CREval: An Automated Interpretable Evaluation for Creative Image Manipulation under Complex Instructions

Supplementary Material

In this supplementary material, we provide additional analysis and experimental results to further support the main paper. The content is organized as follows: Sec. A analyzes the rationale behind the weight selection for the final score. Sec. B presents additional quantitative results and extended experimental analysis; Sec. C provides examples of Question-Answer pairs in CREval; Sec. D presents detailed prompt templates used in our evaluation framework; and Sec. E offers more visual comparisons across the state-of-the-art methods mentioned in the main paper.

A. Weight sensitivity analysis.

As shown in Fig. S1, MLLMs perform suboptimally when evaluating VQ. Therefore, we reduce the weight of VQ. In contrast, IF and VC are more important and discriminative, so we assign them higher weights. Fig. S2 shows experiments with different weight settings, where the 4:4:2 ratio achieves better alignment with human preferences.

B. More Experimental Details

To clarify the implementation logic of the proposed CREval evaluation framework, the pseudocode of the core evaluation method is presented in Table S2.

Table S1 presents a more detailed analysis than Table 2 in Section 4.2, showing the detailed scores for IF, VC, and VQ across nine creative dimensions. This allows for clearer comparisons of different models across all the editing tasks discussed in the main paper.

For the task **Customization**, most models are able to maintain stable editing capabilities on *Derivative Character*. These tasks usually have clear structural constraints and relatively limited modification requirements, so the performance gap between open-source and closed-source models remains small. On the other hand, the *Reimagined Representations* and *Surreal Fantasy* tasks involve structural changes or are highly abstract, and many models struggle to maintain key elements of the source image after editing.

Similar issues also exist in the identity-related modifications tasks, such as the *Identity & Cultural Transformation* dimensions in **Stylization** category, where most models struggle to follow the editing instructions or to preserve essential visual features that should remain unchanged.

In **Contextualization**, especially *Informationization & Narrative Expression* and *Commercial design*, which involve rich contextual information in narrative, these tasks place higher demands on semantic understanding and generative flexibility. It can be observed that different models exhibit noticeable differences in processing and generating rich contextual information.

C. VQA Examples

In Sec. 3.3, we introduce CREval, a MLLM-based VQA automatic evaluation. In this section, we provide some examples of QA pairs. For all questions listed in Table S3, the ideal answer for perfect editing is 'Yes'.

D. Prompt Templates

In the following, we list the prompt templates used in our experiments.

Prompt template for building dataset. After manually selecting high-quality images, we provide some examples of corresponding categories, and then use GPT-4o to generate corresponding editing instructions. The template for the prompt is shown in Figure S3.

Prompt templates for VQA generation CREval utilizes MLLM to generate evaluation question-answer pairs. As illustrated in Fig. S4, S5 and S6, the VQA generated for each metric (IF, VC or VQ) is associated with a specific prompt. All prompts are designed using the Chain-of-Thought (CoT) reasoning scheme to generate structured evaluation question-answer pairs.

Prompt template for evaluation. The prompt template for answer generation is shown in Figure S7.

E. More Visual Comparisons

In Figure S8, we present additional visual comparisons. The full instructions are listed as follows, in order from left to right.

- case 1: "Transform the colossal robot into a miniature 3D articulated model encased in a sleek, circular display case. Add intricate, tiny gears visible under transparent panels on the robot's surface for mechanical depth. Position the display on a futuristic hexagonal black base, etching the robot's model number in a luminous silver font. Surround with a subtly detailed mini landscape evoking the expansive original scene."
- case 2: "Reimagine the bridal scene as a Renaissance portrait, with the central figure as a regal noblewoman in a velvet gown adorned with intricate lacework and pearls, carrying a bouquet of rich, dark roses. The bridesmaids, in brocade dresses with gold embroidery, hold vintage floral arrangements. The setting becomes an elegant arched

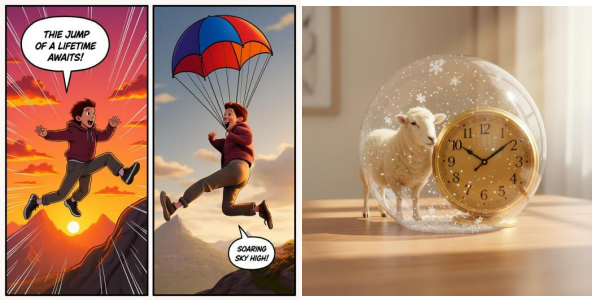

Figure S1. VQ Failure Cases.

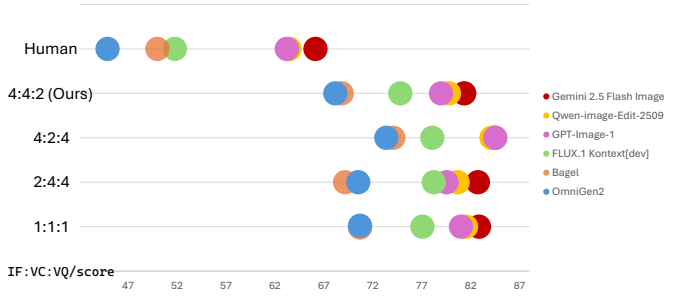

Figure S2. Weight Ratios.

Table S1. More quantitative comparisons on CREval-Bench by GPT-4o. The best performance among closed-source models is highlighted in red, and the second best in blue. For open-source models, the top result is shown in **bold**, and the second best is underlined.

