IntentEdit: Multi-Agent Reasoning for Intent-Driven Complex Image Editing

Supplementary Material

A. More Implementation Details

RISE Bench. RISEBench [2] is a recently proposed benchmark designed to evaluate reasoning-informed visual editing capabilities of generative models. Unlike prior benchmarks that mainly test superficial modifications (e.g., color changes or object replacement), RISEBench systematically focuses on four fundamental dimensions of human-like reasoning in visual editing: *Temporal*, *Causal*, *Spatial*, and *Logical* reasoning. Each dimension is carefully constructed with diverse, high-quality test cases that require models to go beyond pattern matching and demonstrate reasoning-driven visual understanding. For example, temporal reasoning evaluates whether models can capture natural progressions (e.g., fruit ripening), causal reasoning requires models to reflect effects of external forces (e.g., collisions), spatial reasoning involves structural coherence and viewpoint transformations, and logical reasoning tests puzzle-solving and rule-based inference (e.g., tic-tac-toe).

In our experiments, we strictly follow the official implementation and default settings of RISEBench, generating images and conducting evaluations exactly as specified by its authors, thereby ensuring fair and directly comparable results.

Complex-Edit Bench. Complex-Edit [1] is a complexity-controllable benchmark designed to evaluate instruction-based image editing under varying levels of compositional difficulty. Unlike benchmarks with atomic or low-level edits, Complex-Edit constructs instructions in a *chain-of-edit* manner: starting from a set of atomic editing operations (e.g., adding an object, adjusting color, changing background), GPT-4o is employed to generate sequential instructions, which are then simplified and compounded into increasingly complex editing directives. This allows fine-grained control over instruction complexity, ranging from simple single-step edits (`C1`) to highly intricate transformations with multiple dependencies (`C8`).

The benchmark evaluates models across three primary dimensions: *Instruction Following*, *Identity Preservation*, and *Perceptual Quality*. A VLM-based auto-evaluation pipeline is adopted to ensure scalable and reproducible assessment. In our implementation, we specifically focus on the subset of real images provided by Complex-Edit. We evaluate all models at complexity setting (`C5`), and each image is assessed with five independent evaluation runs to mitigate per-sample variance. This setup ensures both robustness and comparability with prior work, while also highlighting model behavior under realistic, challenging editing scenarios.

B. Samples of RISE-MIX Bench

Figure 1 presents representative examples from our RISE-MIX benchmark. Unlike existing datasets that primarily focus on either atomic edits or synthetically composed instructions, RISE-MIX emphasizes realistic scenarios where reasoning and direct edits are inherently intertwined. Each sample requires the model to not only interpret abstract reasoning cues (e.g., imagining object growth, predicting plausible temporal changes) but also to execute precise, visually grounded modifications (e.g., object replacement, background alteration). This dual nature makes RISE-MIX particularly suited for evaluating the capability of multimodal systems to bridge high-level reasoning with low-level visual editing.

C. Generation Process Visualization

To provide a more intuitive understanding of our framework, we present several qualitative examples that illustrate the full generation process. Each case demonstrates how the model interprets the instruction, executes editing steps, and then reflects the output.

As shown in Figure 2 and Figure 3, our framework first analyzes and reasons about the user intent, and then decomposes a complex instruction into multiple intermediate steps. At each step, the model performs targeted editing operations (e.g., duplicating objects, adjusting colors, or modifying spatial arrangements) and subsequently generates a reflection assessing the success of the edit. Reflections are categorized into Success, Refinement, or Replan, which guide the model to either proceed, refine the current output, or retry the step. This step-wise Plan–Edit–Reflect process enables the system to self-monitor its progress and ensures higher fidelity to the original instruction.

D. Additional Qualitative Comparison

As shown in Figures 4 and 5, we present additional qualitative results from the RISE Bench and Complex-Edit Bench. Additionally, we present a qualitative comparison of our two variants against several advanced closed-source models, including GPT-Image, Nano Banana, and SeedEdit3.

