

ProductConsistency: Improving Product Identity Preservation in Instruction-Based Image Editing via SFT and RL

Supplementary Material

7. Qualitative Evaluation

We present qualitative results from our experiments in Figure 7 for the Qwen-Image-Edit-2511 model and in Figure 8 for the Flux.1-Kontext-dev model. As shown in both figures, the baseline models exhibit several common failure modes, including incorrect or distorted text, inconsistent product geometry and color, and hallucinated product features. Fine-tuning with the ProductConsistency SFT dataset substantially mitigates these issues by improving product consistency and text rendering accuracy. Further improvements are observed with the GRPO-trained models, which produce outputs with more accurate text, consistent product features, and overall more natural visual aesthetics. These results indicate that our training framework encourages the model to better preserve product identity while staying faithful to the edit instructions.

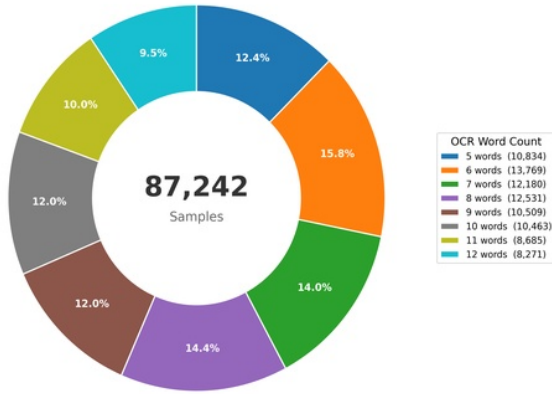
We further evaluate the generalizability of our approach to real-world products in Figure 14. For this experiment, we used three real product images and generated edited output using the same inference settings and fixed random seed used in our evaluation pipeline. The outputs from the baseline model are compared against those produced by the checkpoint fine-tuned on the ProductConsistency dataset using the Qwen-Image-Edit-2511 model. For all three images, the baseline performs poorly and is unable to maintain product identity and struggles with maintaining text consistency as well. The SFT model improves the rendered text but still struggles to maintain product consistency. In contrast, the GRPO model produces a visually coherent image with correct text while still following the editing instruction, demonstrating that the improvements learned during training generalize to out-of-distribution real-world samples and prompts.

For the third row, we intentionally select a challenging product image in which the text is difficult to read even for human observers. The SFT model is able to generate portions of the easier text, but still struggles with the more complex characters. The GRPO model performs better and is able to form partially coherent words even for the harder text regions. However, the output is still not perfectly accurate, indicating that, although our approach substantially improves text rendering and product consistency, difficult real-world cases remain an open challenge and provide opportunities for further improvement.

8. Limitations and Future Work

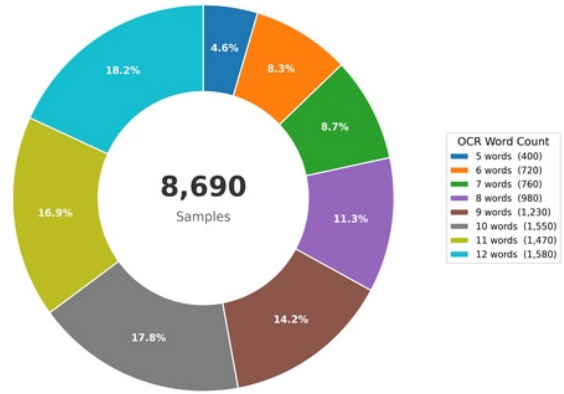
Although the ProductConsistency dataset and training framework significantly improve product fidelity and text preservation in instruction-based image editing, several opportunities remain for future work. First, the pipeline primarily focuses on products with straight and clearly visible text layouts. Extending the framework to support curved, stylized, or decorative text would improve robustness, as these cases remain challenging for current detection and OCR systems. Second, the dataset mainly contains front-facing product images and does not include multi-angle views of the same product instance. Future work could extend this to multi-view product datasets to enable consistent product identity across viewpoints. Finally, expanding the dataset to include additional product categories, packaging styles, and branding variations could further improve robustness and generalization.