Category	Dimension	Metric	Open-source Models									Closed-Source Models			
			OmniGen2 [49]	Bagel [7]	Bagel [7] (think)	UniWorld-V1 [24]	ICEdit [59]	Qwen-Image-Edit [48]	FLUX.1 Kontext[dev] [2]	Step1X-Edit-v1p2 [27]	Step1X-Edit-v1p2 [27] (think)	GPT-Image-1 [53]	Seedream4.0 [38]	Gemini 2.5 Flash Image [10]	FLUX.1 Kontext[pro] [2]
Customization	Derivative Character	IF	79.36	80.76	74.45	61.40	51.11	**85.91**	76.07	78.03	76.82	88.37	90.21	82.68	75.40
		VC	61.23	61.78	69.01	**79.58**	51.41	73.83	74.44	62.49	62.92	74.15	75.64	76.87	75.58
		VQ	85.59	84.02	80.64	74.57	64.88	89.33	**90.91**	85.03	80.07	89.15	90.07	89.12	91.27
		avg	73.35	73.82	73.51	71.31	53.98	**81.76**	78.39	73.21	71.91	82.84	84.35	81.64	78.65
	Reimagined Representations	IF	65.35	71.43	67.28	46.79	40.13	**76.34**	56.12	61.73	65.45	76.28	80.01	72.51	58.47
		VC	44.12	51.33	64.21	**78.50**	62.92	70.38	75.55	59.09	59.72	56.49	74.89	73.31	73.22
		VQ	86.61	82.70	82.41	85.65	80.80	**91.51**	87.64	88.36	87.33	87.20	92.29	93.35	92.43
		avg	61.11	65.64	69.08	67.25	57.38	**76.99**	70.20	66.00	67.53	70.55	80.42	77.00	71.16
	Surreal Fantasy	IF	60.76	76.30	63.13	27.72	31.45	**87.39**	61.28	77.37	76.61	91.06	88.27	83.97	68.61
		VC	53.18	43.56	66.62	**85.37**	70.93	56.67	64.00	46.84	45.55	44.88	66.48	69.56	54.89
		VQ	78.74	73.29	68.54	58.89	58.65	**89.46**	78.82	78.72	75.31	91.08	89.59	87.83	82.58
		avg	61.32	62.60	65.61	57.01	52.68	**75.52**	65.88	65.43	63.93	72.59	79.82	78.98	65.92
	Average	IF	68.49	76.16	68.28	45.30	40.90	**83.21**	64.49	72.38	72.96	85.24	86.16	79.72	67.49
		VC	52.84	52.22	66.61	**81.15**	61.75	66.96	71.33	56.14	56.06	58.50	72.34	73.25	67.90
		VQ	83.64	80.00	77.20	73.04	68.11	**90.10**	85.79	84.04	80.90	89.14	90.65	90.10	88.76
		avg	65.26	67.36	69.40	65.19	54.68	**78.09**	71.49	68.22	67.79	75.32	81.53	79.21	71.91
Contextualization	Containerized scenario	IF	76.07	83.37	71.27	44.23	46.12	**87.10**	70.03	81.11	76.16	90.58	93.40	89.34	73.44
		VC	59.21	59.15	74.06	80.47	35.95	79.14	**87.38**	57.04	54.09	73.13	74.89	79.85	82.37
		VQ	88.96	86.37	85.34	64.02	71.37	**94.29**	88.61	89.60	84.33	96.26	94.03	93.54	90.91
		avg	71.90	74.28	75.20	62.68	47.10	**85.35**	80.69	73.18	68.97	84.74	86.12	86.38	80.51
	Commercial design	IF	68.47	70.95	60.10	45.32	42.32	**84.11**	72.32	69.98	70.91	87.00	88.91	85.07	67.89
		VC	57.91	52.52	59.06	**76.28**	42.87	69.93	75.79	56.75	51.58	63.24	76.31	75.17	74.73
		VQ	80.42	78.52	69.02	71.22	66.69	**90.77**	88.51	79.63	72.72	92.93	93.61	91.50	85.92
		avg	66.64	65.09	61.47	62.88	47.41	**79.77**	76.95	66.62	63.54	78.68	84.81	82.40	74.23
	Informationization& Narrative Expression	IF	60.73	68.71	52.32	35.67	38.09	**85.13**	64.99	58.81	57.79	83.50	86.39	87.02	70.37
		VC	52.59	50.23	64.93	**84.18**	54.22	65.55	68.05	54.57	56.49	59.40	76.40	72.90	65.55
		VQ	74.83	74.41	60.94	60.74	59.68	**87.91**	79.87	77.44	66.89	89.20	91.83	89.12	85.93
		avg	60.29	62.46	59.09	60.09	48.86	**77.85**	69.19	60.84	59.09	75.00	83.48	81.79	71.55
	Average	IF	68.42	74.34	61.23	41.74	42.17	**85.45**	69.11	69.96	68.29	87.03	89.57	87.14	70.57
		VC	56.57	53.97	66.02	**80.31**	44.34	71.54	77.07	56.12	54.05	65.26	75.86	75.97	74.22
		VQ	81.40	79.77	71.77	65.33	65.92	**90.99**	85.67	82.22	74.65	92.80	93.16	91.39	87.59
		avg	66.28	67.28	65.25	61.89	47.79	**80.99**	75.61	66.88	63.87	79.48	84.80	83.52	75.43
Stylization	Artistic Style Transformation	IF	80.65	87.71	83.79	72.04	58.01	**89.29**	80.26	88.30	84.19	92.87	93.12	83.75	74.48
		VC	67.74	60.68	70.00	**82.29**	68.34	74.73	79.53	65.56	65.82	71.36	79.09	82.38	80.05
		VQ	86.59	75.69	75.98	82.93	75.85	**92.48**	87.07	89.27	81.46	92.48	94.15	92.93	90.98
		avg	76.67	74.49	76.71	78.32	65.71	**84.10**	81.33	79.40	76.30	84.19	87.71	85.04	80.01
	Identity&Cultural Transformation	IF	81.46	86.00	83.02	55.00	55.54	**90.25**	77.50	88.03	86.59	91.52	89.90	87.43	78.88
		VC	57.29	51.25	62.00	**70.51**	60.60	57.17	68.51	50.88	50.76	58.39	63.51	66.84	66.02
		VQ	91.29	88.52	85.57	81.10	80.76	93.25	93.18	**96.48**	92.45	94.12	96.22	94.59	93.06
		avg	73.76	72.60	75.12	66.42	62.61	**77.62**	77.04	74.86	73.43	78.79	80.61	80.63	76.57
	Material Transformation	IF	73.66	82.52	81.24	52.62	49.38	**85.80**	71.80	79.05	77.86	92.72	89.93	74.54	71.93
		VC	59.40	51.13	62.32	**77.78**	59.47	66.52	71.09	50.11	45.26	67.02	71.82	74.82	75.85
		VQ	79.91	78.53	73.74	67.75	59.87	**84.10**	80.79	79.96	72.13	87.97	87.44	82.94	80.65
		avg	69.21	69.17	72.17	65.71	55.51	**77.75**	73.31	67.66	63.67	81.49	82.19	76.33	75.24
	Average	IF	78.59	85.41	82.68	59.89	54.31	**88.45**	76.52	85.13	82.88	92.37	90.99	81.91	75.10
		VC	61.48	54.35	64.77	**76.86**	62.81	66.14	73.04	55.52	53.95	65.59	71.47	74.68	73.97
		VQ	85.93	80.91	78.43	77.26	72.16	**89.94**	87.01	88.57	82.01	91.52	92.60	90.15	88.23
		avg	73.21	72.09	74.67	70.15	61.28	**79.82**	77.23	73.97	71.13	81.49	83.50	80.67	77.27
	Overall Average	IF	71.58	78.32	69.82	48.29	45.33	**85.82**	70.13	75.53	74.31	88.34	89.12	83.38	71.24
		VC	57.20	53.69	66.00	**79.74**	55.25	68.50	73.88	55.95	54.65	63.46	73.44	74.79	71.98
		VQ	83.28	80.07	75.27	70.77	67.72	**90.26**	86.03	84.36	78.44	91.23	92.01	90.37	87.98
		avg	68.17	68.82	69.38	65.37	53.78	**79.78**	74.81	69.46	67.27	78.97	83.43	81.34	74.88