E. User Study

To complement the automatic evaluations provided by RISEBench, RISE-MIX Bench and Complex-Edit Bench, we conducted a user study to further assess the perceived quality of model outputs from a human perspective. Our

goal was to examine whether the automatic metrics align with human judgments and to provide a more comprehensive understanding of model performance in realistic usage scenarios.

Table 1. **User study.** Results of human evaluation for complex image editing. Participants assessed model outputs on three criteria: Identity Preservation (IP), Instruction Following (IF), and Visual Quality (VQ). Boldface denotes the best performance, while underlining indicates the second best.

Method	IP↑	IF↑	VQ↑	Overall↑
Step1X-Edit	3.4	3.3	3.5	3.4
ICEdit	3.6	2.9	3.5	3.3
Ovis-U1	3.7	3.9	4.1	3.9
OmniGen2	3.6	3.8	3.9	3.8
IEAP	2.1	2.5	2.3	2.3
CCA	1.9	2.1	2.6	2.2
BAGEL	3.3	4.0	3.8	3.7
BAGEL-CoT	3.5	3.9	3.1	3.5
Kontext-dev	4.0	3.5	**4.5**	4.0
Ours(InternVL3)	**4.2**	4.1	4.1	4.1
Ours(GPT-4o)	4.1	**4.4**	4.2	**4.2**

E.1. Setup

We recruited 10 participants with backgrounds spanning computer science, design, and non-technical fields. All participants reported normal or corrected-to-normal vision. Each participant was presented with pairs of input-output images generated by different models, accompanied by the corresponding editing instruction. To avoid bias, images were shuffled and anonymized, and participants were not informed of which model produced which output. Each participant evaluated 26 randomly sampled editing tasks drawn from both RISE bench, Complex-Edit bench and RISE-MIX bench. Each task were generated from a combination of our method and 9 baseline algorithms. Specifically, the tasks consisted of randomly sampled editing tasks drawn from three distinct benchmark datasets: RISE Bench, Complex-Edit Bench, and RISE-MIX Bench. Participants were asked to rate the model outputs along three dimensions using a 5-point scale:

1. **Instruction Following(IF):** How well does the output reflect the specified instruction?
2. **Identity Preservation(IP):** To what extent are irrelevant aspects of the input image preserved?
3. **Visual Quality(VQ):** How realistic, coherent, and aesthetically pleasing is the output?

In total, the study yielded a comprehensive dataset consisting of 26 tasks evaluated by 10 participants, resulting in 11 ratings per task. This amounted to a total of $26 \times 10 \times 11$ x 3 = 8580 ratings, providing a robust set of evaluations for further analysis.

E.2. Results

The user study results, as shown in Table 1, demonstrate the superior performance of our proposed system. The GPT-4o variant achieves the highest overall score of 4.2. In comparison, the InternVL3 version scores 4.1 overall, ranking second. Ovis-U1 and Kontext-dev show competitive results, with Kontext-dev achieving the best VQ score of 4.5, while Ovis-U1 ranks high in IF and VQ. BAGEL and BAGEL-CoT perform well in IF but do not surpass our models in overall performance. Baseline methods such as Step1X-Edit, ICEdit, and IEAP score significantly lower, particularly in IF and IP, highlighting their limitations in handling complex image editing tasks. These results underline the effectiveness of our approach in intent-driven complex image editing tasks.

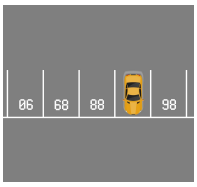
F. Limitation

Our multi-agent editing framework inevitably introduces a trade-off between multi-step capability and image quality: successive multi-turn edits may accumulate artifacts, even though we partially mitigate this with restoration. In addition, iterative planning increases inference time compared to single-pass approaches. Future work will focus on exploring acceleration techniques for diffusion models and designing diffusion architectures that better preserve image quality under multi-step editing.

