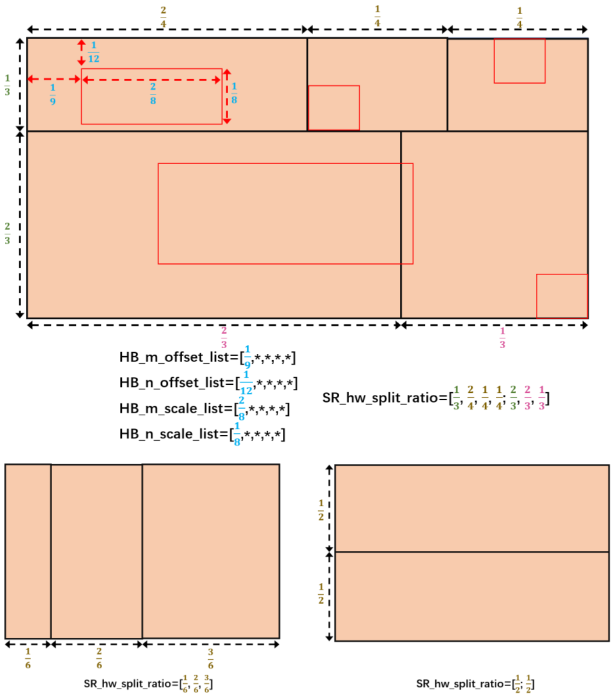
RAGD: Regional-Aware Diffusion Model for Text-to-Image Generation

Supplementary Material

1. More Model Details

In Figure 1, we present example images of $m^i = \{m^i_{offset}, m^i_{scale}\}$, $n^i = \{n^i_{offset}, n^i_{scale}\}$ in regional hard binding and $o^i = \{h^i_{split}, w^i_{split}\}$ in regional soft refinement. As shown, m^i and n^i are primarily responsible for precise positional control of objects, ensuring that the generated content adheres to the specified spatial layout. In contrast, o^i focuses on expanding the global area of objects, thereby enhancing interaction and fusion with adjacent regions. This collaborative design significantly improves the attribute fidelity of the generated images while promoting harmonious coordination across different regions, resulting in more cohesive and visually consistent outputs.



$\frac{2}{4}$
$\frac{1}{4}$
$\frac{1}{4}$
$\frac{1}{12}$
$\frac{1}{3}$
$\frac{1}{9}$
$\frac{2}{8}$
$\frac{1}{8}$
$\frac{2}{3}$
$\frac{2}{3}$
$\frac{1}{3}$
HB_m_offset_list=[$\frac{1}{9}$,*,*,*,*]
HB_n_offset_list=[$\frac{1}{12}$,*,*,*,*]
HB_m_scale_list=[$\frac{2}{8}$,*,*,*,*]
HB_n_scale_list=[$\frac{1}{8}$,*,*,*,*]
SR_hw_split_ratio=[$\frac{1}{3}$, $\frac{2}{4}$, $\frac{1}{4}$, $\frac{1}{4}$; $\frac{2}{3}$, $\frac{2}{3}$, $\frac{1}{3}$]
$\frac{1}{2}$
$\frac{1}{2}$
$\frac{1}{6}$
$\frac{2}{6}$
$\frac{3}{6}$
SR_hw_split_ratio=[$\frac{1}{6}$, $\frac{2}{6}$, $\frac{3}{6}$]
SR_hw_split_ratio=[$\frac{1}{2}$; $\frac{1}{2}$]

Figure 1. Several schematic diagrams of regional division.

2. Usage of MLLM.

In order to facilitate quantitative testing and user use, we use MLLM to automatically segment spatial regions and generate bounding boxes and sub-prompts from a long prompt, as shown in Figure 2. It is worth noting that our method does not inherently rely on MLLM, as this process can also be performed manually.

3. More Analysis

3.1. Effectiveness of Parameter r.

We introduced a parameter r to control the frequency of hard binding applications. As shown in Figure 3, excessive hard binding limits the opportunity for subsequent soft refinement, weakening its impact. This may diminish interactions between adjacent objects, potentially causing noticeable boundaries to reappear. By adjusting r, we aim to balance precise object placement with smooth integration, optimizing both the structure and cohesion of the generated image.

3.2. Effectiveness of Coefficient δ.

We introduced the coefficient δ to modulate the intensity of regional soft refinement. As illustrated in Figure 3, excessively low or high δ values may result in noticeable partitioning or slight misalignment between text and image. Setting δ to an optimal level enhances the fusion of image regions, yielding a more natural overall output.