garden with classical statues and stone pathways, capturing an opulent, timeless ambiance."

- case 3: "Create a whimsical infographic titled T̈he Magical Pumpkin Spice Popcorn Journey.Ïllustrate a popcorn kernel's transformation: 1) Kernel in cozy autumn attire, 2) Bursting from the jar with cartoon energy lines, 3) A popcorn piece donning an explorer hat interacting with pumpkin and spices, 4) A celebratory popcorn parade into ceramic bowls. Use vibrant oranges and browns, with playful icons and engaging typography."
- case 4: "Design a set of chibi-style stickers centered on a horseback rider theme, showcasing the following six poses:\n\n1. Cheerfully flashing a peace sign with one hand while softly gripping the horse's reins with the other. \n2. Tearful, dramatic chibi eyes while leaning in close to the horse for emotional support. \n3. Arms outstretched beside the horse in an excited "welcome" motion. \n4. Peacefully asleep against the horse with a tiny pillow and a sweet, happy expression. \n5. Boldly pointing toward a distant horizon, featuring sparkling accents, with the horse standing majestically behind. \n6. Sending a kiss toward the horse, surrounded by floating hearts for a loving effect. \n\nEnsure the design stays true to the chibi aesthetic: \n– Oversized, expressive eyes \n–

You are a specialized assistant in generating free-form image editing instructions—your output needs to be closely tied to the original image's visual features, have diverse editing logic, and be detailed enough to guide specific editing operations.
You will receive:
1. One original image - {image}
2. Image editing category - including but not limited to {class_description}
Core Requirements (must be fully met):
1. The editing content must cover 4-5 dimensional attributes (e.g., "style + color + details") to avoid single-dimensional adjustments. Each attribute needs specific operational descriptions , not vague statements like "make the image nicer".
2. Logical clarity: The instruction must include operation object and specific modification method.
3. Based on the visual features of the original image, perform creative depth transformation and output detailed and specific editing instructions to avoid making simple modifications or preserving the original overall structure.
4. Instruction length not exceeding 65 words, with no redundant content and no ambiguity.
5. When generating editing instructions, try to avoid mentioning light-related expressions.
Output format:
{
instruction:
}

{Examples}

Figure S3. Prompt for generating instructions.