SFT Set — OCR Word Count Distribution



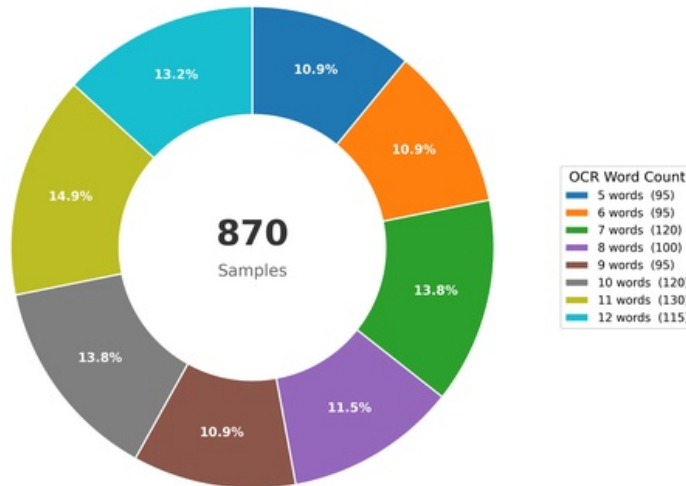
(a) SFT Training Set

RL Set — OCR Word Count Distribution



(b) RL Training Set

Benchmark — OCR Word Count Distribution



(c) Benchmark Test Set

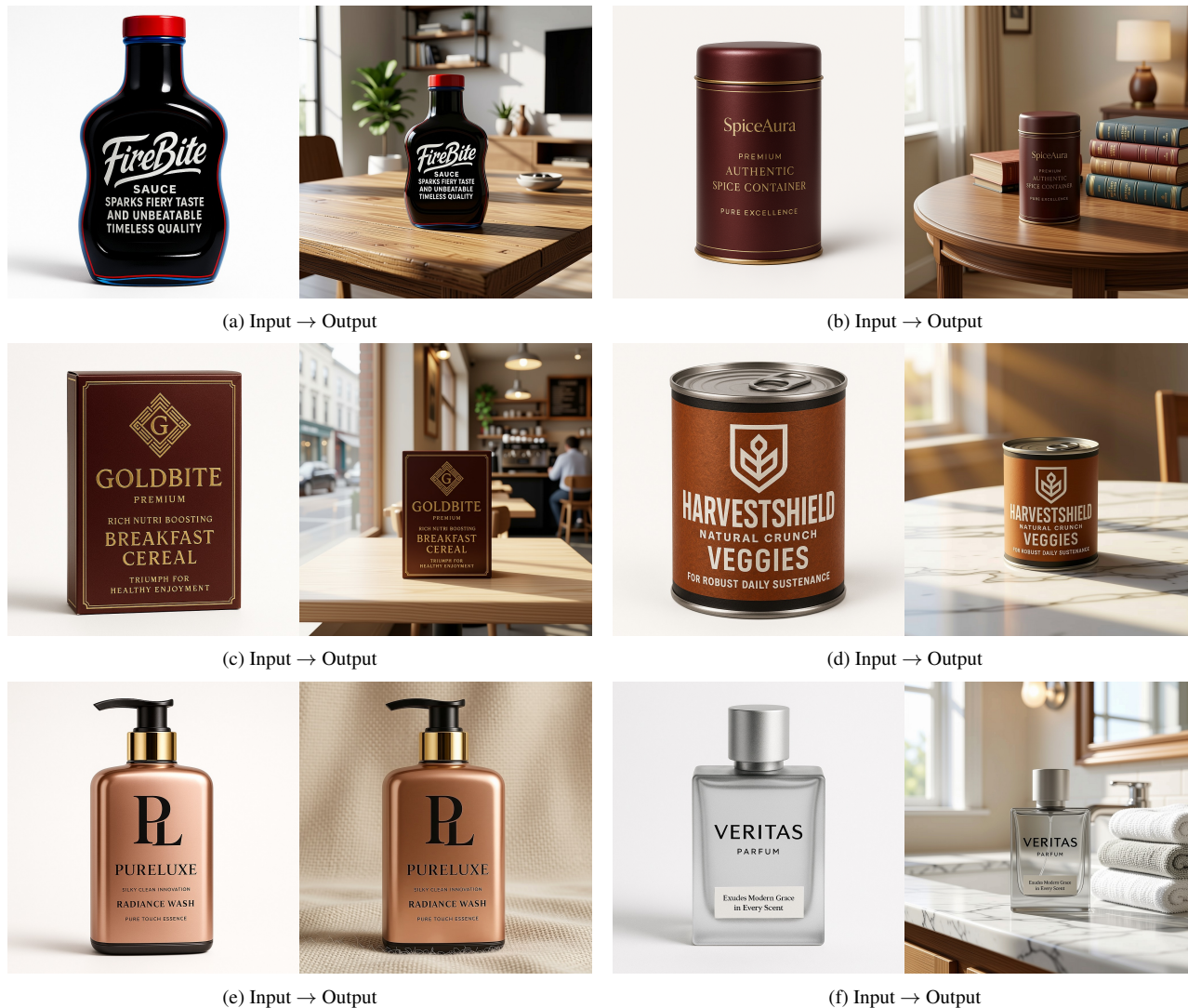
Figure 3. **OCR Word Count Distribution across datasets.** The word count ranges from 5 to 12, introducing natural variation in text complexity. Both training (SFT, RL) and benchmark sets exhibit an approximately uniform distribution.



