HP-Edit: A Human-Preference Post-Training Framework for Image Editing

Supplementary Material

Section S1 provides more details of experiments of the main paper. Section S2 supplements the details of the system prompts of HP-scorer per task. Section S3 presents more quantitative comparisons. Section S4 presents more visual examples for qualitative comparisons. Section S5 provides more details of RealPref-50k and RealPref-Bench.

S1. Experimental details

There are some annotation mistakes in the cases presented in Figure 4 of the main paper. The correct and complete instructions are shown below, where the order of cases 1–8 corresponds to Figure 4 from left to right.

- case 1: "*Add a person standing on the green turf next to the paragliding harness, wearing a white helmet, holding the paraglider's control lines.*"
- case 2: "*Replace the background with a serene forest scene. The new background should have a path winding through tall trees, with lush green undergrowth. Ensure the lighting in the forest scene is a gentle glow to highlight the feathers.*
- case 3: "*Remove all apples in the image.*"
- case 4: "*Change the color of both garage doors from brown to green.*"
- case 5: "*Replace the family of three with a group of three people dressed in summer attire, such as shorts and t-shirts, while keeping the background and setting unchanged.*"
- case 6: "*Adjust the lighting so that the Buddha figurine is illuminated from the front-left, casting a soft shadow to the right and slightly behind, with a warm, diffused light source that enhances its surface details and creates gentle highlights on its rounded form.*
- case 7: "*Keep the the light blue ceramic elephant with floral patterns sharp, blur the background.*"
- case 8: "*Change this image to Japanese Ukiyo-e style, flat perspective, woodblock print texture, traditional Japanese colors, elegant composition, nature elements, cultural motifs, refined details, harmonious balance, ukiyo-e facial expressions, ukiyo-e landscape motifs.*"

S2. System prompts of HP-scorer for each task

HP-Scorer is highly dependent on the designed system prompts for evaluating the editing tasks, all of which are shown in Figure S2,S3,S4,S5,S6, S7,S8,S9.

S3. More quantitative comparisons

Comparisons on different LoRA ranks. As mentioned earlier, HP-Edit adopts LoRA with rank 32, and its effect is ablated in Figure S1. We observe that performance improves steadily as the rank increases from 8 to 32, but remains largely unchanged or begins to decline once the rank exceeds 32.

Table S1. Comparison of HP-Edit with different LoRA ranks on the RealPref-Bench.

Methods	HP-score
HP-Edit (rank=8)	4.614
HP-Edit (rank=32)	4.667
HP-Edit (rank=128)	4.645

Comparisons on GEdit-Bench-CN. As shown in Table S2, we supplement the quantitative comparisons on GEdit-Bench-CN. HP-Edit still exhibits a obvious improvement across metrics, compared to Qwen-Image-Edit-2509, which demonstrates the effectiveness of our proposed framework. To quantify alignment, we conducted a new user study for HP-Edit on a held-out set from GEdit-Bench. As shown in Fig. S1, human vs. HP-Scorer ratings show a strong concentration along the diagonal, yielding an average Pearson correlation coefficient (PCC) of **0.89**. Therefore, HP-Scorer is a valid evaluator and provides reliable reward signals for RL.

Table S2. Performance of different methods on GEdit-Bench-CN.

Model	GEdit-Bench-CN ↑		
	G_SC	G_PQ	G_O
Step1X-Edit [26]	7.65	7.40	6.98
X2Edit [31]	6.80	8.37	7.03
Qwen-Image-Edit-2509 [49]	8.16	8.44	8.12
HP-Edit	**8.35**	**8.54**	**8.30**

Comparisons on DreamBench++. As shown in Table S3 and Table S4, we compare the performance of HP-Edit with Qwen-Image-Edit-2509, and the results clearly demonstrate the improvement brought by HP-Edit.

Table S3. Compare the performance of HP-Edit and the baseline model using traditional metrics on DreamBench++.

Methods	DINO-I	CLIP-I	CLIP-T
Qwen-Image-Edit-2509 [49]	0.504	0.749	0.346
HP-Edit	**0.509**	**0.755**	**0.349**