References

[1] Siwei Yang, Mude Hui, Bingchen Zhao, Yuyin Zhou, Nataniel Ruiz, and Cihang Xie. Complex-edit: Cot-like instruction generation for complexity-controllable image editing benchmark. arXiv preprint arXiv:2504.13143, 2025. 1

[2] Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, et al. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. arXiv preprint arXiv:2504.02826, 2025. 1



Draw what it will look like after the car moves out from the parking slot and add a 'Reserved' sign to the parking slot where the car was previously parked. Additionally, change the color of the background to a lighter shade of gray.

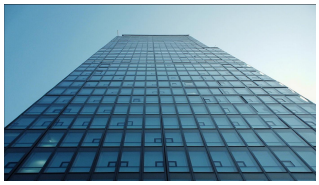
Draw what it will look like after being sewn together and add decorative stitching along all edges of the sewn fabric. Then place the sewn fabric onto a rustic wooden table as the background.

Draw what it will look like during the next king tide, add a beach umbrella in the foreground, and place a lifeguard tower near the left side of the beach.

Draw what it looks like after being violently shaken and then suddenly opened. Additionally, replace the setting with a festive party environment and add colorful confetti flying in the air.

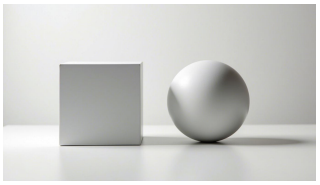
Draw what it will look like after a week in a damp basement and add scattered crumbs around the plate along with a faint watermark on the bread that says 'For Display Only.'

Draw what it will look like 50 years after abandonment, add graffiti art to some parts of the carousel, and replace the background with an overgrown forest.



Draw the front view of this building as seen from a distance, add a fountain in front of the building, and create a cloudy sky in the background.

Draw what the cutting board will look like after three months of daily use and add knife marks on its surface. Also, place a small bowl of chopped vegetables on top of it.



Draw the left view of the objects in the image, and add a label that says 'Cube' on the cube and 'Sphere' on the sphere. Additionally, change the background color to light blue.

Draw what it will look like after many people have walked on it, add a snowman near the front door, and place a wooden bench with snow piled on it under one of the pine trees.

Draw what it will look like in spring, with leaves fully grown on the tree branches, and add blooming flowers around the base of the tree along with a wooden bench placed nearby.

Draw the computer after the side cover has been removed, add a glowing LED light strip inside the case that changes colors, and place the computer on a wooden desk with a mouse pad and a wireless mouse next to it.

Draw what it will look like after being deep-fried, add a dusting of chopped parsley on top, and place it on a rustic wooden platter.

Draw what it will look like after a century of global warming, add a large abandoned signpost with 'Restricted Area' written on it near the glacier, and change the sky to an ominous orange hue.

Draw what the keyboard will look like after five years of continuous use, add a coffee stain on one corner of the keyboard, and replace the background with a wooden desk.

Draw what it will look like after it is stepped on by muddy boots and add a small potted plant in the corner near the chair. Also, place a pair of neatly arranged slippers next to the ottoman.

Draw what it will look like after being splashed with ink. Additionally, change the color of the wooden table to a dark mahogany and fold the napkin into an origami rose shape.

Generate an image assembling the provided disassembled clock components into a complete clock displaying the time as 9:45. Additionally, replace the clock's metallic frame with a wooden frame and add a small engraving on the wooden frame that reads 'Timeless Elegance'.

Draw what the building will look like in twenty years, add solar panels to the roof, and replace the tree on the left with a fountain.

Draw what it will look like immediately after one of its edges is lightly burned, add a small handwritten note in the top-left corner saying 'Fragile', and place a subtle coffee stain in the bottom-right corner.

Draw what it looked like during the Jurassic period, and add a lush prehistoric jungle as the background with tall cycads and ferns. Include a flying Pterosaur gliding through the sky above.

Figure 1. Samples from the proposed **RISE-MIX Bench**, illustrating real-world, reasoning-intensive editing instructions that combine both abstract reasoning and concrete visual modifications. Each example pairs an input image with complex instructions and the corresponding target edits, showcasing diverse reasoning types.