3.3. Quantitative Experiment of Repainting.

In contrast to conventional inpainting approaches, the proposed repainting scheme in this paper eliminates the dependency on original image inputs, enabling iterative local adjustments through the reuse of parameter configurations from prior generation processes. To further validate the feasibility of our methodology, we conducted quantitative comparative experiments with advanced inpainting model BrushNet [2] and RAGD. We randomly selected 100 distinct text prompts from the Geneval and T2I-CompBench benchmark to generate initial images. The resulting outputs served as image inputs for BrushNet, while retaining corresponding generation parameters as input conditions for RAGD. As demonstrated in Table 1, without requiring secondary training or external inpainting models, RAGD exhibits significant advantages across all evaluation metrics.

Metrics	Image Quality		Masked Region Preservation		Text Align
Models	$IR_{\times 10} \uparrow$	$AS \uparrow$	$PSNR \uparrow$	$LPIPS_{\times 10^3} \downarrow$	$CLIPSim \uparrow$
BrushNet	5.1006	5.8474	22.4292	8.7020	21.6796
RAG	**6.3304**	**6.1093**	**26.5401**	**7.7600**	**24.3563**

Table 1. Comparison with state-of-the-art inpainting solutions.

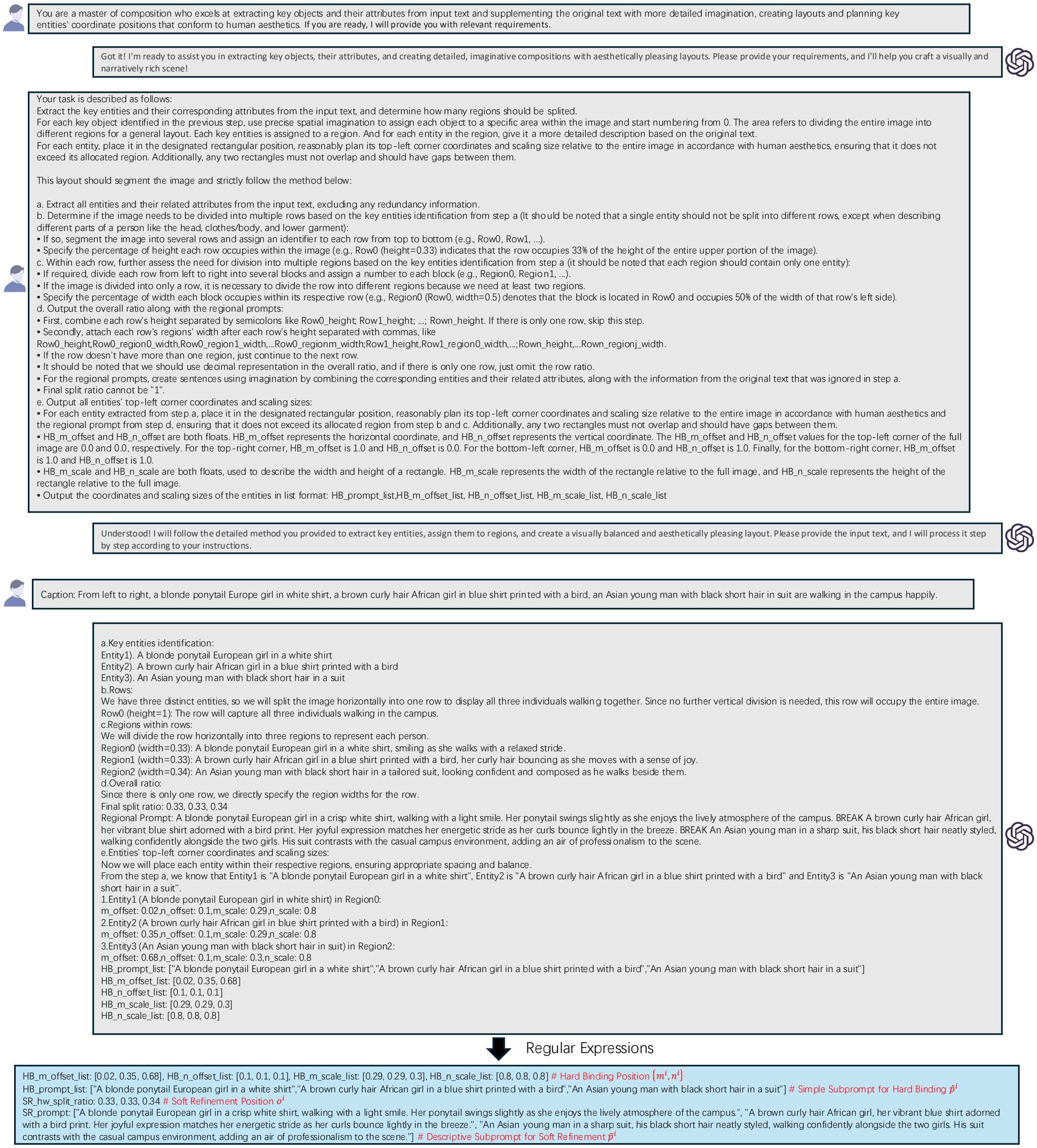
3.4. Stability of RAGD

We validated stability by applying different random seeds under identical parameter settings. As shown in Figure 4, varying seeds generated results with diverse characteristics, highlighting the model's creative potential. However, in scenarios with a higher number of regions, occasional suboptimal outcomes were observed, such as incorrect object counts or pronounced boundaries between regions. These cases can be regarded as acceptable variations, given that similar issues are inherent in most existing T2I models during image generation.

3.5. More Qualitative Results

As shown in Figure 5, we present additional qualitative results comparing RAGD, Flux [1] and RPG [3] that further highlighting RAGD's superior performance in attribute depiction and spatial control. Through the generation of multi-region complex scenes, RAGD demonstrates its ability to accurately capture the details described in the input text while maintaining overall coherence and naturalness in the generated content.

Figure 6 showcases the potential applications of RAGD in image repainting tasks, such as removing unsatisfactory areas from previously generated results or adding new objects. These examples underscore RAGD's flexibility and controllability in image repainting, particularly its capability to execute complex modifications without requiring model fine-tuning.



You are a master of composition who excels at extracting key objects and their attributes from input text and supplementing the original text with more detailed imagination, creating layouts and planning key entities' coordinate positions that conform to human aesthetics. If you are ready, I will provide you with relevant requirements.

Got it! I'm ready to assist you in extracting key objects, their attributes, and creating detailed, imaginative compositions with aesthetically pleasing layouts. Please provide your requirements, and I'll help you craft a visually and narratively rich scene!

Your task is described as follows:
Extract the key entities and their corresponding attributes from the input text, and determine how many regions should be splited.
For each key object identified in the previous step, use precise spatial imagination to assign each object to a specific area within the image and start numbering from 0. The area refers to dividing the entire image into different regions for a general layout. Each key entities is assigned to a region. And for each entity in the region, give it a more detailed description based on the original text.
For each entity, place it in the designated rectangular position, reasonably plan its top-left corner coordinates and scaling size relative to the entire image in accordance with human aesthetics, ensuring that it does not exceed its allocated region. Additionally, any two rectangles must not overlap and should have gaps between them.

This layout should segment the image and strictly follow the method below:

a. Extract all entities and their related attributes from the input text, excluding any redundancy information.
b. Determine if the image needs to be divided into multiple rows based on the key entities identification from step a (It should be noted that a single entity should not be split into different rows, except when describing different parts of a person like the head, clothes/body, and lower garment):
- If so, segment the image into several rows and assign an identifier to each row from top to bottom (e.g., Row0, Row1, ...).
- Specify the percentage of height each row occupies within the image (e.g., Row0 (height=0.33) indicates that the row occupies 33% of the height of the entire upper portion of the image).

c. Within each row, further assess the need for division into multiple regions based on the key entities identification from step a (it should be noted that each region should contain only one entity):
- If required, divide each row from left to right into several blocks and assign a number to each block (e.g., Region0, Region1, ...).
- If the image is divided into only a row, it is necessary to divide the row into different regions because we need at least two regions.
- Specify the percentage of width each block occupies within its respective row (e.g., Region0 (Row0, width=0.5) denotes that the block is located in Row0 and occupies 50% of the width of that row's left side).

d. Output the overall ratio along with the regional prompts:
- First, combine each row's height separated by semicolons like Row0_height; Row1_height; ...; Rown_height. If there is only one row, skip this step.
- Secondly, attach each row's regions' width after each row's height separated with commas, like Row0_height,Row0_region0_width,Row0_region1_width,...Row0_regionm_width;Row1_height,Row1_region0_width,...;Rown_height,...Rown_regionj_width.
- If the row doesn't have more than one region, just continue to the next row.
- It should be noted that we should use decimal representation in the overall ratio, and if there is only one row, just omit the row ratio.
- For the regional prompts, create sentences using imagination by combining the corresponding entities and their related attributes, along with the information from the original text that was ignored in step a.
- Final split ratio cannot be "1".