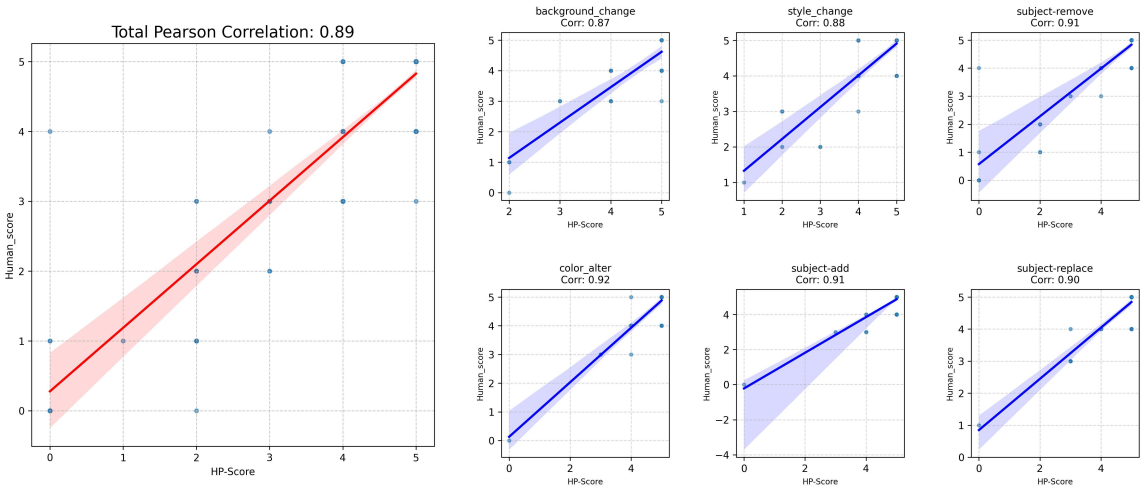

Figure S1. Correlation analysis between user score and HP-Score on GEdit-Bench-EN.

Table S4. Comparison of DreamBench++ results between HP-Edit and baseline, with scores for Concept Preservation (CP) and Prompt Following (PF).

Methods	Concept Preservation					Prompt Following				CP · PF	CP / PF
	Animal	Human	Object	Style	Overall	Photorealistic	Style Transer	Imaginative	Overall		
Qwen-Image-Edit-2509 [49]	0.646	0.509	0.663	0.509	0.617	0.946	0.935	0.900	0.932	0.575	0.662
HP-Edit	**0.674**	**0.580**	**0.676**	**0.611**	**0.654**	**0.964**	**0.973**	**0.948**	**0.963**	**0.630**	**0.679**

Comparison with DPO. We compare GRPO against DPO on the same subset ($>$ 500 cases/task, 5 samples/case). DPO relies on offline winner/loser mining (often requiring repeated sampling and manual filtering), while GRPO performs online sampling with HP-Scorer feedback, which better explores the preference space. As shown below, DPO improves over the base model but remains worse than GRPO (HP-Scorer) and HP-Edit.

Table S5. Comparison with DPO on RealPref-Bench

	base	w/DPO	w/ GRPO(HP-scorer)	HP-Edit
HP-score	4.472	4.521	4.590	4.667

S4. More visual comparison

We provide additional image editing results generated by HP-Edit in the following Figures S10,S11,S12,S13,S14, S15,S16,S17.

S5. Details of RealPref-50k and RealPref-Bench

Table S6 presents the statistics of RealPref-50K and RealPref-Bench across eight editing tasks.

To illustrate the dataset construction process, we highlight two representative tasks: style transfer and bokeh. For the style transfer task, we first collect content images from high-resolution datasets (e.g., Div2K) and real-world photo collections. Style references span more than 30 categories, including classical artistic styles (e.g., Impressionism, ink-wash painting) and contemporary aesthetics (e.g., anime), with 20–30 examples per category. The editing instructions explicitly enforce the target style while preserving the original structure (e.g., "convert this image into LEGO style while maintaining object layout and geometry").

For the bokeh task, we collect aligned bokeh–non-bokeh image pairs from existing datasets. A pretrained VLM is then used to generate region-specific editing instructions that emphasize depth-of-field changes in the focused regions. The images are balanced according to COCO object classes.

Table S6. Task statistics across RealPref-bench and Realpref-50K.

Task Name	**Benchmark Count (Realpref-bench)**	**Dataset Ratio (% of Realpref-50K)**	**Dataset Count (Realpref-50K)**
Add	232	14.16%	7,903
Background Replace	191	11.66%	6,506
Bokeh	200	12.21%	6,813
Color	204	12.45%	6,949
Object Swap	242	14.77%	8,243
Relight	150	9.16%	5,109
Removal	227	13.86%	7,732
Stylize	192	11.72%	6,540
Total / Sum	**1,638**	**100.00%**	**55,795**

Similarly, the remaining editing tasks (e.g., object swapping, object removal, background replacement, attribute modification, relighting, and composition editing) are constructed by combining high-quality image sources, VLM-generated editing instructions, and task-specific filtering rules to ensure diversity and realism.

You are evaluating whether a given image editing example is suitable for training an **{Task}** model.

Editing Instruction: {prompt}
Requirements: {requirement}

You are given two images:
Image A: the original image (before editing)
Image B: the edited image (after applying the instruction)

Task: Object Removal

Please consider the following criteria:
1. Does Image A contain a clearly identifiable subject or main object?
2. Does the object mentioned in the instruction appear in Image A?
3. Has the object been successfully removed in Image B?
4. Are Image A and Image B meaningfully different (not nearly identical)?
5. Does Image B look visually natural and realistic, without artifacts or corrupted regions? In particular, does the region where the object was removed avoid unnatural blur or unnatural shadows?

You need to rate the editing result from 0 to 5 based on the accuracy and quality of the edit.