Table S2. Our Pipeline Algorithm

Algorithm
I_i denotes the original input image, P denotes the prompt used for generating questions, and $Model$ represents the image generation model.
Benchmark Construction
$instruction = MLLM_1(I_i)$
$input_i = \{(I_i, instruction), P_i\}, i = IF, VC, VQ$
$Q_{IF} \leftarrow MLLM_2(input_{IF})$
$Q_{VC} \leftarrow MLLM_2(input_{VC})$
$Q_{VQ} \leftarrow MLLM_2(input_{VQ})$
Evaluation
$I_o = Model(I_i, instruction)$
$Pairs = \{I_i, I_o, Q\}$
$Score_i = MLLM_3(Pairs), i = IF, VC, VQ$
$S = 0.4 * Score_{IF} + 0.4 * Score_{VC} + 0.2 * Score_{VQ}$

Smooth and rounded facial features \n– Fun and playful short hairstyle matching the rider's look \n– Chibi-style depictions of the rider's beige shirt and detailed, miniature representation of the horse. \n\nBackground elements should feature warm, earthy tones paired with subtle stars or confetti for a natural, outdoor-inspired ambiance. Include clean white space surrounding each individual sticker to frame them neatly. \nAspect ratio required: 9:16."

- case 5: "Transform this sterile site map into a playful children's treasure map, using a whimsical visual style. Replace page names with fun icons, like a castle for the Homepage, and path lines as winding journey trails. Add imaginative decorative elements such as colorful trees and mystical creatures along the paths, using bold, child-friendly colors and dynamic, storybook-style fonts for any text."
- case 6: "Transform the Hagia Sophia into a monumental sky guardian creature with metallic domes forming the wings, and the minarets morphing into long, elegant legs. Add intricate Byzantine patterns glowing with mystical energy across its structure, and replace the central dome with a large, ever-watchful jewel that reflects cosmic wonders."

More detailed results for each category are shown in Figures S9, S10, S11, S12, S13, S14.

Table S3. Examples of question-answer pairs, where the ideal answer in the table showing cases is 'Yes'.

	**Instruction:** Transform the figure into a chibi-style decorative resin bust for a tabletop display. Feature an embroidered brown robe with exaggerated patterns and a colorful headdress with vivid stripes. Accentuate the pomegranate by making it shine with gloss, and use minimalistic facial lines, ensuring the bust captures charm and elegance, sitting atop a smooth wooden base.
Questions-IF	Q1: Is the figure now a chibi-style bust with an oversized head and simplified body proportions, suitable for tabletop display? Q2: Does the figure wear a brown robe with visibly exaggerated and stylized embroidery patterns? Q3: Does the figure wear a headdress featuring vividly colored, clearly defined stripes? Q4: Is the pomegranate rendered with a glossy, reflective surface that highlights its shine? Q5: Does the figure's face feature minimalistic lines with simplified facial features such as small eyes and soft contours? Q6: Is the bust mounted on a smooth, flat wooden base?
Questions-VC	Q1: Is the pomegranate present and clearly identifiable as a red, round fruit with visible internal seeds, held in the hands? Weight: 3 Q2: Does the headscarf retain its golden-yellow base color with diagonal stripes in purple, blue, and red? Weight: 3 Q3: Is the left ear adorned with a silver earring featuring a red gemstone and dangling components? Weight: 3 Q4: Does the brown outer vest have blue embroidered detailing along the collar and shoulder seams? Weight: 2 Q5: Are both hands visibly positioned around the pomegranate, showing a clear grip? Weight: 2 Q6: Is there a silver-colored bracelet visible on the right wrist? Weight: 1 Q7: Is the inner garment under the vest primarily off-white in color? Weight: 1
Questions-VQ	Q1: Does the bust have a visibly coherent head-to-body proportion where the head is enlarged relative to the torso but remains structurally plausible? Q2: Are the facial features simplified but still clearly defined, with no missing or distorted elements such as eyes or mouth? Q3: Do the embroidered patterns on the robe appear continuous and logically structured, without jagged edges or disrupted textures? Q4: Are the stripes on the headdress evenly spaced and smoothly colored, without visible artifacts such as color bleeding or misalignment? Q5: Does the pomegranate have a glossy surface with natural-looking highlights that do not distort its shape or create false reflections? Q6: Is the wooden base fully attached to the bust and visually grounded, with no floating or misaligned sections? Q7: Does the hand holding the pomegranate have exactly five fingers with natural joint angles and no deformation?

You are an expert in instruction tracking and evaluation for visual editing tasks, and need to generate targeted questions to determine whether the image editing results fully comply with the instructions.
Input:
Image A (original): {Image-A}
Edit instructions: {Edit instructions}

Your task:
Generate a set of binary (yes/no) questions, each aimed at evaluating whether the edited version of image A is correct and fully satisfies the given editing instructions applied to image A.
For each question, an ideal answer should also be provided, that is, if the edited version of image A is a perfect execution of instructions, the expected answer should be provided.

Step 1: Decompose Editing Instructions
……
Step 2: Generate evaluation questions
……

Output Format
Firstly, print the decomposed instruction elements as bullet points in the block. For example:
```
-<Instruction Element 1>
-<Instruction Element 2>
- . ..
```
For each sub requirement, write down the yes/no question and its ideal answer in another Markdown block. For example:
```
Q1:
Thinking process: The thinking process requires judging the proprietary concepts in the instructions (such as "cyberpunk style", clarifying its iconic visual features: neon lights, high-tech low life scenes, etc.); Based on the knowledge of the world, suggest modifications (such as "changing to nesting dolls", based on common sense of nesting dolls, the shape should be consistent, the size should decrease and be able to nest, the bottom opening must be retained, and only the size should be proportionally reduced).
Question:<Question>
Choices: Yes, No
A: <Ideal Answer>
...
```

When generating questions, please follow the following guidelines:
-Ensure that the issue is objective and directly related to the instruction requirements.
-Do not use any asterisks, bold text, or special characters (such as * * Q1: * * or * * A: * *).
-Do not ask vague or subjective questions, transform abstract style or emotional terms into concrete, observable features.
-Generate at least 5 questions (adjusted according to the complexity of the image to fully cover the key observable content of the image)

Figure S4. Prompt for generating IF Questions.