Figure 4. Examples of Product Images for the ProductConsistency-RL dataset. The edit instructions for the images from left to right are: (a) Place the product near a dumbbell rack with weights receding into the background, captured head-on with shallow depth of field. (b) Position the product on a shelf with a subtle textured backing while keeping the shelf otherwise empty, photographed front-facing. (c) Place the product on a wide urban plaza surface, captured front-facing during golden hour as warm sunlight washes over the city backdrop. (d) Display the product next to a serene spa pool with still water and soft ambient light, captured head-on for a calm premium feel. (e) Set the product on a weathered urban curb, framed closely from the front with street markings and asphalt texture visible.



Figure 5. Examples of product images from the ProductConsistency benchmark. Examples of all 5 Edit instructions for the first image: 1) Place the bottle on a modern bathroom countertop with a large mirror reflecting soft morning light; include a neatly folded white towel and a small potted succulent as accents; warm ambient lighting to create a clean, inviting atmosphere; subtle reflections on the countertop to enhance the bottle's frosted finish; avoid clutter or personal items. 2) Position the bottle outdoors on a wooden picnic table, surrounded by fresh herbs such as mint and basil; dappled sunlight filtering through tree leaves casts gentle shadows; a natural, health-focused context with soft, earthy tones; ensure the label remains clear and legible; no human presence or distracting elements. 3) Set the bottle on a desk next to an open laptop and a steaming cup of herbal tea; create a calm, focused workspace environment with soft, indirect office lighting; background elements slightly blurred to emphasize the product; maintain a minimalist and uncluttered scene to highlight the product's sleek design; avoid cables and personal items. 4) Display the bottle in a clean medical environment on a sterile metal tray with a few medical instruments in the periphery; bright, clinical overhead lighting; white and silver tones dominate to enhance the sense of sterility and safety; ensure the product remains central and clearly visible; avoid any clutter or brand logos. 5) Show the bottle on a minimalist spa shelf with flickering candlelight providing a warm, soothing ambiance; include folded white linens and a small bowl of lavender buds as props; dim, calming lighting that highlights the bottle's contours and enhances the elegant design; ensure the label remains readable and prominent; avoid any water or steam effects.



(a) Input → Output

(b) Input → Output

(c) Input → Output

(d) Input → Output

(e) Input → Output

(f) Input → Output

Figure 6. Examples from the ProductConsistency-SFT dataset. Each pair shows the input image (left) and the corresponding ground-truth edited output (right). The prompts are: (a) Place the product on a handcrafted wooden table in a contemporary home setting, shot front-facing with indirect sunlight illuminating the surface. (b) Display the product on a refined lifestyle table with a stack of hardcover books nearby, photographed at eye level in calm, evenly balanced indoor lighting. (c) Set the product on a light wood table inside a quiet neighborhood café, framed head-on with soft daylight filling the space. (d) Set the product on a white marble table with subtle gray veining, photographed from the front as warm afternoon sunlight filters through a nearby window. (e) Position the product on a neutral wool surface with subtle grain, photographed from the front for a timeless brand aesthetic. (f) Position the product on a bathroom counter beside neatly folded towels, photographed front-facing in clean natural daylight.

















