WiseEdit: Benchmarking Cognition- and Creativity-Informed Image Editing

Supplementary Material

Appendix Overview

In this supplementary material, we present:
- Benchmark Details in Section A.
- Data Collection in Section B.
- Evaluation Details in Section C.
- More Experimental Results in Section D.

A. Benchmark Details

In this section, we provide a detailed explanation of the various tasks and knowledge types incorporated within the WiseEdit benchmark.

A.1. Task Type

The WiseEdit dataset features four distinct types of challenging tasks integrated across the three editing steps: **Awareness Task**, **Interpretation Task**, **Imagination Task**, and **WiseEdit-Complex**. Case examples illustrating each of these tasks are provided in Fig. 3 of the main paper.

Awareness Task. We design Awareness tasks, challenging the model to deduce the target modification area without explicit spatial information provided in the instruction, which necessitates the application of reasoning capabilities. Concrete examples of sub-tasks within the Awareness Task category include (***but are not limited to***):
- **Property Comparison**: Identifying the subject to be edited by comparing the differential properties among multiple similar subjects within the input image.
- **Regional Perception**: Determining the target edit region by analyzing the spatial relationship between given detection boxes and a specific object in the image.
- **Fine-grained OCR Search**: Localizing the exact word mentioned in the instruction among dense text using Optical Character Recognition (OCR) capabilities.
- **Camera Depth**: Pinpointing the subject to be edited by comparing the relative camera depths of various objects in the input image.
- **Function Recognition**: Identifying the object to be edited in the image based on a functional description provided in the instruction, differentiating it from other objects with distinct functions.
- **Geographical Locating**: Given a map (whether of a country, continent, or the world), the primary task is to identify the specific provinces or countries that require editing by interpreting the accompanying knowledge-based instructions or descriptive cues.
- **Visual Correspondence via Reference Image**: Locating the object in the editing image that has the correct relationship described in the instruction to an object in a separate reference image to perform the subsequent edit.

Interpretation Task. In the Interpretation task, instructions often do not explicitly state the required modification. The models must leverage world knowledge to parse the implicit intent into a directly executable action sequence. Concrete examples of sub-tasks within the Interpretation Task category include (***but are not limited to***):
- **Error Correctness**: Given an image that contains elements inconsistent with real-world plausibility, perform appropriate modifications to correct the error.
- **Violation Correctness**: Given an image that contains violations (*e.g.*, policy, safety), perform appropriate deletions and modifications to correct the violation.
- **Common Sense Reasoning**: Based on common sense knowledge, predict the resulting changes an implicit instruction will introduce to the image.
- **Cascaded Change Prediction**: An instruction that appears straightforward triggers a sequence of dependent, cascaded reactions within the scene upon execution.
- **Rule-Based Reasoning**: In scenarios governed by explicit rules, such as board games (*e.g.*, chess), correct content in the image that violates those established rules.
- **Complex Referencing**: The target object for editing is not given directly but is described through complex descriptions or indirect references (*e.g.*, modifying an object to a "traditional dish eaten during the Lantern Festival which is round", where the model should first refer the description to tangyuan).
- **Biology-Related Reasoning**: Perform inference based on biological knowledge (*e.g.*, predicting the mature form of bamboo shoots).
- **Mathematical Reasoning**: Perform inference based on mathematical knowledge (*e.g.*, solving an equation, or predicting the function graph after a periodic function completes half a cycle).
- **Physics-Related Reasoning**: Perform inference based on physical knowledge (*e.g.*, predicting the direction of a scale's tilt after adding or removing weights from a balance beam).
- **Chemistry-Related Reasoning**: Perform relevant inference based on chemical knowledge (*e.g.*, predicting the color change of a solution after adding a specific chemical substance).
- **Logical Reasoning**: Deduce the edited image through abstract logical reasoning.
- **Conditional Inference**: The given instruction has multiple conditional branches, requiring the model to select

and execute the correct branch based on its understanding of the image scenario.
- **Spatial Ordering**: Given multiple objects in an image, the model must re-order and reposition them from left to right according to a complex instruction.
- **Multi-Instruction Composition**: The instruction involves a combination of multiple sub-instructions, requiring the model to complete several operations in a single, cohesive edit.