You are a professional image consistency assessment expert, skilled in capturing visual elements in various types of images (including but not limited to daily scenes, portraits, object close ups, abstract graphics, well-known artworks, etc.), accurately judging whether the edited image retains the core and detail elements that should not be changed.
You will receive:
Image A (original image)
Editing instructions (describing how to visually modify image A): {Edit instructions}
Your task is:
Focusing on the original visual content (exclusive textures, iconic forms, unique colors, etc.) of the edited subject (such as specific objects in instructions) in image A, as well as the scene elements required by the instructions to be retained, combined with the analysis of the editing instructions to allow for modifications, accurately identify the core and detail elements that should be retained in the original image. By analyzing the allowed modification range of instructions and combining image specific features (such as iconic symbols of well-known works and unique shapes of daily images), generate question answer pairs to detect the retention of edited elements, and assign 1-3 weight points to the questions (3 points: core identity/functional elements, difficult to identify due to loss; 2 points: important scenes/related elements, affecting overall consistency; 1 point: minor details, slightly affecting coherence).
Please strictly follow the following three steps, with each step including a thought process.

Step 1 - Carefully read the editing instructions, analyze the modification requirements mentioned word by word, and clearly distinguish the changeable scope of the subject and scene……
Step 2 - Determine the elements that should remain unchanged
Refer to the "Allow Change List" in the first step and use exclusion method combined with the following logic to deduce that elements should be retained in the original image……
Step 3 – Generate evaluation questions
……
The output format is as follows:
{
Q1:
Thinking process: Explain how the problem accurately corresponds to key identifying elements (such as "Tiger pattern is the core identity identifier of tigers, and if the brown texture and distribution are lost, it is difficult to identify that the subject is a tiger, which is a key element with a weight of 3 points").
Question:<Question Content>?
Choices: Yes, No
A: Yes
Weight: <1-3>
}
...

Important Reminder
Each question must directly correspond to the 'key visual elements of the original image', and strict checks must be made to ensure that they are retained after editing
The retained elements need to cover the core recognition visual elements, associated logical visual elements, and secondary detail visual elements of the original image;
The number of generated questions shall not be less than 5 (adjusted according to the complexity of the image), and the questions shall have no cross overlap;
The output strictly follows the format, disabling unnecessary symbols such as * and - to ensure clarity and standardization.

Figure S5. Prompt for generating VC questions.

You are an image editing evaluation expert. Please generate a clear and objective set of "yes/no questions" based on the following input information to check the visual quality of the edited image, with a particular focus on the perceptual quality of the generated image, with a particular emphasis on overall authenticity, natural appearance, and the absence of obvious artifacts. It evaluates whether the output maintains structural coherence and visual plausibility without introducing distortions such as unnatural textures, broken geometric shapes, or degraded fine details.
You will receive the following:
Original Image (Image A): Reference Image
Editing instructions (describing how to edit images based on reference images): {Edit instructions}
Please note that you must combine the visual content of the original image (Image A) to clarify all specific positions that need to be edited, assist in understanding and refining editing instructions, ensure that the generated questions are visual, targeted, and reasonable for each position that needs to be edited, and reflect the evaluation of perceived quality.
Your output must be generated strictly according to the following two steps and formats:
Step 1: Disassemble Editing Instructions
……
Step 2: Generate Yes/No Questions
……

Ideal answer (assuming the edited image is faithful): Yes, each question must be output in the following format (without special symbols, bolding, numbering, or listing symbols):
```
Q1:
Thinking process:
Question:<Question 1>
Choices: Yes, No
A: <Ideal Answer>
```
For example,
Q1:
Thinking process: Based on the sub requirement of "adding and modifying elements at the character's top (changing red to blue), whether editing at this position affects the overall sense of reality and natural appearance", judge whether the color dimension of the editing target at this position has a natural appearance, and evaluate the perceived quality dimension of natural appearance.
Question: Is the blue color on the character's shirt natural without any obvious unnatural color blocks?
Choices: Yes, No
A: Yes

Precautions
-The problem must be semantically clear and can be judged by observing the edited image
-Try not to use subjective evaluation words (such as "beautiful", "good-looking", "realistic", etc.) in the questions
-The default ideal answer is yes (i.e. the ideal answer for perfect editing)
-Do not output content beyond the list and question pairs
-Generate at least 5 questions (adjusted according to the complexity of the image), and the questions shall have no cross overlap

Figure S6. Prompt for generating VQ questions.

SYSTEM_PROMPT = """
You will be provided with two images:
- Image A: the original image
- Image B: the edited image

Your task is to answer questions related to Image B. When answering, you must combine your world knowledge (such as common sense, background information, etc.) for analysis and need to reflect the thinking process.
Follow these strict steps:
1. Based on the visual cues in Image B, independently analyze and describe the relevant region or object by combining world knowledge (without referring to Image A for the time being), and briefly explain the analysis process (e.g., how to draw conclusions through visual information combined with common sense).
2. If the question requires a comparison, first analyze and describe the same region in Image A by combining world knowledge based on its visual cues, then compare it with the analysis result of Image B, and explain the knowledge and reasoning logic used in the comparison process.
3. Strictly output in the following format, without irrelevant symbols such as *,%, etc:
{
"Question": "Q1:Does the chibi keychain feature a large, singular pearl earring on the girl's left ear?",
"explanation": "A detailed description of the analysis process, including how to reason and judge by combining visual cues with world knowledge; if comparison is involved, also explain the thinking logic in the comparison.",
"answer": "{correct option}"
}

When answering the questions, please follow these guidelines:
- On the basis of adhering to the given image content, fully use world knowledge to assist in understanding and analysis.
- The thinking process must be clear and coherent, reflecting the reasoning logic from visual information to conclusions, ensuring the rationality and understandability of the explanation.
- Your answer must be from the provided choices.
"""
{Questions}

Figure S7. Prompt for evaluation.