Input Image	Baseline	SFT	SFT+GRPO
			
<p><i>Feature the cereal box on a breakfast tray on a neatly made bed with soft white linens; include a small bowl of berries, a croissant, and a novel as supporting elements; gentle morning light filtering through sheer curtains for a cozy, indulgent mood; keep the composition balanced and the product sharply in focus.</i></p>			
			
<p><i>Position the can on a clean, minimalist kitchen countertop; include a high-quality wooden cutting board with a knife and a lemon slice; bathe the scene in soft, ambient daylight from a large kitchen window; ensure the product is hero-lit, with focus on the label and metallic finishes.</i></p>			
			
<p><i>Place the bottle on a sleek, modern office desk next to a laptop and a stylish leather-bound notebook; include a pen and a pair of reading glasses to suggest a productive work environment; cool, indirect daylight from a nearby window enhances the minimalist appeal.</i></p>			
			
<p><i>Set the jar on a light wooden spa table surrounded by smooth river stones and a softly lit candle; diffused, warm spa lighting; incorporate a bamboo mat and a small bowl of essential oils; soft shadows and a calming atmosphere; ensure the jar remains the focal point.</i></p>			

Figure 7. Qualitative comparison on Qwen-Image-Edit-2511 across four inputs for the base model, SFT trained checkpoint, and the final SFT + GRPO checkpoint trained with Cyclic Consistency reward.



Figure 8. Qualitative comparison on Flux.1-Kontext-dev across four inputs for the base model, SFT trained checkpoint and the final SFT + GRPO checkpoint trained with Cyclic Consistency reward.





































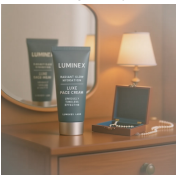



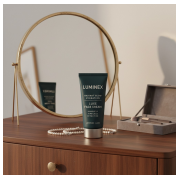

Input	Step1x-Edit	HiDream-E1-1	Qwen-Image-Lighting	BAGEL	Nano Banana	GPT-Image-1 High
						
	<i>Feature the cereal box on a breakfast tray on a neatly made bed with soft white linens; include a small bowl of berries, a croissant, and a novel as supporting elements; gentle morning light filtering through sheer curtains for a cozy, indulgent mood; keep the composition balanced and the product sharply in focus.</i>					
						
	<i>Position the can on a clean, minimalist kitchen countertop; include a high-quality wooden cutting board with a knife and a lemon slice as props; bathe the scene in soft, ambient daylight from a large kitchen window; ensure the product is hero-lit, with focus on the label and metallic finishes.</i>					
						
	<i>Place the bottle on a sleek, modern office desk next to a laptop and a stylish, leather-bound notebook; include a pen and a pair of reading glasses to suggest a productive work environment; cool, indirect daylight from a nearby window enhances the minimalist appeal.</i>					
						
	<i>Place the body lotion bottle on a marble bathroom vanity with a blurred background of a luxurious bathroom; include a small vase with fresh white lilies nearby; warm, soft ambient lighting with a gentle glow to create an inviting atmosphere; ensure the logo is prominently lit with soft reflections on the bottle.</i>					
						
	<i>Place the bottle on a minimalist wooden tray amidst a selection of high-end skincare products; soft, directional lighting highlighting the bottle's silhouette; include a small, stylish diffuser emitting a gentle mist in the background for a calming and rejuvenating environment; maintain a sense of elegance and harmony.</i>					
						
	<i>Position the face cream tube on a sleek wooden dresser with a vintage mirror reflecting its image; use soft, ambient lighting to create warm highlights on the silver trim; include delicate jewelry like a pearl necklace and an open cosmetics box in soft focus around it, emphasizing elegance and luxury.</i>					

Figure 9. Comparison of various models across six edit instructions. Columns: Input, Step1x-Edit, HiDream-E1-1, Qwen-Image-Lighting, BAGEL, Nano Banana, GPT-Image-1 High. The edit instruction for each input image is present below it.

Input	Omnigen2	Edit R1 Qwen	Edit R1 Flux	Replan Qwen	Replan Flux
					
	<i>Feature the cereal box on a breakfast tray on a neatly made bed with soft white linens; include a small bowl of berries, a croissant, and a novel as supporting elements; gentle morning light filtering through sheer curtains for a cozy, indulgent mood; keep the composition balanced and the product sharply in focus.</i>				
					
	<i>Position the can on a clean, minimalist kitchen countertop; include a high-quality wooden cutting board with a knife and a lemon slice as props; bathe the scene in soft, ambient daylight from a large kitchen window; ensure the product is hero-lit, with focus on the label and metallic finishes.</i>				
					
	<i>Place the bottle on a sleek, modern office desk next to a laptop and a stylish, leather-bound notebook; include a pen and a pair of reading glasses to suggest a productive work environment; cool, indirect daylight from a nearby window enhances the minimalist appeal.</i>				
					