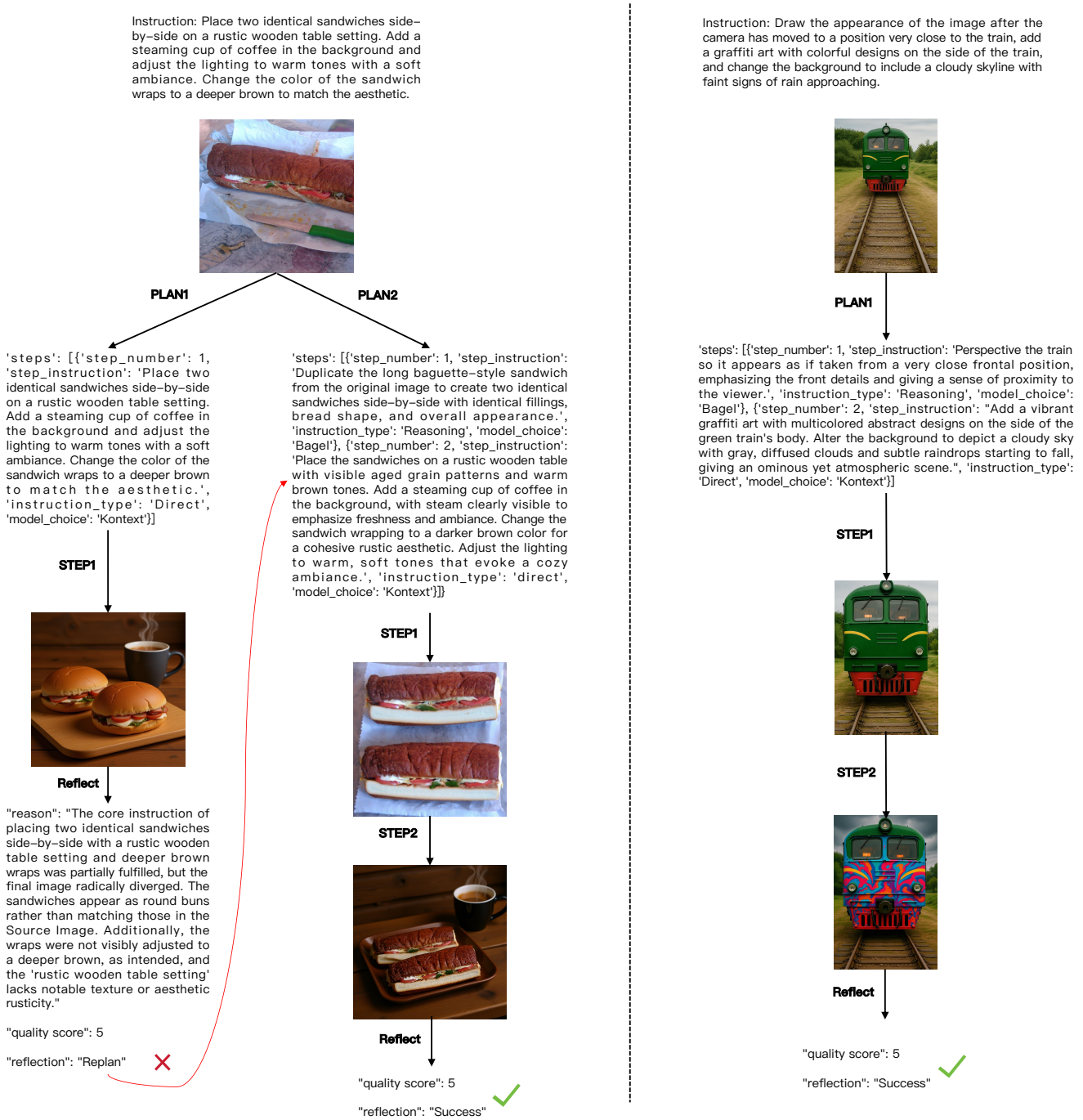


Figure 2. **Visualization of our system's planning, generation and reflection.** The model executes the planned editing instructions sequentially. Reflection feedback determines the next action, ensuring better alignment with complex user instructions.