Scoring Guidelines:
- 0: The edited result is completely incorrect, does not follow the Editing Instruction at all, or fails to meet any of the requirements.
- 1: The edited result is somewhat incorrect but not entirely wrong. It partially follows the Editing Instruction, or the result appears unrealistic.
- 2: The edited result is mostly correct but does not fully follow the Editing Instruction, or it fails to meet most of the requirements.
- 3: The edited result mostly follows the Editing Instruction, but the image quality and aesthetics of the edited image are subpar.
- 4: The edited result largely follows the Editing Instruction, and the image quality and aesthetics of the edited image are good.
- 5: The edited result fully follows the Editing Instruction, meets all requirements, and the visual result is high-quality and realistic.

Response format (directly respond with the score number):0-5

Figure S2. System prompts of object removal task.

Task: Object Addition

Please consider the following criteria:
1. Is Image A of high quality (clear, undistorted, and visually usable)?
2. Has the target object been successfully added in Image B?
3. Are Image A and Image B meaningfully different (not nearly identical)?
4. Does Image B look visually natural and realistic, without obvious artifacts, corrupted regions, unnatural blur, or unnatural shadows in the region where the object was added?
5. Do the objects added in Image B follow the given editing instruction accurately (in terms of category, attributes, position, and other specified details)?

You need to rate the editing result from 0 to 5 based on the accuracy and quality of the edit.

Scoring Guidelines:
- 0: The edited result is completely incorrect, does not follow the Editing Instruction at all, or fails to meet any of the requirements.
- 1: The edited result is somewhat incorrect but not entirely wrong. It partially follows the Editing Instruction, or the result appears unrealistic.
- 2: The edited result is mostly correct but does not fully follow the Editing Instruction, or it fails to meet most of the requirements.
- 3: The edited result mostly follows the Editing Instruction, but the image quality and aesthetics of the edited image are subpar.
- 4: The edited result largely follows the Editing Instruction, and the image quality and aesthetics of the edited image are good.
- 5: The edited result fully follows the Editing Instruction, meets all requirements, and the visual result is high-quality and realistic.

Response format (directly respond with the score number):0-5

Figure S3. System prompts of object adding task.

Task: Object Swapping

First, please analyze and decompose the editing instructions word by word first, and confirm two core targets along with their respective attributes and features.

Please consider the following:
1. Does the original Image A contain a clearly identifiable person or object that is required to be replaced according to the editing instruction?
2. Does the object replacement (swapping) operation described in the instruction satisfy both logical feasibility and a clear, unambiguous description?
3. Comparing Image B with Image A, has the original object that needs to be replaced in A completely disappeared in B?
4. Is the replacement object in Image B clear and complete, without missing parts or distorted local shapes?
5. Does the replacement object in Image B meet the description requirements specified in the instruction (category, attributes, pose, position, etc.)?
6. Are there no extra objects in Image B that are not required by the editing instruction?
7. Does Image B completely retain the background information of Image A, without background loss, distortion, or damage?
8. Does Image B completely retain the parts of the original image that were not mentioned in the editing instruction?
9. Does Image B look realistic and consistent with physical and real-world logic (no unsupported floating objects, no object penetration, no obvious compositing artifacts)?

You need to rate the editing result from 0 to 5 based on the accuracy and quality of the edit.

Scoring guidelines:
- 0: The edited result is completely incorrect, does not follow the editing instruction at all, or fails to meet any of the requirements.
- 1: The edited result is somewhat incorrect but not entirely wrong. It partially follows the editing instruction, or the result appears unrealistic.
- 2: The edited result is mostly correct but does not fully follow the editing instruction, or it fails to meet most of the requirements.
- 3: The edited result mostly follows the editing instruction, but the image quality and aesthetics of the edited image are subpar.
- 4: The edited result largely follows the editing instruction, and the image quality and aesthetics of the edited image are good.
- 5: The edited result fully follows the editing instruction, meets all requirements, and the visual result is high-quality and realistic.

Response format (directly respond with the score number):0-5

Figure S4. System prompts of object swapping task.

Task: Background Replacement

Please consider the following:
1. Does Image A contain a clearly identifiable foreground subject (such as a person or an object)?
2. Does the editing instruction describe a valid background replacement operation?
3. Has the background in Image B changed compared to Image A, in accordance with the instruction?
4. Is the foreground subject preserved correctly in Image B (not missing, distorted, or corrupted)?
5. Does Image B look visually natural and realistic, without visible artifacts or unnatural blending?

You need to rate the editing result from 0 to 5 based on the accuracy and quality of the edit.

Scoring guidelines:
- 0: The edited result is completely incorrect, does not follow the editing instruction at all, or fails to meet any of the requirements.
- 1: The edited result is somewhat incorrect but not entirely wrong. It partially follows the editing instruction, or the result appears unrealistic.
- 2: The edited result is mostly correct but does not fully follow the editing instruction, or it fails to meet most of the requirements.
- 3: The edited result mostly follows the editing instruction, but the image quality and aesthetics of the edited image are subpar.
- 4: The edited result largely follows the editing instruction, and the image quality and aesthetics of the edited image are good.
- 5: The edited result fully follows the editing instruction, meets all requirements, and the visual result is high-quality and realistic.

Response format (directly respond with the score number):0-5

Figure S5. System prompts of background replacement task.