e. Output all entities' top-left corner coordinates and scaling sizes:
- For each entity extracted from step a, place it in the designated rectangular position, reasonably plan its top-left corner coordinates and scaling size relative to the entire image in accordance with human aesthetics and the regional prompt from step d, ensuring that it does not exceed its allocated region from step b and c. Additionally, any two rectangles must not overlap and should have gaps between them.
- HB_m_offset and HB_n_offset are both floats. HB_m_offset represents the horizontal coordinate, and HB_n_offset represents the vertical coordinate. The HB_m_offset and HB_n_offset values for the top-left corner of the full image are 0.0 and 0.0, respectively. For the top-right corner, HB_m_offset is 1.0 and HB_n_offset is 0.0. For the bottom-left corner, HB_m_offset is 0.0 and HB_n_offset is 1.0. Finally, for the bottom-right corner, HB_m_offset is 1.0 and HB_n_offset is 1.0.
- HB_m_scale and HB_n_scale are both floats, used to describe the width and height of a rectangle. HB_m_scale represents the width of the rectangle relative to the full image, and HB_n_scale represents the height of the rectangle relative to the full image.
- Output the coordinates and scaling sizes of the entities in list format: HB_prompt_list,HB_m_offset_list, HB_n_offset_list, HB_m_scale_list, HB_n_scale_list

Understood! I will follow the detailed method you provided to extract key entities, assign them to regions, and create a visually balanced and aesthetically pleasing layout. Please provide the input text, and I will process it step by step according to your instructions.

Caption: From left to right, a blonde ponytail Europe girl in white shirt, a brown curly hair African girl in blue shirt printed with a bird, an Asian young man with black short hair in suit are walking in the campus happily.

a.Key entities identification:
Entity1). A blonde ponytail European girl in a white shirt
Entity2). A brown curly hair African girl in a blue shirt printed with a bird
Entity3). An Asian young man with black short hair in a suit
b.Rows:
We have three distinct entities, so we will split the image horizontally into one row to display all three individuals walking together. Since no further vertical division is needed, this row will occupy the entire image.
Row0 (height=1): The row will capture all three individuals walking in the campus.
c.Regions within rows:
We will divide the row horizontally into three regions to represent each person.
Region0 (width=0.33): A blonde ponytail European girl in a white shirt, smiling as she walks with a relaxed stride.
Region1 (width=0.33): A brown curly hair African girl in a blue shirt printed with a bird, her curly hair bouncing as she moves with a sense of joy.
Region2 (width=0.34): An Asian young man with black short hair in a tailored suit, looking confident and composed as he walks beside them.
d.Overall ratio:
Since there is only one row, we directly specify the region widths for the row.
Final split ratio: 0.33, 0.33, 0.34
Regional Prompt: A blonde ponytail European girl in a crisp white shirt, walking with a light smile. Her ponytail swings slightly as she enjoys the lively atmosphere of the campus. BREAK A brown curly hair African girl, her vibrant blue shirt adorned with a bird print. Her joyful expression matches her energetic stride as her curls bounce lightly in the breeze. BREAK An Asian young man in a sharp suit, his black short hair neatly styled, walking confidently alongside the two girls. His suit contrasts with the casual campus environment, adding an air of professionalism to the scene.
e.Entities' top-left corner coordinates and scaling sizes:
Now we will place each entity within their respective regions, ensuring appropriate spacing and balance.
From the step a, we know that Entity1 is "A blonde ponytail European girl in a white shirt", Entity2 is "A brown curly hair African girl in a blue shirt printed with a bird" and Entity3 is "An Asian young man with black short hair in a suit".
1.Entity1 (A blonde ponytail European girl in white shirt) in Region0:
m_offset: 0.02,n_offset: 0.1,m_scale: 0.29,n_scale: 0.8
2.Entity2 (A brown curly hair African girl in blue shirt printed with a bird) in Region1:
m_offset: 0.35,n_offset: 0.1,m_scale: 0.29,n_scale: 0.8
3.Entity3 (An Asian young man with black short hair in suit) in Region2:
m_offset: 0.68,n_offset: 0.1,m_scale: 0.3,n_scale: 0.8
HB_prompt_list: ["A blonde ponytail European girl in a white shirt","A brown curly hair African girl in a blue shirt printed with a bird","An Asian young man with black short hair in a suit"]
HB_m_offset_list: [0.02, 0.35, 0.68]
HB_n_offset_list: [0.1, 0.1, 0.1]
HB_m_scale_list: [0.29, 0.29, 0.3]
HB_n_scale_list: [0.8, 0.8, 0.8]