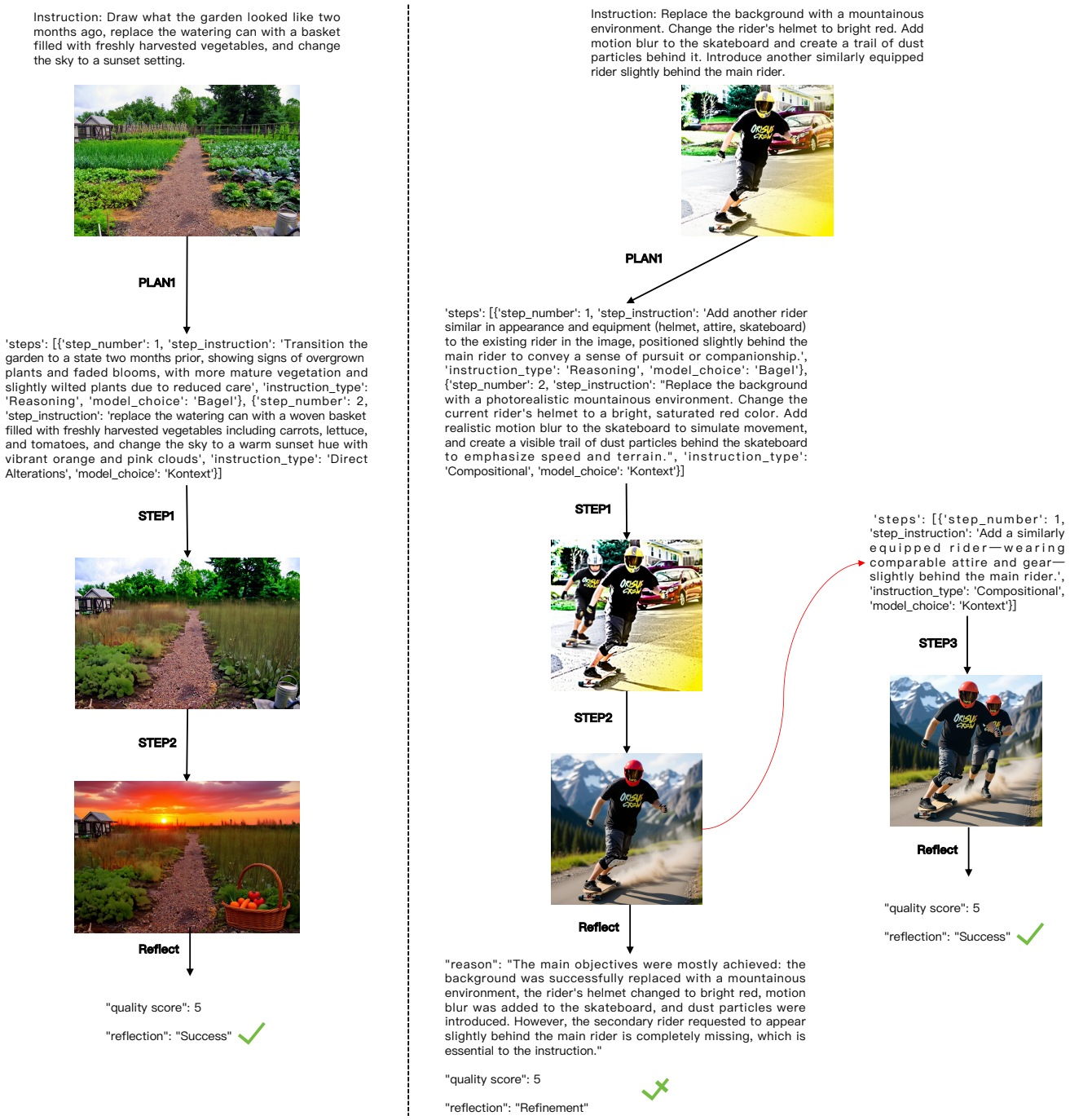


Figure 3. **Visualization of our system's planning, generation and reflection.** The model executes the planned editing instructions sequentially. Reflection feedback determines the next action, ensuring better alignment with complex user instructions.