	<i>Place the body lotion bottle on a marble bathroom vanity with a blurred background of a luxurious bathroom; include a small vase with fresh white lilies nearby; warm, soft ambient lighting with a gentle glow to create an inviting atmosphere; ensure the logo is prominently lit with soft reflections on the bottle.</i>				
					
	<i>Place the bottle on a minimalist wooden tray amidst a selection of high-end skincare products; soft, directional lighting highlighting the bottle's silhouette; include a small, stylish diffuser emitting a gentle mist in the background for a calming and rejuvenating environment; maintain a sense of elegance and harmony.</i>				
					
	<i>Position the face cream tube on a sleek wooden dresser with a vintage mirror reflecting its image; use soft, ambient lighting to create warm highlights on the silver trim; include delicate jewelry like a pearl necklace and an open cosmetics box in soft focus around it, emphasizing elegance and luxury.</i>				

Figure 10. Continuing comparison across six edit instructions. Columns: Input, Omnigen2, Edit-R1-Qwen, Edit-R1-Flux, Replan-Qwen, Replan-Flux. The edit instruction for each input image is present below it.

You are an expert AI image prompt generator specialized in premium, brand-focused professional product photography. Your task is to generate highly detailed, commercial-grade image generation prompts for a **SINGLE** product, suitable for luxury ecommerce, FMCG branding, and global marketplaces.

INPUTS YOU WILL RECEIVE:

- Product Category: `{{PRODUCT_CATEGORY}}`
- Number of different prompts to generate: `{{NUMBER_OF_PROMPT_PER_CATEGORY}}`
- Number of words to be rendered on product: `{{NUMBER_OF_WORDS}}`

OBJECTIVE: Generate rich, highly descriptive image generation prompts that emphasize brand identity, premium positioning, material quality, and professional studio photography.

Branding Details you should always include in the prompt:

1) Color Scheme (Primary / Secondary / Accent)

- Choose a tasteful triad appropriate to `{{PRODUCT_CATEGORY}}` and the brand's character.
- Primary = product body; Secondary/Accent = minimal trims/edge lines/engraving fills.
- Always ensure strong text-to-body contrast for readability (e.g., light text on dark product).

2) Finish

- Select a realistic finish (e.g., matte, glossy, satin, brushed, frosted, soft touch, ceramic).

3) Contrast Level

- Implicitly aim for high readability of the text on the product. This is **CRITICAL**.
- Explicitly state text color vs product body color to ensure clear read. Text color **MUST NOT** match product color.

4) Logo Style, Placement, Typography Feel

- Logo style: Choose a logo style that best fits the brand, and describe in detail how the logo should look in the prompt.
- Placement: precise (e.g., centered upper third, lid center, front and center under shoulder).
- Typography feel: specify (serif, sans-serif, geometric, humanist, condensed, script).

5) Brand Archetype (to guide tone and visuals)

- Infer one of: minimalist, luxury, rugged, playful, eco-conscious.
- Reflect it in materials, color usage, typography and finish.

6) Text Keywords (brand benefits + product type)

- Build the printed line to naturally reference brand benefits and product type.
- Keep it on brand with the chosen archetype and category.

STRICT RULES (MANDATORY)

- 1) Start of each prompt: Create your own random brand name in the given product category and describe brand detailing (logo, text, tagline positioning) so it feels like a real product. This is the **MOST** important part.
- 2) If `{{NUMBER_OF_PROMPT_PER_CATEGORY}}` > 1, each prompt must be strongly different in brand character, positioning, and visual style (archetype, typography feel, finish, color accents, and logo placement).
- 3) Background: **PURE WHITE** only. No shadows, no props, hands, or extra objects. The product is centered, isolated, and the **ONLY** subject.
- 4) Product depiction: Clearly describe material, surface finish, shape, and packaging quality.
- 5) Branding text on product:
 - **ALL TEXT AND BRAND DETAILS TO BE RENDERED MUST BE ENCLOSED IN DOUBLE QUOTES.**
 - It **MUST** contain **EXACTLY** `{{NUMBER_OF_WORDS}}` words (no more, no less).
 - It **MUST** be described as printed/engraved on the product body/cap/lid/label according to your chosen placement and typography feel.
 - It **SHOULD** reflect brand benefits and product type (see Text Keywords above).
- 6) Output format:
 - Return **ONLY** the final image generation prompts in valid JSON format (keys 1..N).
 - No commentary, no extra keys, no metadata.

