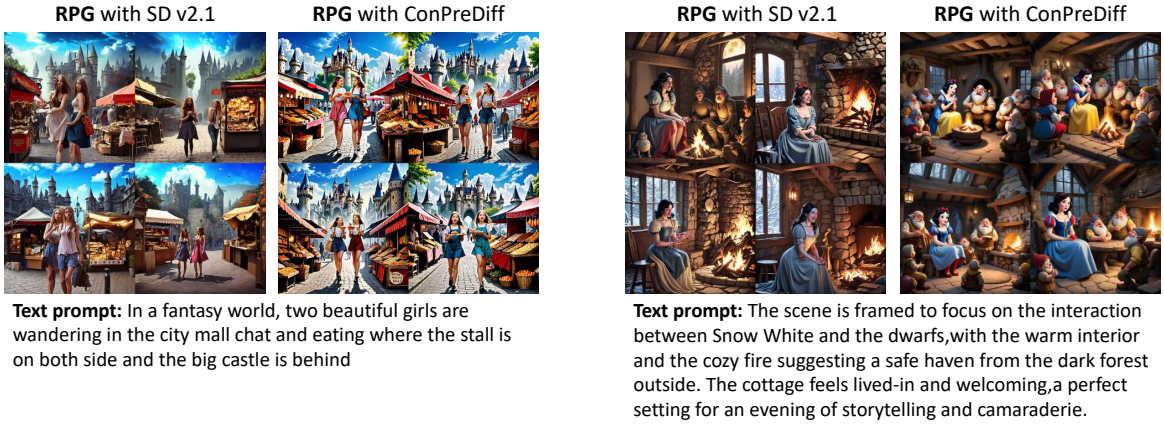


Figure 12. Generalizing RPG to different diffusion backbones, Stable Diffusion v2.1 (Rombach et al., 2022) and recent SOTA diffusion model ConPreDiff (Yang et al., 2023b).

Effect of Recaptioning We conduct ablation study about the recaptioning, and show the result in Figure 14. From the result, we observe that without recaptioning, the model tends to ignore some key words in the generated images. Our recaptioning can describe these key words with high-informative and denser details, thus generating more delicate and precise images.

Effect of CoT Planning In the ablation study about CoT planning, as demonstrated in Figure 15, we observe that the model without CoT planning fail to parse and convey complex relationships from text prompt. In contrast, our CoT planning can help the model better identify fine-grained attributes and relationships from text prompt, and express them through a more realistic planned composition.

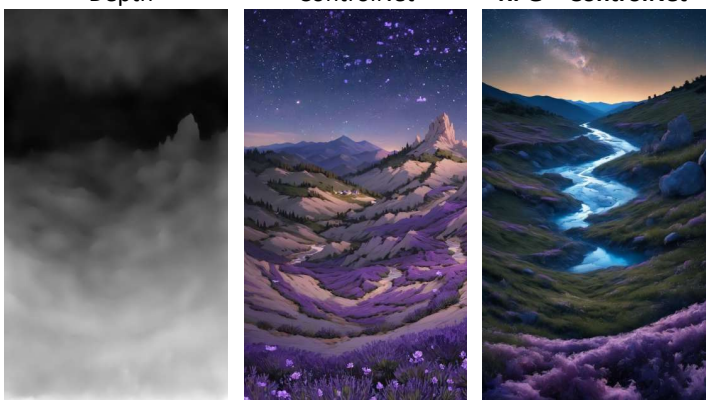
Effect of Base Prompt In RPG, we leverage the generated latent from base prompt in diffusion models to improve the coherence of image compositions. Here we conduct more analysis on it in Figure 16. From the results, we find that the

Pose | ControlNet | **RPG + ControlNet**

User input: A beautiful black hair girl with her eyes closed in champagne long sleeved formal dress standing in her bright room with delicate blue vases with pink roses on the left and some white roses, filled with upgraded growth all around on the right

Base prompt: A beautiful girl standing in her bright room
Region 0: A delicate blue vases with pink roses
Region 1: A beautiful black hair girl with her eyes closed in champagne long sleeved formal dress
Region 2: Some white roses, filled with upgraded growth all around

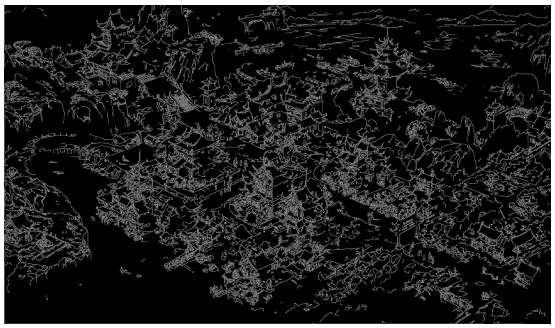
Depth | ControlNet | **RPG + ControlNet**



User input: Under the clear starry sky, clear river water flows in the mountains, and the lavender flower sea in front of me dances with the wind, a peaceful, beautiful, and harmonious atmosphere

Base prompt: Under the clear starry sky, clear river water flows in the mountains, and the lavender flower sea in front of me dances with the wind, A peaceful, beautiful, and harmonious atmosphere
Region 0: Clear starry sky, many stars twinkling in the night sky
Region 1: The blue stream flows through the valleys covered with colorful flowers ,reflecting the starry sky
Region 2: The sea of lavender dancing in the wind,