Figure S8. More visual comparison.

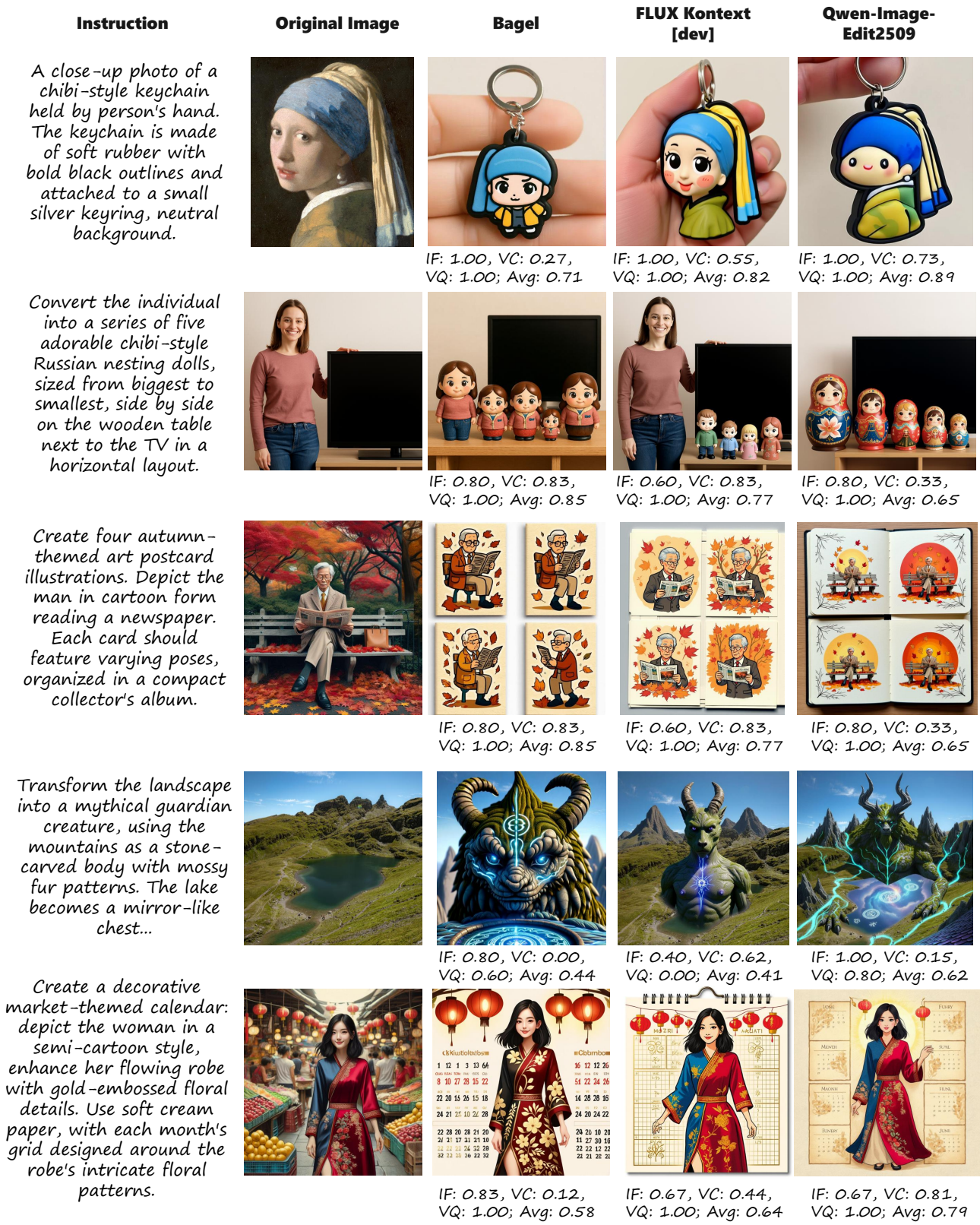

Figure S9. Visual comparison of open-source models in Customization.