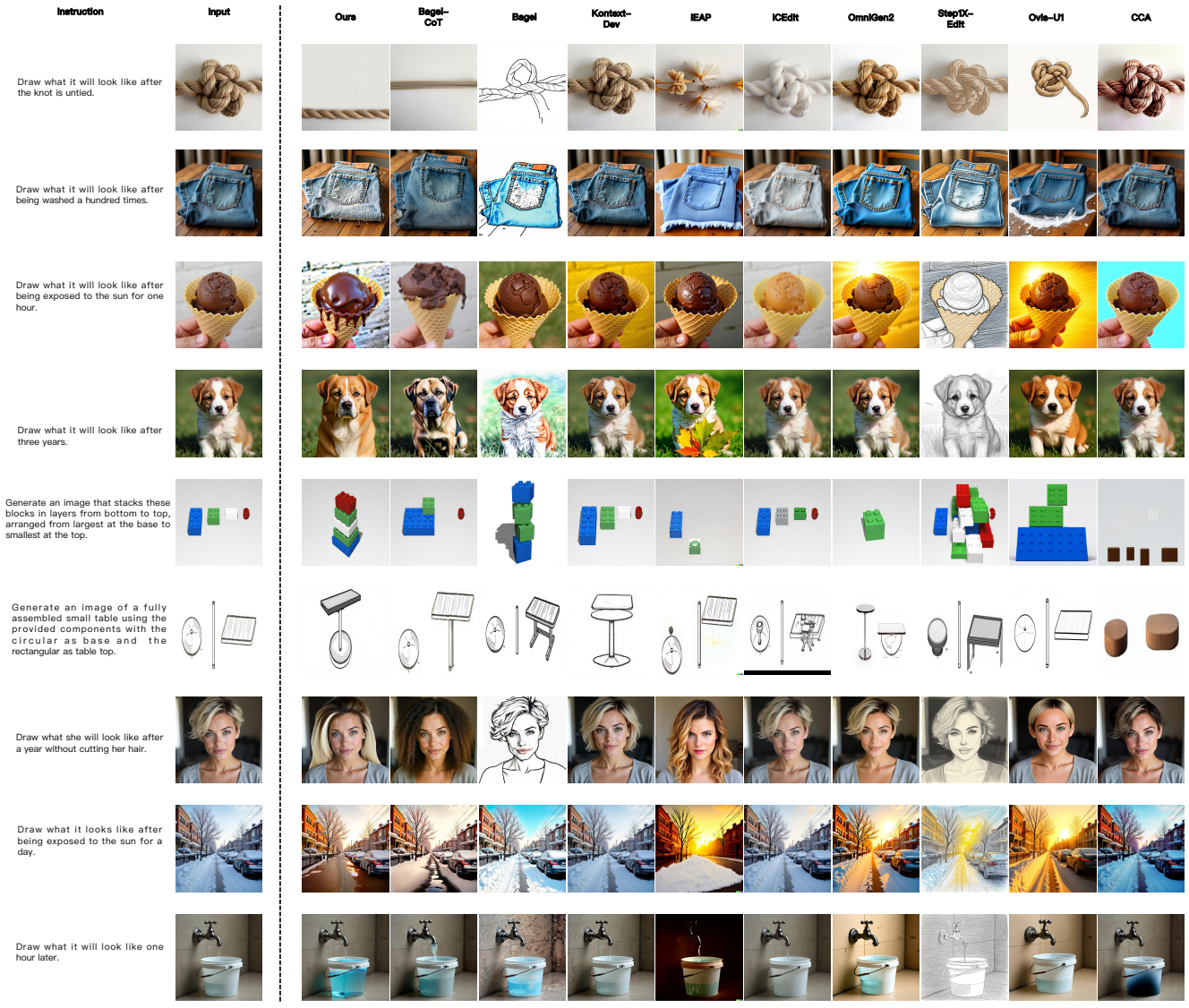


Figure 4. More visualization results on RISE Bench.