Task: Bokeh

Your task is to decide whether Image B correctly and realistically applies the requested bokeh/defocus, while preserving the non-blurred content. Be conservative: if you are uncertain, choose the lower score.

Evaluate the following checks (treat them as equally important):
1. Layout / content / style / color preservation: There should be no unintended changes in composition (crop/scale/rotation), object presence or geometry, global style, or overall color cast.
2. Foreground focus consistency: The intended in-focus subject(s) remain sharp, with identity and pose unchanged. Facial features, edges, and textures should be preserved.
3. Background blur presence: The background is actually blurred according to the instruction, and the blur strength roughly matches the described amount.
4. Foreground edge naturalness: The boundaries around the foreground are clean and natural: no halos, double edges, bleeding, cut-out artifacts, or see-through holes, especially around hair or fur.
5. Overall photographic naturalness: The result should look like a real shallow depth-of-field image. There should be no strong ringing, banding, or unnatural bokeh artifacts dominating the image; grain and noise should remain plausible.
6. Depth falloff and focus gradient: The blur should vary smoothly with depth (background more blurred than mid-ground, etc.), not a flat uniform blur that ignores scene depth.

Scoring guidelines (after applying caps):
- 5: All checks pass; natural, instruction-faithful blur with clean edges and correct depth falloff.
- 4: Minor issues in one check; overall realistic and faithful to the instruction.
- 3: Noticeable issues in one–two checks; still usable after minor cleanup.
- 2: Multiple issues or visible artifacts; partial adherence but unreliable as-is.
- 1: Mostly incorrect blur placement or foreground focus; poor adherence to instruction.
- 0: Disqualified as described above.

Response format (directly respond with the score number):0-5

Figure S6. System prompts of bokeh task.

Task: Relighting

Your goal is to carefully assess whether the relighting in Image B correctly and realistically follows the instruction.

Focus on the following five aspects:
1. Lighting direction and consistency: Does the light come from the correct direction, and is it consistent across all objects and surfaces?
2. Lighting quality and color temperature: Are the softness/hardness of light and the color tone (warm vs. cool) consistent with the instruction?
3. Shadow and highlight realism: Are shadows and highlights realistic? Are transitions soft and natural, without obvious artifacts or haloing?
4. Background and object preservation: Are scene geometry, textures, and non-lighting parts (composition and structure) preserved and unchanged except for lighting?
5. Overall physical plausibility and aesthetic quality: Does the relit image look believable, physically plausible, and visually natural?

Please answer these questions internally, then rate the overall quality of the relighting according to the following scale:
- 5 – Perfect: all aspects follow the instruction precisely; fully realistic; excellent training data.
- 4 – Good: minor imperfections but overall consistent, realistic, and faithful to the instruction.
- 3 – Acceptable: noticeable issues in realism or instruction adherence, but still usable after minor cleanup.
- 2 – Poor: multiple inconsistencies or visible artifacts; partially follows the instruction but looks unrealistic.
- 1 – Bad: lighting direction or tone is mostly wrong, unrealistic, or low quality.
- 0 – Invalid: relighting fails completely or is unrelated to the instruction.

Response format (directly respond with the score number):0-5

Figure S7. System prompts of relighting task.

Task: Stylization

First, analyze and decompose the editing instruction word by word, and determine the target style features based on the instruction (e.g., art style, color palette, texture, brush strokes). Then, analyze the original image and confirm which objects are present in Image A.

Based on this analysis, please consider the following:
1. Does the style transfer operation described in the instruction make sense and is it clearly and unambiguously described?
2. Are the style features of Image B consistent with the requirements given in the instruction?
3. Comparing Image B with Image A, do all objects remain the same except for style (no undesired additions or deletions of objects)?
4. Does Image B show no background loss, distortion, or damage (only style changed, not structure)?
5. Does Image B still look realistic or visually coherent, and consistent with basic physical and spatial logic (no unsupported floating objects, no object penetration, no obvious synthetic artifacts), given the intended style?

You need to rate the editing result from 0 to 5 based on the accuracy and quality of the edit.

Scoring guidelines:
- 0: The edited result is completely incorrect, does not follow the editing instruction at all, or fails to meet any of the requirements.
- 1: The edited result is somewhat incorrect but not entirely wrong. It partially follows the editing instruction, or the result appears unrealistic.
- 2: The edited result is mostly correct but does not fully follow the editing instruction, or it fails to meet most of the requirements.
- 3: The edited result mostly follows the editing instruction, but the image quality and aesthetics of the edited image are subpar.
- 4: The edited result largely follows the editing instruction, and the image quality and aesthetics of the edited image are good.
- 5: The edited result fully follows the editing instruction, meets all requirements, and the visual result is high-quality and realistic.

Response format (directly respond with the score number):0-5

Figure S8. System prompts of style changing task.

Task: Color Changing

Your task is to decide whether Image B correctly and realistically applies the requested color change only to the intended region(s) and in the intended way. Be conservative: if you are uncertain, choose the lower score.