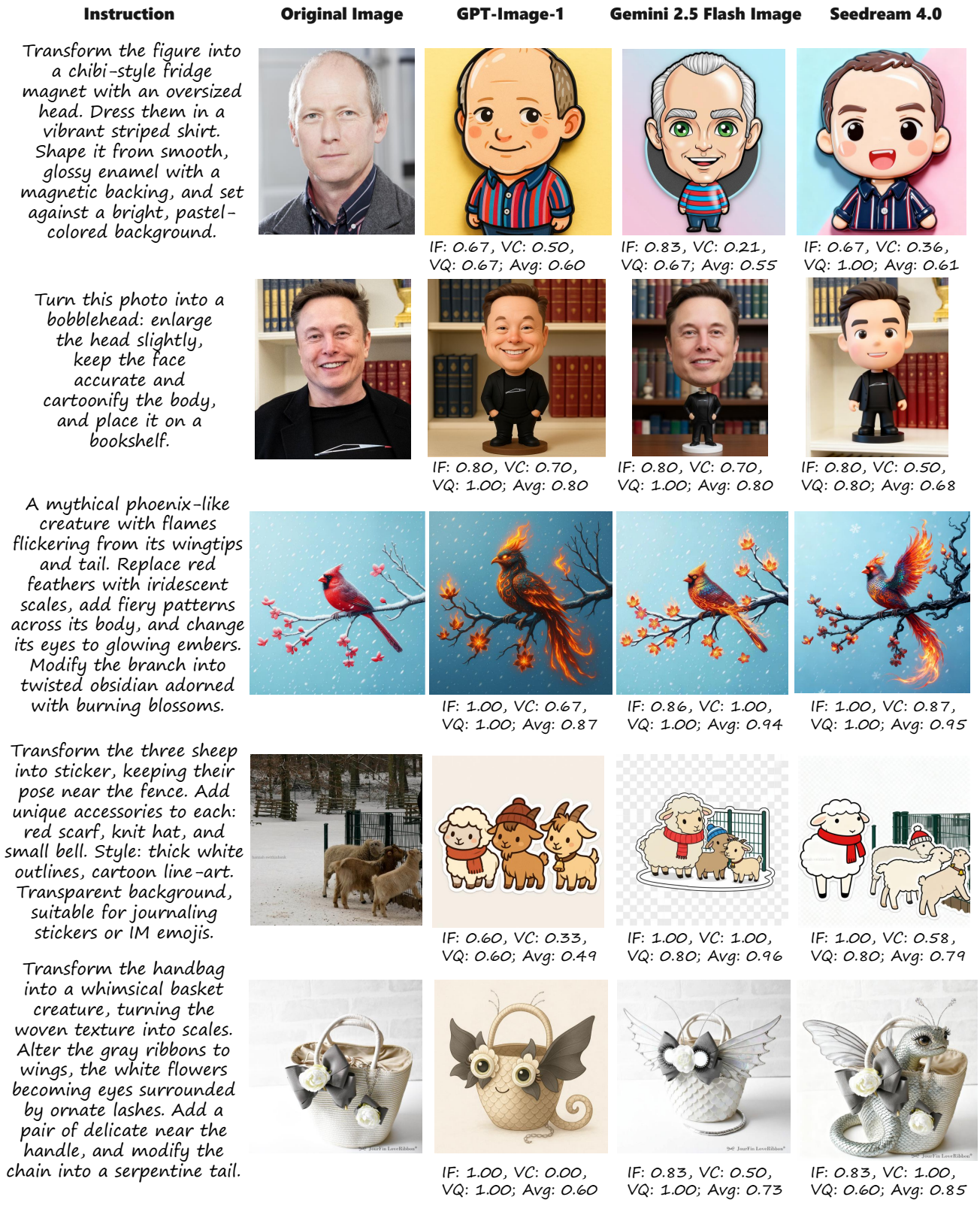

Figure S10. Visual comparison of closed-source models in Customization.