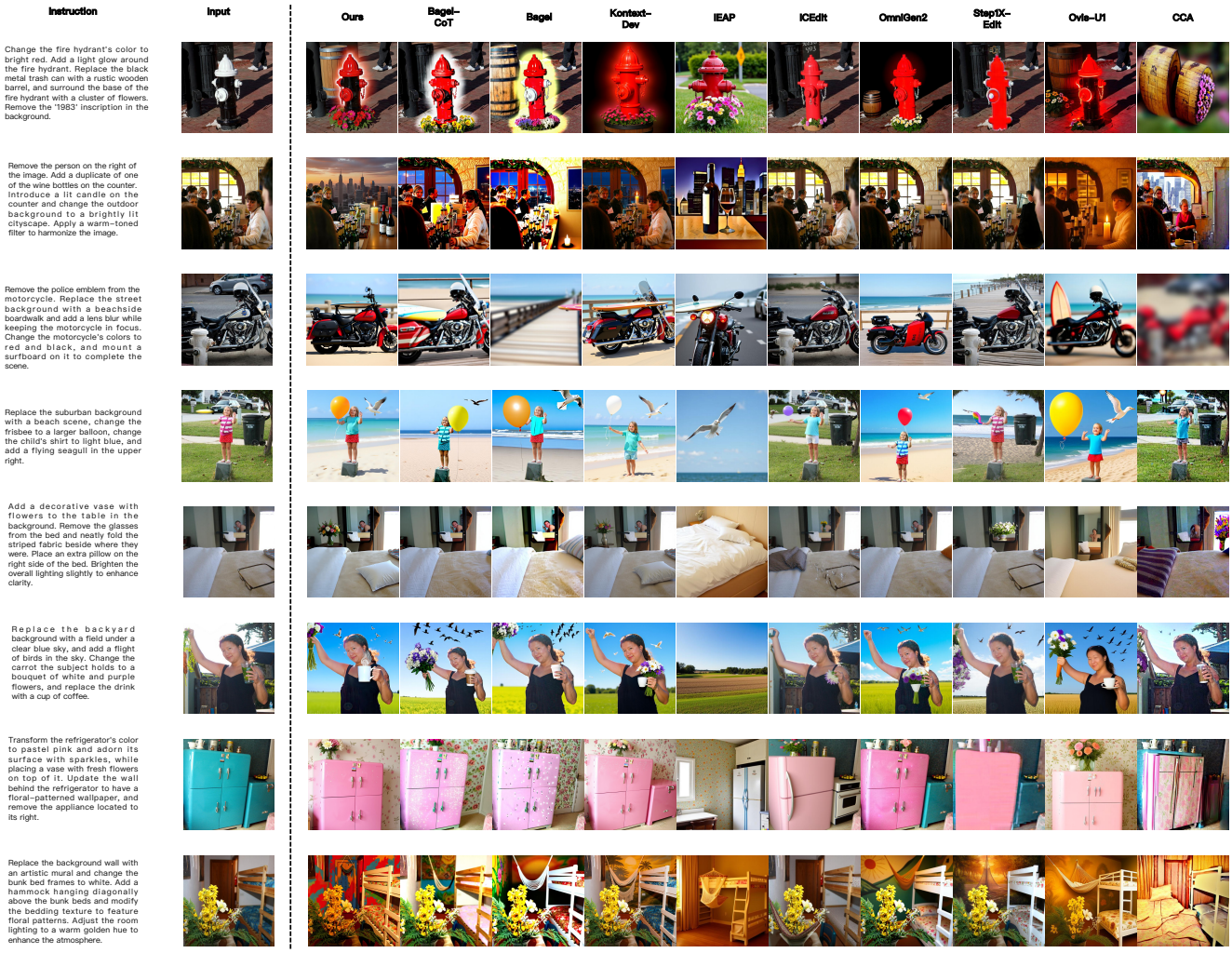


Figure 5. More visualization results on Complex-Edit Bench.