HB_m_offset_list: [0.02, 0.35, 0.68], HB_n_offset_list: [0.1, 0.1, 0.1], HB_m_scale_list: [0.29, 0.29, 0.3], HB_n_scale_list: [0.8, 0.8, 0.8] # Hard Binding Position $\{m^i, n^i\}$
HB_prompt_list: ["A blonde ponytail European girl in a white shirt","A brown curly hair African girl in a blue shirt printed with a bird","An Asian young man with black short hair in a suit"] # Simple Subprompt for Hard Binding $\hat{p}^i$
SR_hw_split_ratio: 0.33, 0.33, 0.34 # Soft Refinement Position o^i
SR_prompt: ["A blonde ponytail European girl in a crisp white shirt, walking with a light smile. Her ponytail swings slightly as she enjoys the lively atmosphere of the campus.", "A brown curly hair African girl, her vibrant blue shirt adorned with a bird print. Her joyful expression matches her energetic stride as her curls bounce lightly in the breeze.", "An Asian young man in a sharp suit, his black short hair neatly styled, walking confidently alongside the two girls. His suit contrasts with the casual campus environment, adding an air of professionalism to the scene."] # Descriptive Subprompt for Soft Refinement $\tilde{p}^i$

Figure 2. Usage of MLLM.

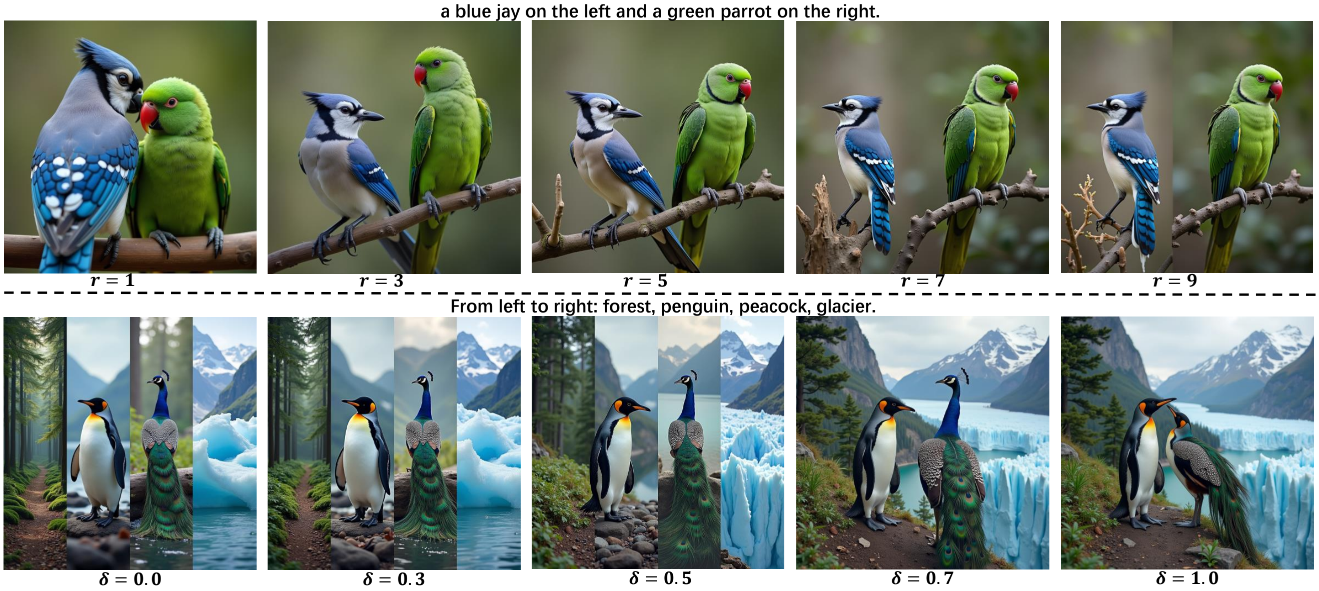

a blue jay on the left and a green parrot on the right.
$r = 1$
$r = 3$
$r = 5$
$r = 7$
$r = 9$
From left to right: forest, penguin, peacock, glacier.
$\delta = 0.0$
$\delta = 0.3$
$\delta = 0.5$
$\delta = 0.7$
$\delta = 1.0$

Figure 3. Qualitative analysis of hard binding steps r and soft refinement strength δ. A few steps of binding is sufficient for regional completeness, and appropriate soft refinement intensity leads to improved regional coherence.