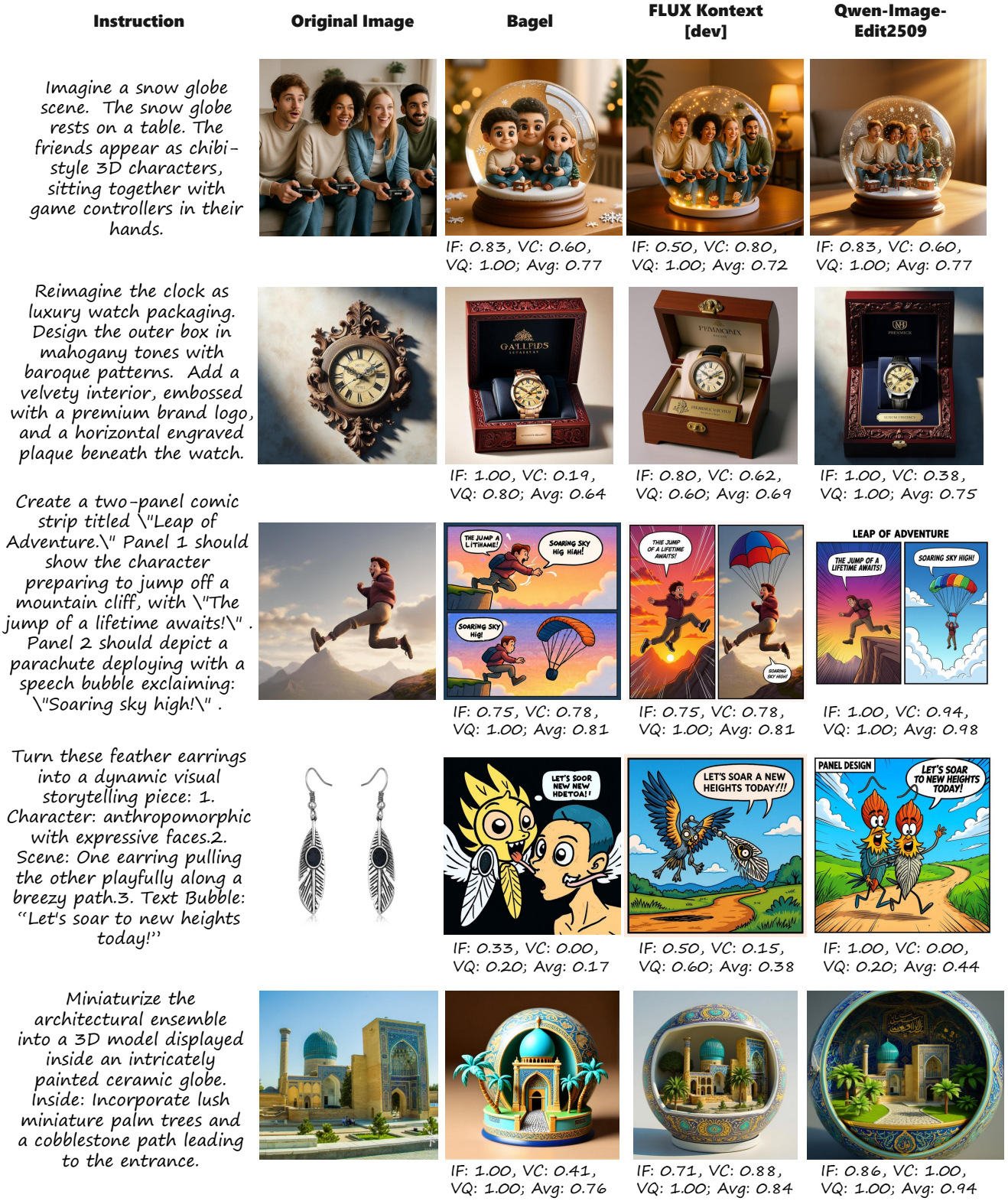

Figure S11. Visual comparison of open-source models in Contextualization.

Figure S12. Visual comparison of closed-source models in Contextualization.

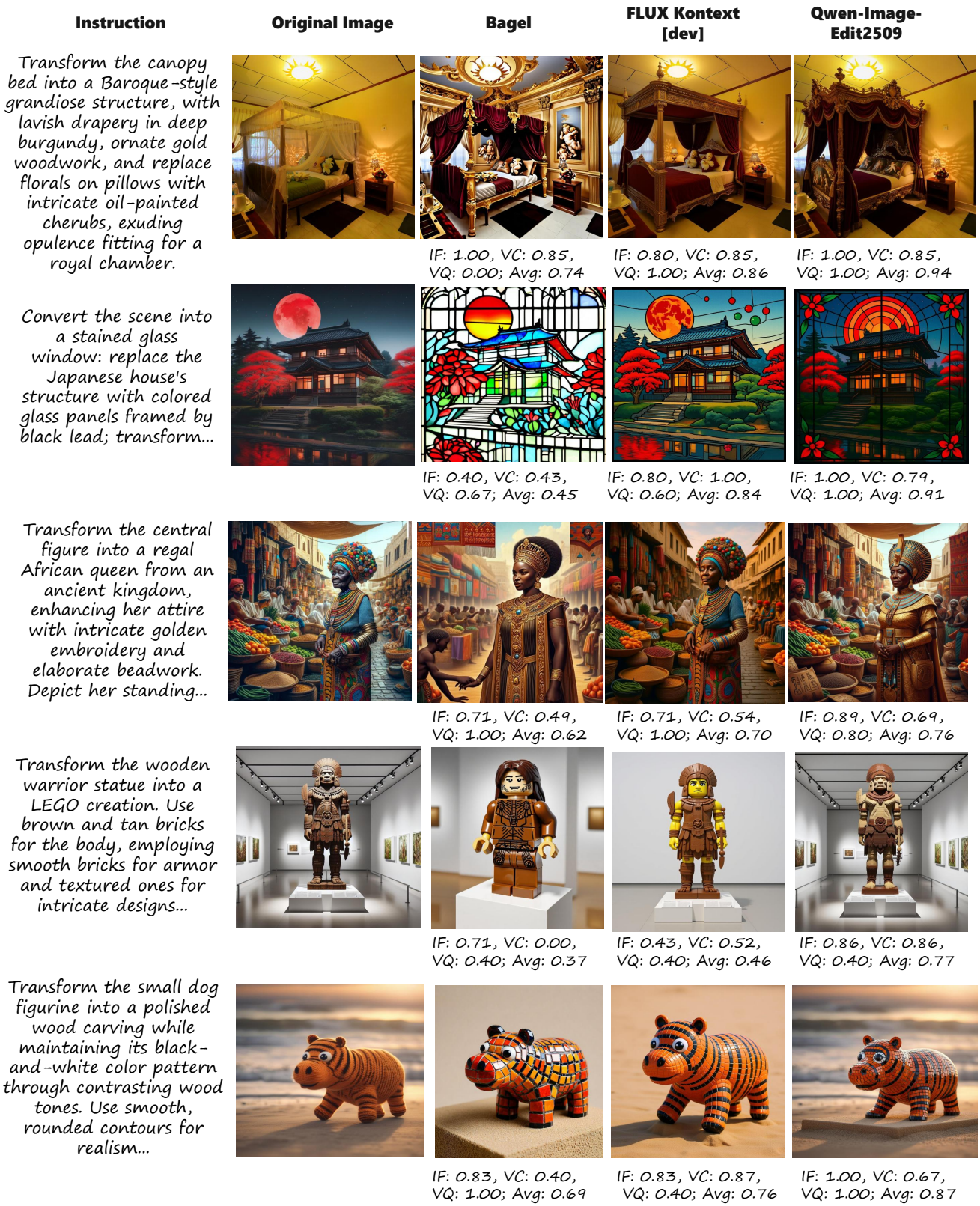

Figure S13. Visual comparison of open-source models in Stylization.

Figure S14. Visual comparison of closed-source models in Stylization.