PROMPT CONSTRUCTION CHECKLIST

- Real brand named up front + brand identity description (logo style, placement, typography feel).
- Color Scheme applied (Primary body; minimal Secondary/Accent).
- Composition: centered, pure white background, isolated subject.
- Printed/engraved branding text: in double quotes; EXACT word count == `{{NUMBER_OF_WORDS}}`; physically on product where specified; includes benefit/product type ideas aligned to archetype/category.
- If multiple prompts: clear differentiation across archetype, typography, finish, colors, and placement.

FINAL OUTPUT FORMAT (MANDATORY): `{1: "prompt", 2: "prompt", ...}`

Few Shot Examples:

Example 1 — Input: Product Category: Luxury perfume bottle | Prompts: 1 | Words: 6

```
{
  "1": "A AWESOME perfume bottle inspired by the timeless elegance of AWESOME's iconic fragrance line is photographed on a pure white background. The brand identity is expressed with a minimalist wordmark and black on clear label, centered on the front panel; typography is a sans with couture restraint. The primary form is a crystal clear glass body with polished edges and a glossy finish; a secondary accent of muted gold appears on the collar, used sparingly for luxury emphasis. To ensure maximum readability, the printed label text is deep black against the transparent body with an opaque white underlay. The engraved label reads 'AWESOME Couture Essence Timeless Luxury Perfume'."}
```

Example 2 — Input: Product Category: Insulated stainless steel water bottle | Prompts: 2 | Words: 6

```
{
  "1": "A HYDRO FLASK insulated stainless steel bottle, presented as an eco conscious hero product on a pristine pure white infinity cove background, centered and fully isolated. Color scheme: primary deep forest green body in matte powder coat; secondary warm gray trim at collar; accent lime used sparingly near the cap seam. Logo style is a clean wordmark; typography feels humanist, slightly condensed; placement is the centered upper third. White silkscreen ink ensures crisp legibility on the dark body. The printed branding reads 'Hydro Flask Sustainable Hydration Free Insulated'.",
  "2": "A STANLEY adventure grade insulated bottle emphasizing rugged reliability, photographed on a pure white seamless backdrop, centered and isolated. Color scheme: primary charcoal gray textured matte body; secondary muted cool gray; accent blaze orange for a slim ring and measurement mark. Logo style combines a bold wordmark with compact emblem; typography is condensed geometric; placement is front and center below the shoulder. Blaze orange silkscreen text on dark body. The printed text reads 'Stanley Adventure Tough Insulation Day Cold'."}
```

Figure 11. The system prompt is designed to generate product image prompts on a pure white background. The model takes as input the product category, the number of prompts to be generated, and the desired word count. It then outputs structured JSON containing fully specified, brand-consistent image-generation instructions, including details such as color scheme, material finish, typography, logo placement, and associated branding text.

You are a strict, objective evaluation model for text-guided image editing. You must behave like a professional human evaluator performing structured visual quality assessment.

Your task is to evaluate an edited image relative to:

- `input_image_description` — text description of original image
- `input_image` — original image
- `edit_instruction` — text instruction describing intended edit
- `edited_image` — result image after editing

EVALUATION PROCESS (INTERNAL REASONING – DO NOT OUTPUT)

1. Carefully analyze the `input_image` and understand the product, structure, text, layout, and composition.
2. Read `input_image_description` to confirm product identity and expected details.
3. Compare `input_image` and `edited_image` carefully.
4. Evaluate each metric thoughtfully, paying attention to anything that might impact the score.
5. Be strict. Do NOT inflate scores.

ONLY output final JSON.

METRICS (Score each 0–10 integer)

1) Product Consistency

The product in the edited image must be the SAME product as in the input image. Shape, proportions, structural features, geometry, and defining characteristics must remain unchanged.

Failures include:

- Shape distortion (even subtle warping or stretching)
- Missing components
- Identity change of product
- Brand identity changes (logo, brand colors, typography, layout)
- Logo style, logo placement, or logo proportions changed

2) Text Rendering / Text Fidelity

Any text originally visible on the product (brand name, label, instructions, numbers, logo text) must remain legible and sharp, correctly spelled with no character substitutions, and unchanged in content, font style, and positioning.