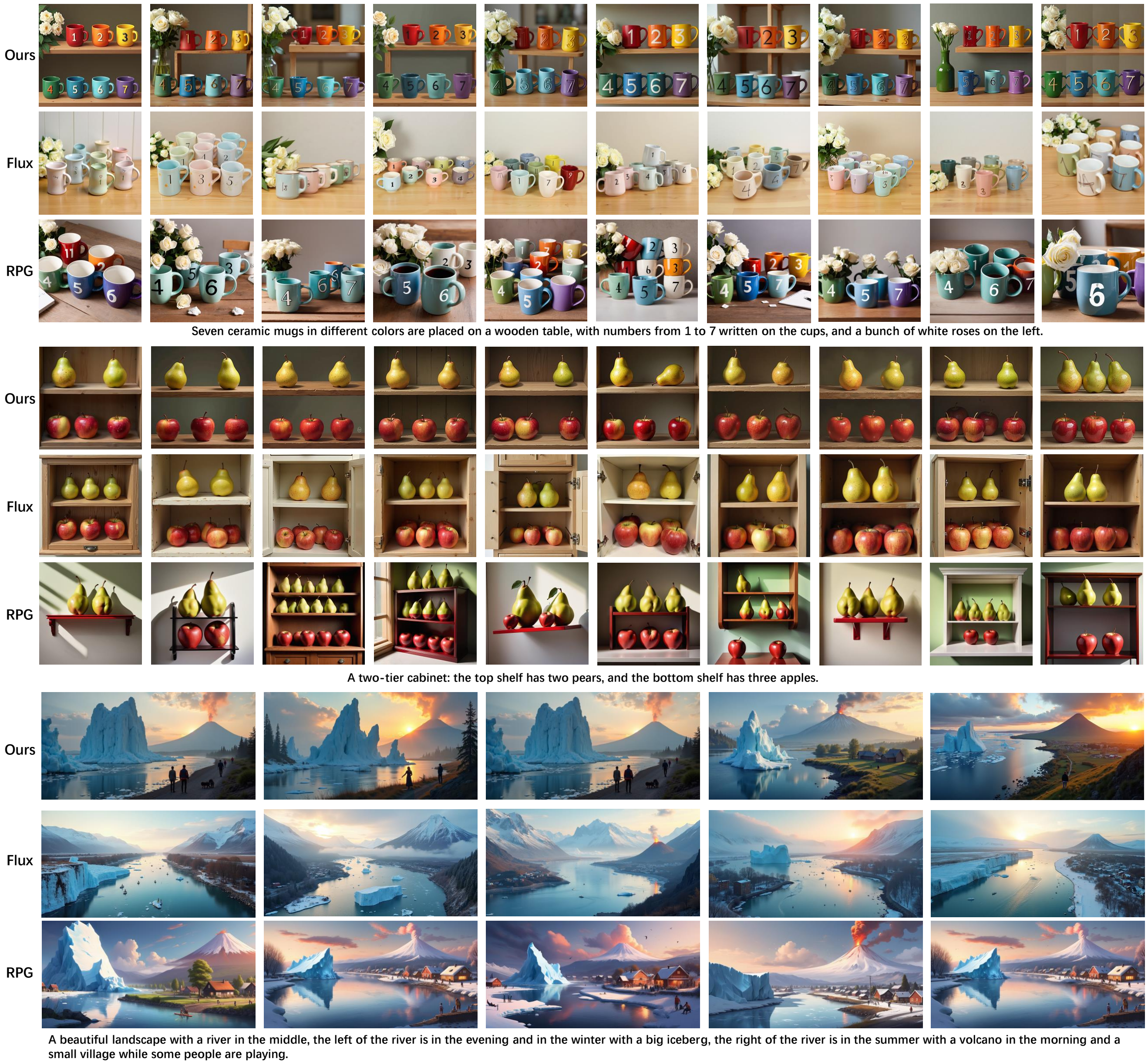
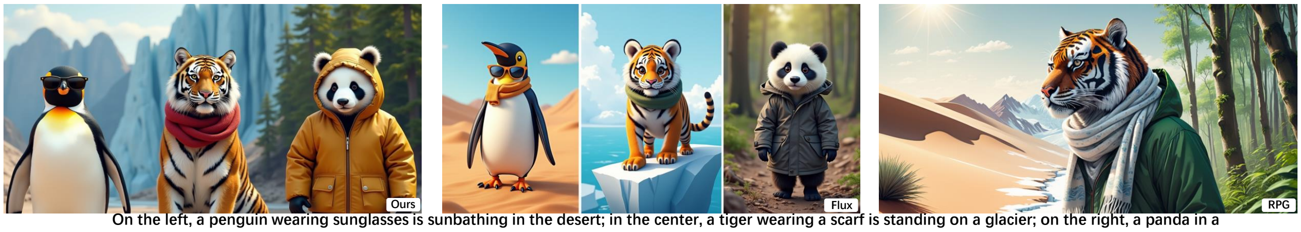
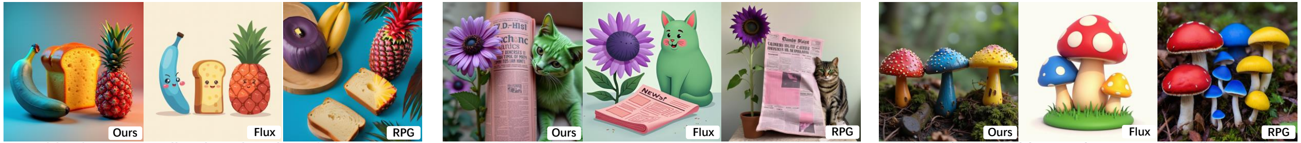