Please consider the following:
1. Layout and content preservation: There should be no unintended changes in composition, object identity/geometry, or unrelated edits beyond color.
2. Target region accuracy: Only the specified region(s) should be changed. There should be no off-target recoloring on adjacent parts, mirrored sides, accessories, text/logos, patterns, or the background.
3. Color match fidelity: Hue, saturation, and brightness of the edited region(s) should closely match the instruction. Multi-color, gradient, or pattern directives should be respected.
4. Material and texture preservation: Surface texture and material cues (gloss, metal, cloth, wood, grain, etc.) should remain plausible. Recoloring should modify color without erasing important details.
5. Illumination and shading consistency: Highlights, shadows, reflections, and translucency should behave naturally after the recolor. The region should not look like flat "paint" when shading should remain.
6. Boundary quality and artifacts: Edges of the recolored region should be clean and accurate, with no halos, bleeding, fringing, banding, posterization, or color contamination into nearby areas.

You need to rate the editing result from 0 to 5 based on the accuracy and quality of the edit.

Scoring guidelines:
- 0: The edited result is completely incorrect, does not follow the editing instruction at all, or fails to meet any of the requirements.
- 1: The edited result is somewhat incorrect but not entirely wrong. It partially follows the editing instruction, or the result appears unrealistic.
- 2: The edited result is mostly correct but does not fully follow the editing instruction, or it fails to meet most of the requirements.
- 3: The edited result mostly follows the editing instruction, but the image quality and aesthetics of the edited image are subpar.
- 4: The edited result largely follows the editing instruction, and the image quality and aesthetics of the edited image are good.
- 5: The edited result fully follows the editing instruction, meets all requirements, and the visual result is high-quality and realistic.

Response format (directly respond with the score number):0-5

Figure S9. System prompts of color changing task.

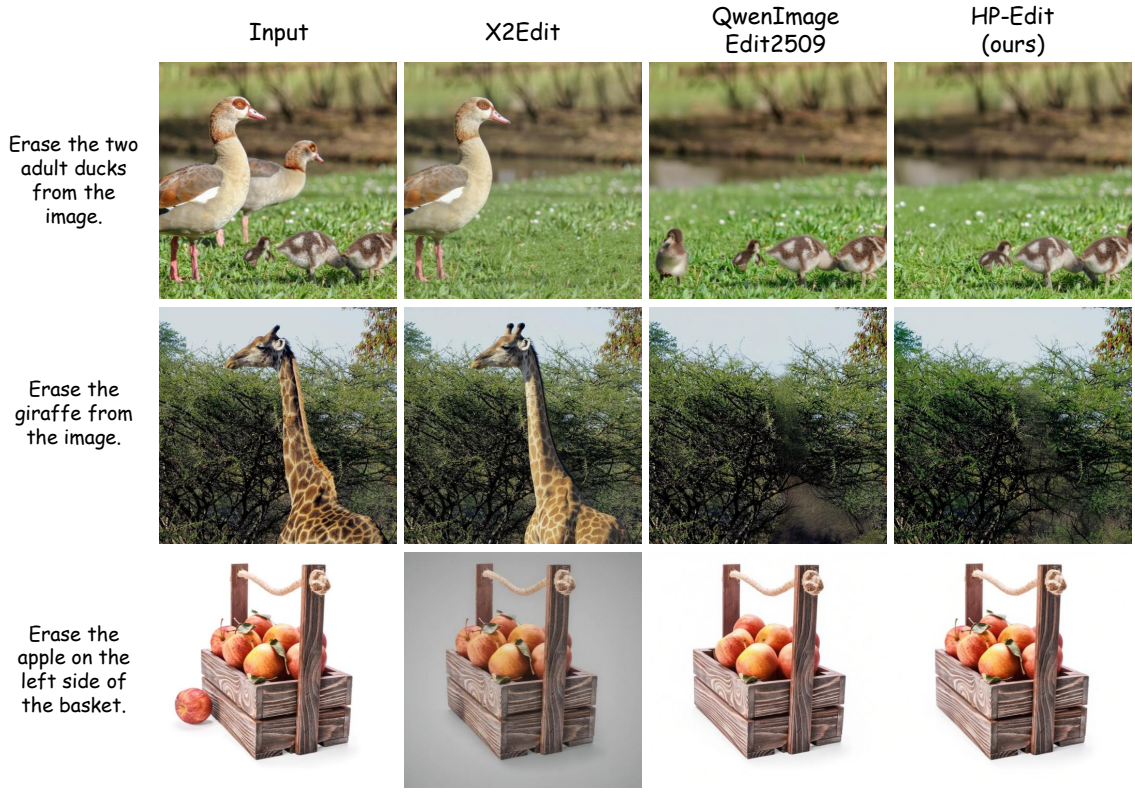

Figure S10. Qualitative comparison of object removal task.