Imagination Task. Imagination serves as the generative step, rendering the visual edits parsed previously onto the target regions of the original image. For this task, we introduce highly challenging subject-driven generation tasks that require models to perform imaginative and artistic creations while preserving the subject's identity. Concrete examples of sub-tasks within the Interpretation Task category include (***but are not limited to***):

- **Texture Transfer**: Migrating a 3D-aware texture onto a 2D object to enhance its three-dimensional sense under a specific material property.
- **Image Style Transfer**: Modifying the design style of an image, such as converting a watercolor painting to a line drawing, or coloring a line drawing with a specified palette.
- **Watermark Addition and Removal**: Adding a watermark to or removing a watermark from an image.
- **Counterfactual Scenario Imagination**: Generating imaginative counterfactual scenes, such as making a car grow wings and fly.
- **Temporal Prediction**: Given a multi-frame video, predict the subsequent frame.
- **Pose Alteration**: Changing a subject's pose based on a reference image; this may involve combining multiple subjects.
- **Scene Alteration**: Changing the background scene where the subject is located, based on a reference image.
- **Viewpoint Transformation**: Changing the viewpoint of the subject or scene, *e.g.*, from a frontal view to a top-down (aerial) view.
- **Clothing/Attire Alteration**: Changing the subject's clothing based on a reference image.
- **Core Subject Identity Alteration**: Changing the subject's core appearance, such as having the subject cosplay a specific character, altering stylistic details like beards and hairstyles, or generating an image representing their childhood or elderly appearance.
- **Compositional Subject Alteration**: Combining the above subject alteration requirement for complex designs.
- **Person-to-Object and Object-to-Person Transformation**: This includes transforming a person into a doll, statue, or virtual/anime character, and vice versa, transforming a doll or virtual/anime character into a realistic person.
- **Object State Modification**: Changing the physical state of an object, such as making it transparent or causing it to explode.
- **Object Restoration**: Restoring a damaged or incomplete object to its original, complete state.
- **Themed Style Scene Design**: Designing a corresponding scene based on the theme or subject appearance provided in a reference image.
- **Graphic Design**: Creating flat/2D designs for a given character IP, such as posters, mobile wallpapers, comics, game interfaces, etc.
- **Logo Design**: Designing a logo semantically related to a specific concept.
- **Architectural Design**: Designing a 3D building based on 2D planning information (*e.g.*, blueprints).

WiseEdit-Complex. We further designed WiseEdit-Complex, a set of tasks where none of the three editing steps—awareness, interpretation, or imagination—can be easily completed. This requires the model to fully leverage its capacity for complex reasoning and creative generation. Each example in WiseEdit-Complex combines the difficulty processes of the three aforementioned tasks. Here are a few illustrative cases:

Example 1:

- **Instruction:** *Have the person with longer hair in the first image hold the national flag of the country where the second image was taken, wear the clothes of the person in the third image, and wear the hat of the person in the fourth image.*
- **Required Steps:**
  - **Awareness:** Identify which person in the first image has longer hair.
  - **Interpretation:** Deduce the country where the second image was taken and retrieve its national flag.
  - **Imagination:** Modify the identified person to wear the specific clothing and hat from the third and fourth images, and have them hold the correct national flag.

Example 2:

- **Instruction:** *Identify the club logo in the second image which is most relevant to the third image. Then let the person in the first image put on this club's home jersey and strike the pose in the fourth image.*
- **Required Steps:**
  - **Awareness:** Identify the club logo in the second image and determine its relevance to the third image.
  - **Interpretation:** Use external knowledge to correctly determine the appearance of that club's home jersey.
  - **Imagination:** Render the person in the first image wearing the determined jersey and adopting the specific pose from the fourth image.

Example 3:

- **Instruction:** *Let the person in the first image hold the longest object from the second image. If the object can be used for paper-cutting, have the person face the camera; otherwise, have him back to the camera.*
- **Required Steps:**
  - **Awareness:** Determine the longest object visible in the second image.
  - **Interpretation:** Analyze the functionality of the identified object to determine if it is typically used for paper-cutting. This determines the correct conditional branch.
  - **Imagination**: Modify the person in the first image to hold the object and correctly adjust their viewpoint (facing or backing the camera) based on the interpretation result.

A.2. Knowledge Type

In WiseEdit, we structure our tasks around three core types of knowledge to emulate human cognitive learning: **Declarative**, **Procedural**, and **Metacognitive knowledge**. Building upon this, we further broaden the knowledge scope by covering diverse domains, including **natural science**, **cultural common sense**, and **spatio-temporal-logic reasoning**.

Declarative knowledge. It is often described as "***knowing what***", encompasses facts and concepts that can be explicitly stated and defined. In the context of elaborating on different task types, several sub-categories fall under Declarative knowledge. These include: mathematical laws, physical principles, chemical reactions, descriptive definitions of objects, rules for games (like chess), social policies, comparative perceptions, functional descriptions of object properties and so on. For example, given the instruction, "Remove the illegal piece according to the Chinese chess rule", the model must possess the declarative knowledge regarding the movement constraints and forbidden zones for each piece in Chinese chess.