Ours
Flux
RPG
Seven ceramic mugs in different colors are placed on a wooden table, with numbers from 1 to 7 written on the cups, and a bunch of white roses on the left.
Ours
Flux
RPG
A two-tier cabinet: the top shelf has two pears, and the bottom shelf has three apples.
Ours
Flux
RPG
A beautiful landscape with a river in the middle, the left of the river is in the evening and in the winter with a big iceberg, the right of the river is in the summer with a volcano in the morning and a small village while some people are playing.

Figure 4. RAGD generation results using different seeds show more stable generation results than other methods.



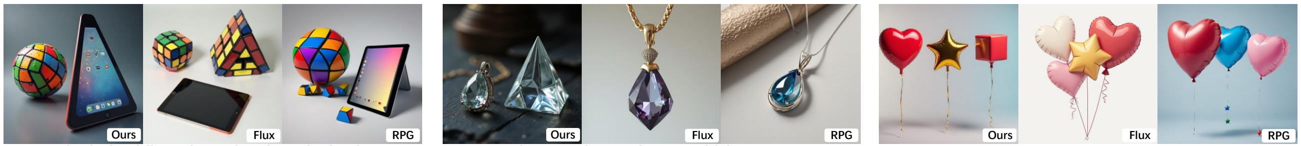
On the left, a penguin wearing sunglasses is sunbathing in the desert; in the center, a tiger wearing a scarf is standing on a glacier; on the right, a panda in a windbreaker is walking through the forest.



A blue banana, a yellow bread, and a red pineapple.

A purple sunflower, a pink newspaper, and a green cat.

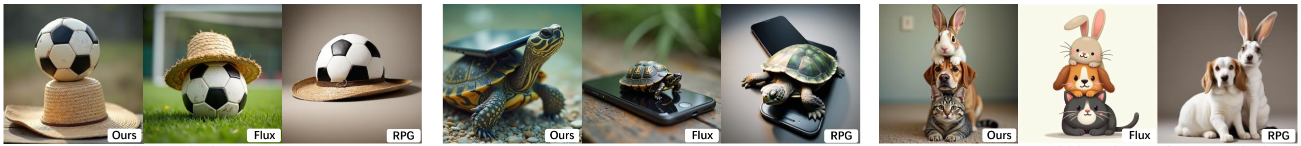
A red mushroom, a blue mushroom, and a yellow mushroom.



A spherical Rubik's cube and a triangular iPad.

A teardrop pendant and a pyramidal gemstone.

A heart-shaped balloon, a star-shaped balloon, and a square balloon.



A straw hat on the bottom of a soccer.

A turtle on the bottom of a phone.

From top to bottom, a rabbit, a dog, and a cat stacked together.

Figure 5. More qualitative results of RAGD.