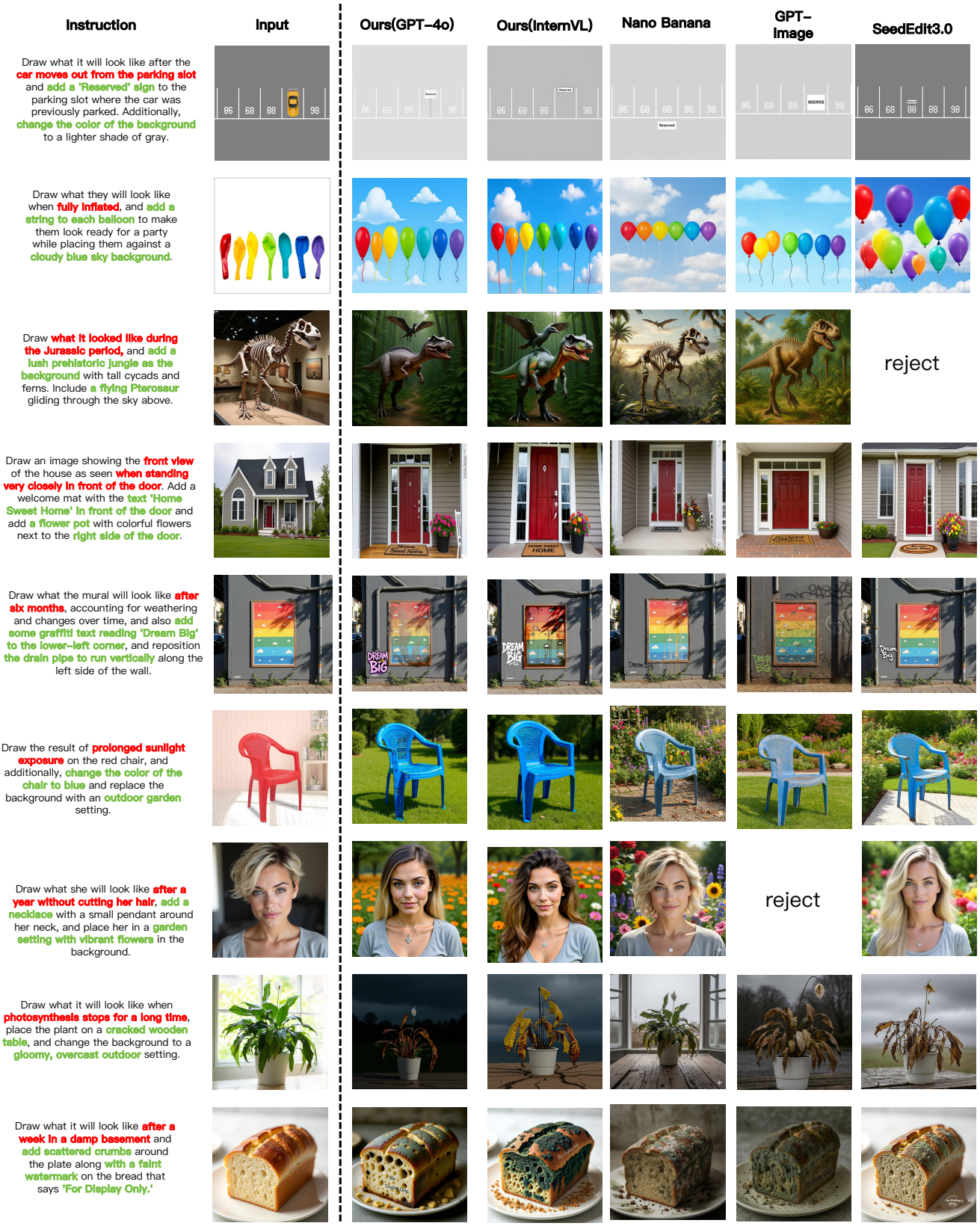


Figure 6. More qualitative comparison results of our two variants against several advanced closed-source models on RISE-MIX Bench.