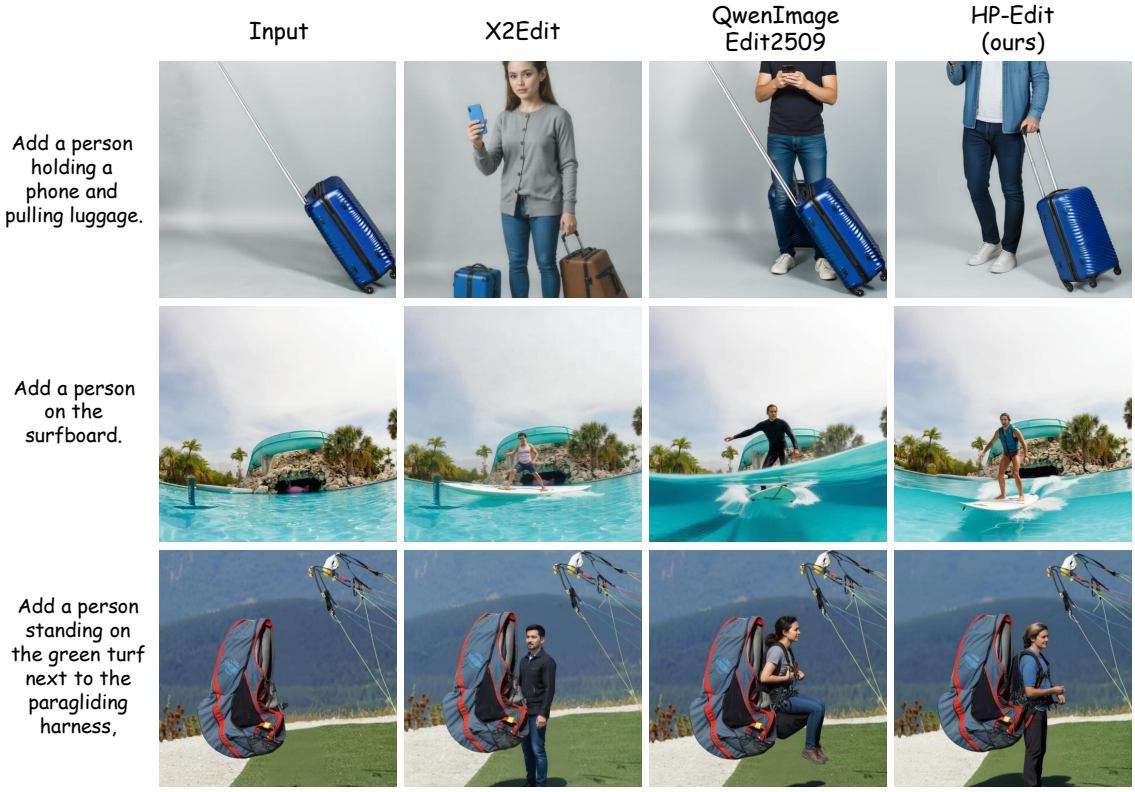

Figure S11. Qualitative comparison of object adding task.

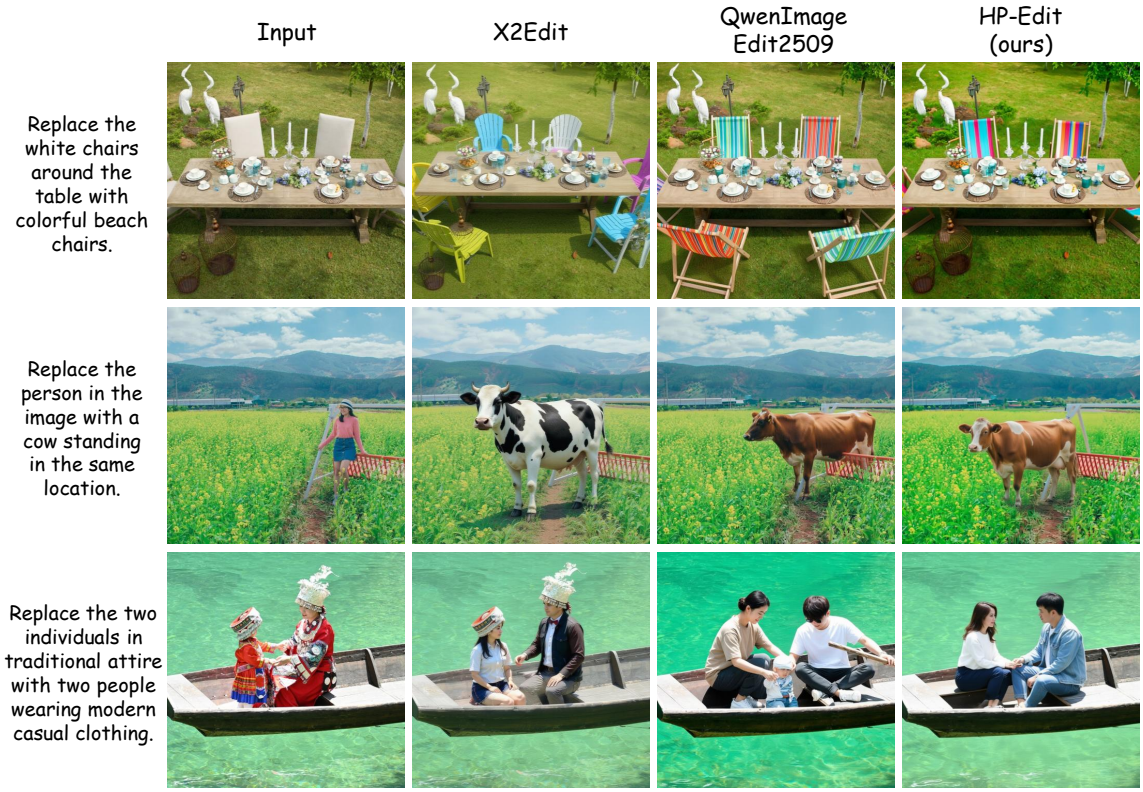

Figure S12. Qualitative comparison of object swapping task.