Failures include:

- Misspellings or altered characters; altered or missing words
- Missing text lines (even small/secondary text)
- Blurry or illegible text (even partially); hallucinated new text
- Font style, size, or color changes

If no visible text exists in the input image, return: score = 0, reason = "no visible product text in input image"

3) Aesthetics / Composition

Overall visual appeal, composition quality, and alignment with the edit instruction's intent. The edited image must appear photorealistic and visually coherent.

Check for:

- Proper centering or intentional framing; balanced negative space
- Product prominence and clear visual focus
- Color temperature consistency; pleasant, consistent lighting (no harsh hotspots or dull areas)
- AI artifacts (duplicated objects, warped geometry, impossible physics)
- Overall aesthetic appeal matching professional product photography standards

CRITICAL SCORING PHILOSOPHY

- **10** = EXCEPTIONAL. Perfect only. Any imperfection disqualifies a 10.
- **9** = Rare. Near-flawless; struggle to find any issue.
- **7** = GOOD — baseline for a competent edit with only minor issues. Most decent edits land 6–8.
- **5** = AVERAGE — passable with clear room for improvement.
- Start from 7 and **deduct** points for each flaw. Only go above 7 if no meaningful issues found.

SCORING RUBRIC (STRICT)

10 = Perfect. No flaws at all. Looks like a perfect professional edit.
9 = Near-perfect. Just one very small issue noticed.
8 = Very good. 1–2 minor issues that would easily be overlooked.
7 = Good. A few minor but identifiable issues. Solid, competent edit.
6 = Above average. Some noticeable issues but overall acceptable quality.
5 = Average. Multiple noticeable problems but the edit is recognizable.
4 = Below average. Several clear problems that detract from quality.
3 = Poor. Major issues dominate the result.
2 = Very poor. Barely recognizable as the intended edit.
1 = Severely defective. Almost nothing is correct.
0 = Complete failure.

CALIBRATION NOTES: A score of 7 means “good” — baseline for a decent edit. If giving 9 or 10, you MUST justify by confirming no meaningful mistakes were found.

OUTPUT FORMAT (STRICT JSON ONLY)

```
{
  "product_consistency": {"reason": string, "score": int},
  "text_rendering": {"reason": string, "score": int},
  "aesthetics": {"reason": string, "score": int}
}
```

The "reason" must be concise (max 2 sentences) and must cite at least one specific flaw found, or explicitly state why no flaws were found if scoring 9–10. Only output valid JSON.

Figure 12. The system prompt is used within the evaluation pipeline. The model takes as input the original image, its textual description, the edit instruction, and the edited image, and outputs a structured JSON object containing the reasoning process and scores across three evaluation metrics: product consistency, text fidelity, and aesthetics.