Canny Edge



ControlNet

RPG + ControlNet

User input: From left to right, an ancient Chinese city in spring, summer, autumn and winter in four different regions

Region 0: The city in spring where every where is lively and colorful with pink flowers mountains
Region 1: The city with buildings with green rooves in summer, where a lake is beautiful
Region 2: The city in autumn with buildings with orange rooves, where every where is light orange and golden
Region 3: The city in winter with trees covered with snow and blue rooves covered with snow, where every is white and covered with snow

Figure 13. Our RPG framework can extend text-to-image generation with more conditions (e.g., pose, depth and canny edge) by utilizing ControlNet (Zhang et al., 2023a). Compared to original ControlNet, RPG significantly improves its prompt understanding by decomposing "user input" into the combination of base prompt and subprompts, and further enhance its compositional semantic alignment of generated images by performing region-wise diffusion generation (in Section 2.2).

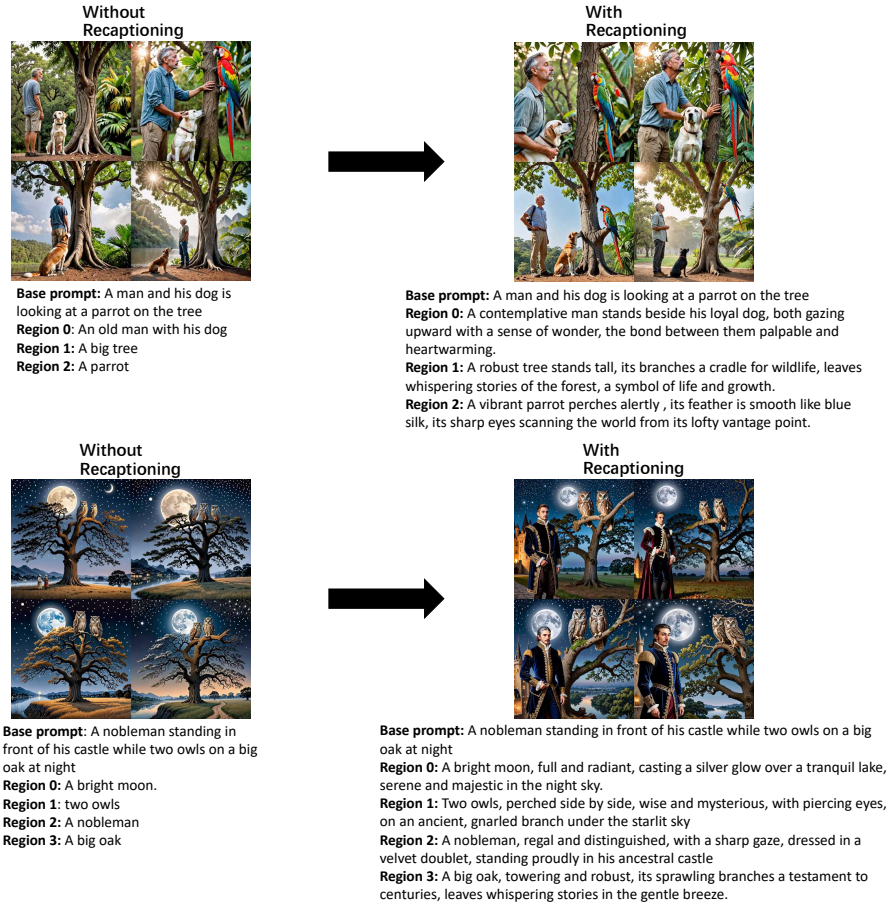


Figure 14. Ablation study of recaptioning in RPG.

proper ratio of base prompt can benefit the conjunction of different subregions, enabling more natural composition. Another finding is that excessive base ratio may result in undesirable results because of the confusion between the base prompt and regional prompt.

Figure 15. Ablation study of CoT planning in RPG.

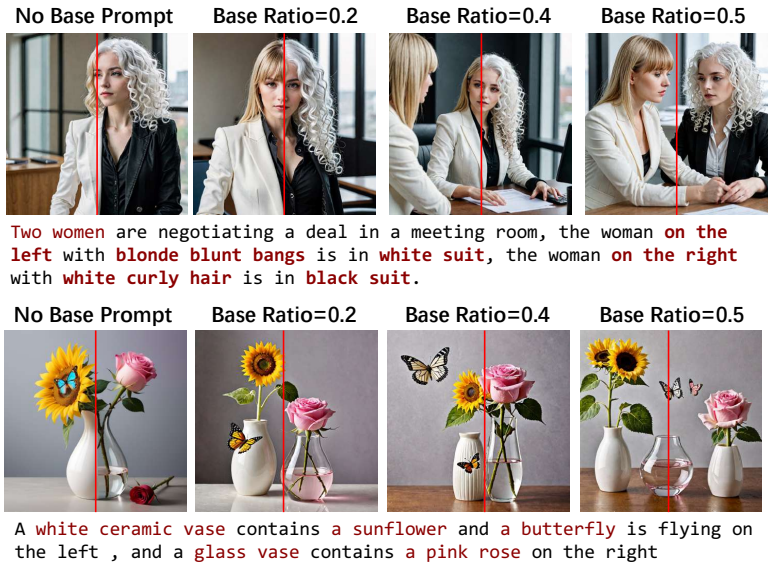


Figure 16. Ablation study of base prompt in complementary regional diffusion.