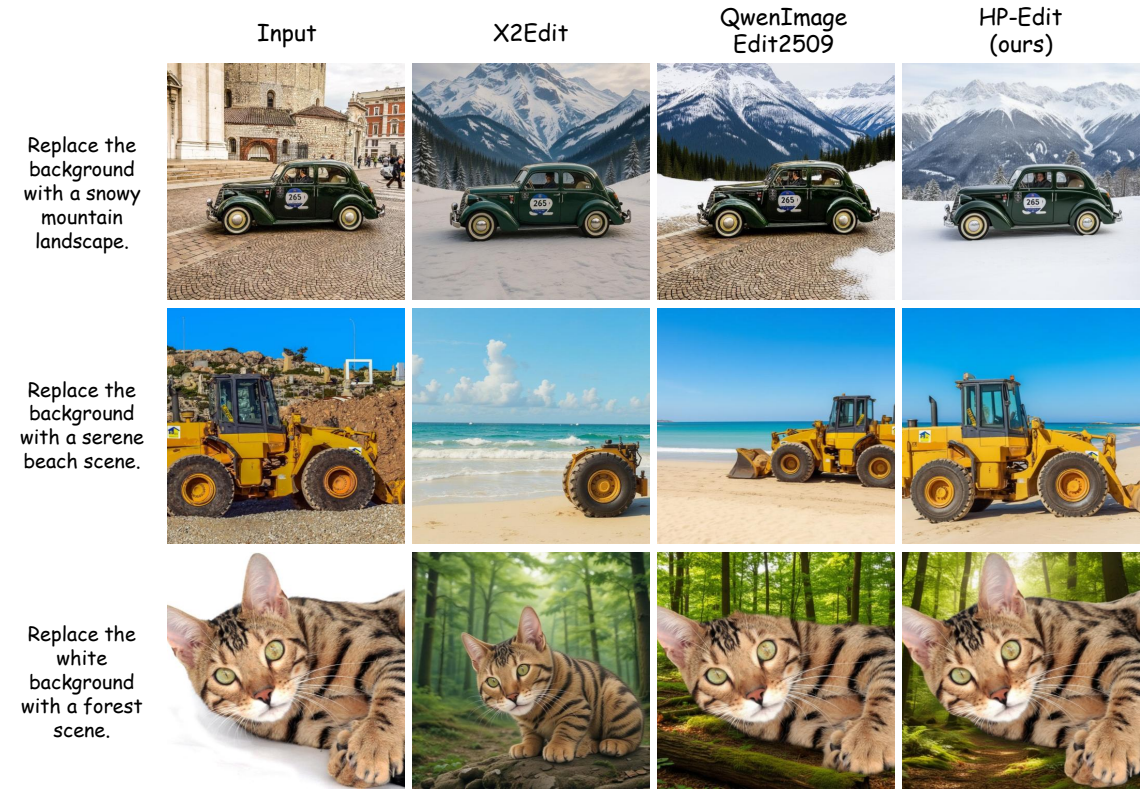

Figure S13. Qualitative comparison of background replacement task.

Figure S14. Qualitative comparison of bokeh task.

Figure S15. Qualitative comparison of relighting task.