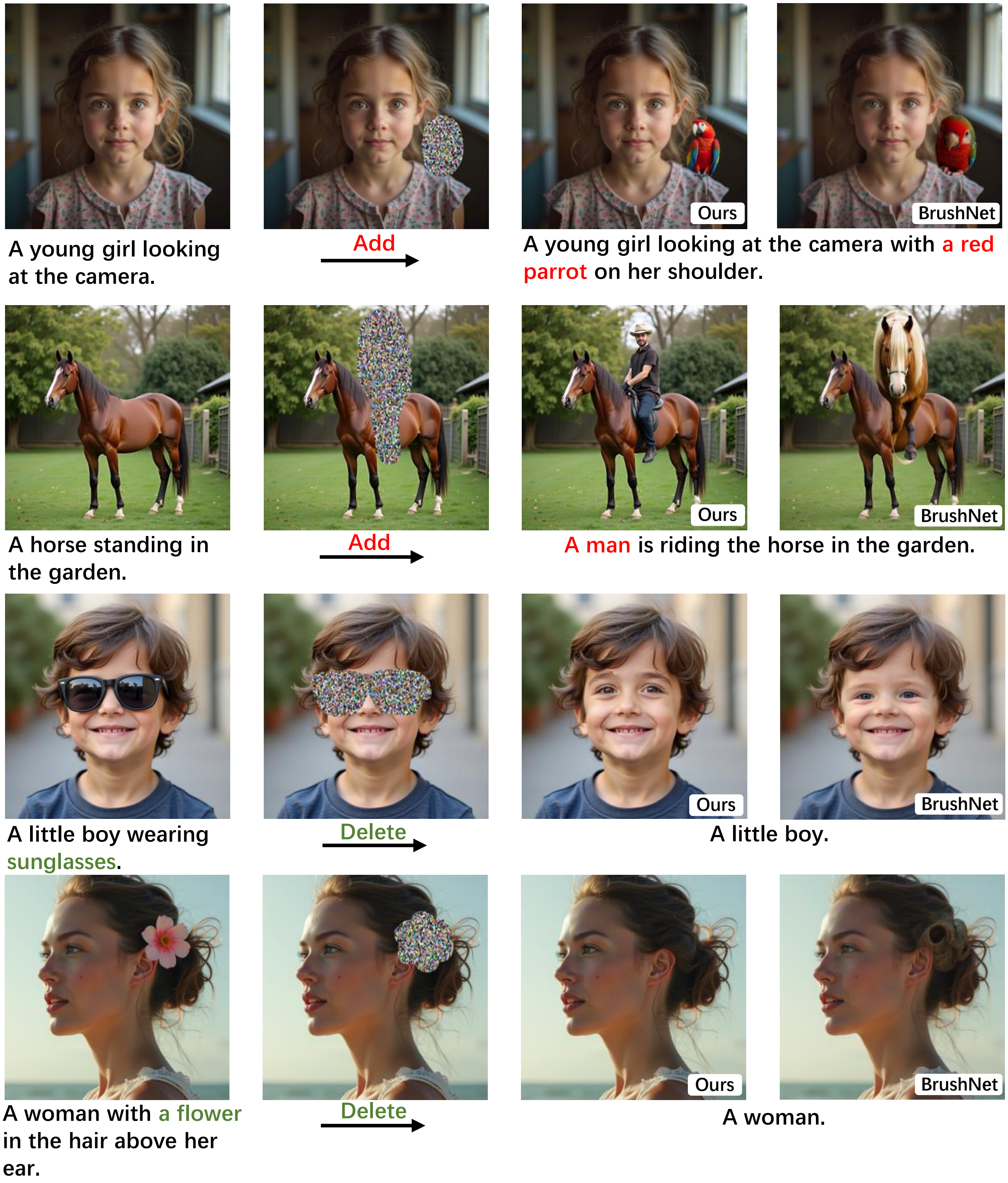

A young girl looking at the camera.
Add
Ours
BrushNet
A young girl looking at the camera with a red parrot on her shoulder.
A horse standing in the garden.
Add
Ours
BrushNet
A man is riding the horse in the garden.
A little boy wearing sunglasses.
Delete
Ours
BrushNet
A little boy.
A woman with a flower in the hair above her ear.
Delete
Ours
BrushNet
A woman.

Figure 6. The diverse applications of image repainting demonstrate the flexibility of this method.

References

[1] BlackForest. Black forest labs; frontier ai lab, 2024. 2

[2] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. *arXiv preprint arXiv:2403.06976*, 2024. 1

[3] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and CUI Bin. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*, 2024. 2