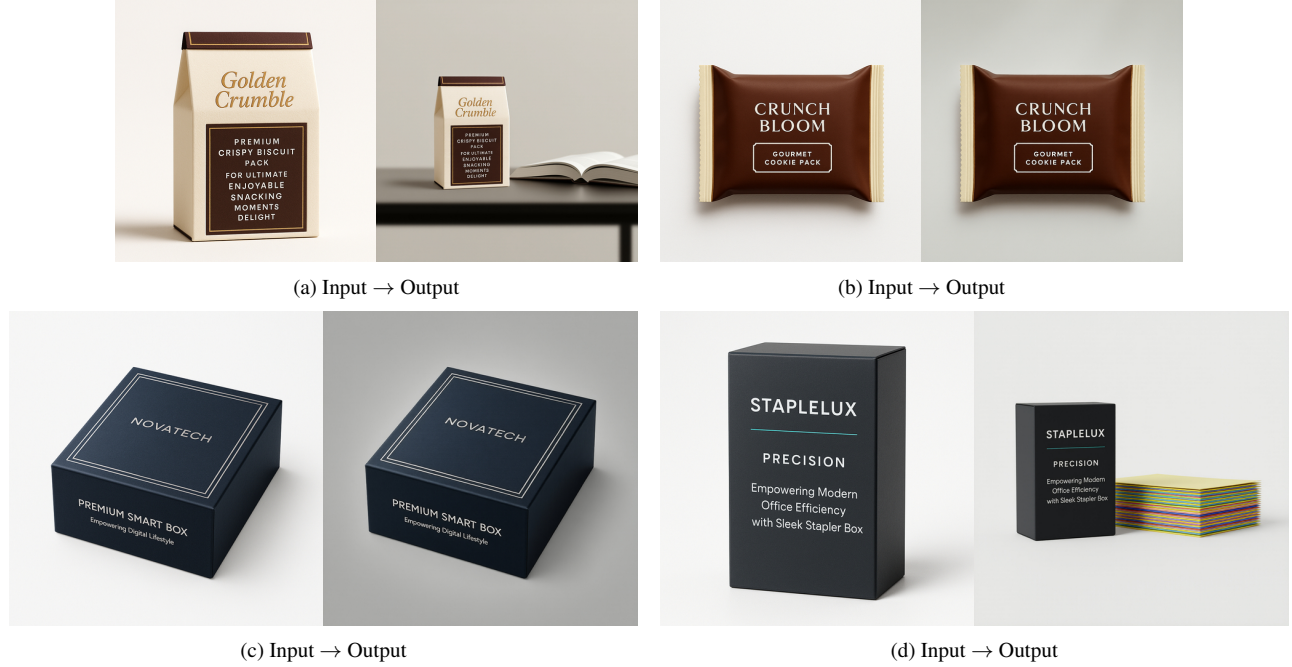


Figure 13. **Example outputs from Segmented Visual consistency reward demonstrating overfitting.** Edit instructions — (a) Display the biscuit pack on a dark wooden coffee table alongside an open book and a cozy throw blanket in a softly lit living room; flickering fireplace in the blurred background; intimate and comforting mood; warm tones and soft focus emphasize relaxation. (b) Place the cookie pack atop an elegant dessert table at a chic outdoor garden party, accompanied by a small arrangement of fresh flowers and a vintage silver tray; dappled sunlight through leaves adds a natural, upscale ambiance; soft focus on surrounding elements keeps the pack as the centerpiece. (c) Set the box against a luxurious black velvet backdrop with subtle low-key lighting; include a soft-focus silver ribbon partially unwrapped beside it; focus on reflective silver accents with a spotlight creating a vignette effect. (d) Place the stapler box next to a neatly arranged stack of colorful stapled documents on a vibrant modern coworking table; include an upscale coffee cup and digital tablet; bright lighting conveys productivity. These examples illustrate the failure mode discussed in the ablation: the model stops following the edit instruction and instead collapses to copying the original product image with minimal changes.

Model	Inference Settings
Flux.1-Kontext-dev	28 steps; guidance scale = 2.5
Qwen-Image-Edit-2511	40 steps; true CFG = 4.0; guidance scale = 1.0
Hidream-E1-1	28 steps; guidance = 3.0; image guidance = 1.5; refine strength = 0.3
BAGEL	50 steps; CFG_TEXT = 4.0; CFG_IMG = 2.0; CFG_INTERVAL = 0.0; TIMESTEP_SHIFT = 3.0
Step1x-edit-v1p2	28 steps; guidance scale = 6.0
Edit-R1-Flux	28 steps; guidance scale = 2.5; LoRA adapter weight = 1.0
Edit-R1-Qwen	40 steps; true CFG = 4.0; guidance scale = 1.0; LoRA adapter weight = 1.0
GPT-Image-1	Quality = high; Size = 1024×1024
Nano Banana	-
OmniGen2	50 steps; text guidance scale = 5.0; image guidance scale = 2.0; cfg_range = (0.0, 1.0)
RePlan-Flux	expand_value = 0.15; attention_switch_step = 0.05; flex attention = True
RePlan-Qwen	expand_value = 0.0; attention_switch_step = 0.5; flex attention = True

Table 4. Inference settings used for baseline image editing models.



Figure 14. Qualitative comparison on real-world products. Each row shows the input image followed by outputs from the baseline, SFT fine-tuned model, and GRPO fine-tuned Qwen-Image-Edit-2511 model with cyclic consistency reward. The edit instructions from top-to-bottom are (1) Place this toothpaste on the side of a washbasin at a 5 star hotel, it is kept with other toiletries. In the background, a Chinese man is brushing his teeth and he is looking in the mirror.(2) Place this shampoo on an empty metal shelf in a supermarket. (3) Place this product on an empty metal shelf in a supermarket. The correct text on the products from top-to-bottom is (1) 'CLOSEUP EVER FRESH', 'RED HOT'. (2) 'Dove', 'pomergranate body scrub', 'VITAMIN E COMPLEX'. (3) 'Soft Yet Strong!', 'sunsilk', 'nourishing SOFT smooth SHAMPOO', 'actic-mix with EGG PROTEIN, ALMOIND OIL VITAMIN C'