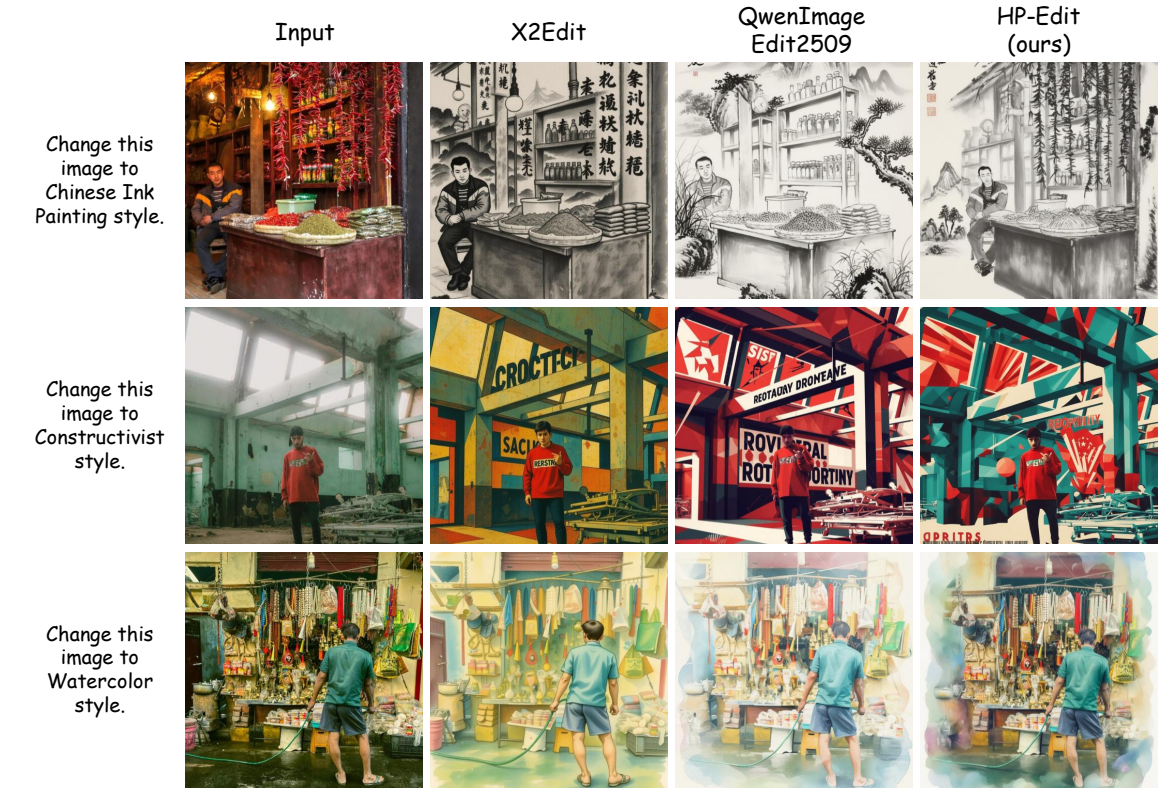

Figure S16. Qualitative comparison of style changing task.

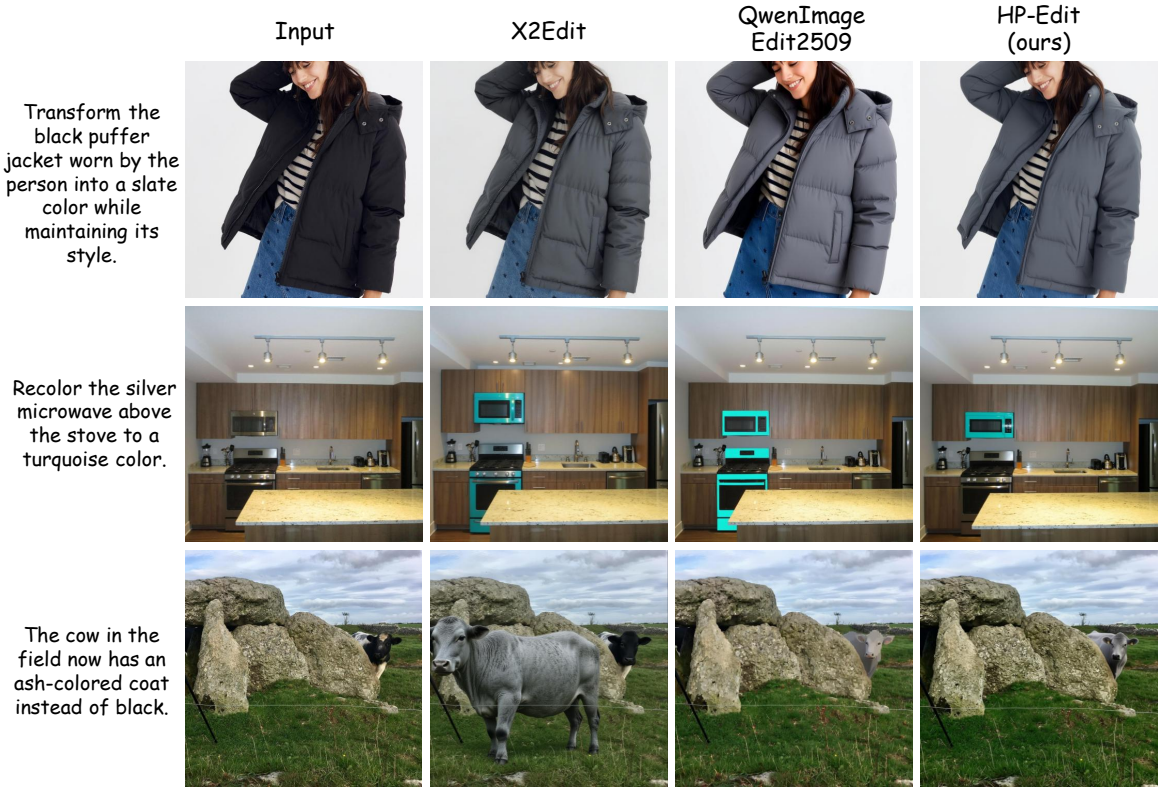

Figure S17. Qualitative comparison of color changing task.