Procedural knowledge. It is defined as "***knowing how***", encompasses the expertise and skills necessary to perform a task, which is often dynamic and difficult to explain verbally. In the context of task types described above, examples such as multi-step instruction following, cascading reaction prediction, time series forecasting, logical reasoning, architectural design, viewpoint transformation, and style transfer are generally categorized as Procedural Knowledge. For example, when required to convert a watercolor painting into a line drawing, the model relies on procedural knowledge to complete the process systematically.

Metacognitive knowledge. It is "***knowing about knowing***", requiring self-awareness and self-regulation during image editing. It dictates the high-level management of when to invoke declarative or procedural knowledge, and how to effectively combine them. In the context of complex tasks—like the types described previously—metacognitive knowledge is essential for conditional branching and controlling subsequent editing actions. For example, executing a conditional instruction "*have the girl hold the longest object in the second image; if the object can be used to brush teeth, let her face the camera, otherwise turn her back to the camera*", requires that the model knows when and why to apply the necessary declarative or procedural steps.

Knowledge Domain. On this basis, we also systematically encompass three critical knowledge domains. **Cultural common-sense** evaluates the model's understanding of nuanced human experiences and common-sense knowledge (*e.g.*, social customs, daily life), ensuring culturally-appropriate image editing. **Natural sciences** assess the model's comprehension of domain-specific principles (*e.g.*, physical laws, chemical reactions) essential for generating physically consistent images. **Spatio-temporal-logical** requires the model to possess a clear cognition of causality, spatial arrangement, and temporal progression, generating images that adhere to coherent spatio-temporal relations and logical consistency.

B. Data Collection

Our benchmark images are primarily sourced from two methods: collected from the Internet or generated using generative models. A very small portion of the images is drawn from existing datasets. For each test case, the editing instruction is initially created by trained human annotators. To ensure clarity, eliminate ambiguity, and enhance the diversity of the instructions, we utilize GPT-4 [1] to augment the original prompts without altering their semantic meaning. Furthermore, for instructions originally written in English, we also use GPT-4 to generate corresponding Chinese translations. Finally, a dual-screening process involving both Ph.D. experts and GPT-4 is employed to filter out unreasonable or unachievable test cases, resulting in a final benchmark of 1,220 test cases.

Specifically, the Awareness task comprises 362 cases, where 281 cases involve a single image input, and 81 cases involve two image inputs. The Interpretation task includes 317 cases, all of which utilize a single image input. The Imagination task consists of 451 cases, distributed as follows: 320 cases with a single image input, 109 cases with two image inputs, 8 cases with three image inputs, 16 cases with four image inputs, and 8 cases with five image inputs. Furthermore, the WiseEdit-complex subset contains 80 cases, all of which involve multiple image inputs: 47 cases utilize two image inputs, 17 cases utilize three image inputs, and 16 cases utilize four image inputs.

C. Evaluation Details

Evaluation Models. We evaluate 20 mainstream image editing models across architectures, covering both open- and closed-source models: **(1) Diffusion models:** MagicBrush [46], OmniGen [43], AnyEdit [45], UltraEdit [48], ICEdit [47], FLUX.1 Kontext Dev [20], FLUX.2 Dev [19]. Specifically, MagicBrush, AnyEdit and UltraEdit utilize the UNet architecture based on SD 1.5 or SDXL [35], while OmniGen, ICEdit, FLUX.1 Kontext Dev, and FLUX.2 Dev are based on the Diffusion Transformer (DiT) [33] architecture. Notably, FLUX.1 Kontext Dev is a large DiT model with 12 billion parameters. While FLUX.2 Dev is a 32B DiT model with another 24B model [2] for text encoding.

(2) Unified comprehension and generation models: UniWorld-V1 [23], HiDream-E1 [5], OmniGen2 [40], Step1X-Edit-v1p2 [25], Echo-4o [44], Bagel [10], Uni-CoT [34], Qwen-Image-Edit [39], and DreamOmni2 [42]. These models integrate the powerful visual comprehension [3, 7, 21, 29] capabilities of autoregressive Vision-Language Models (VLMs) with diffusion models. They employ three primary architectures: **(a) Quantized AR**: leveraging autoregressive visual generation [9, 24, 28, 30–32, 38] with discrete visual tokenizers, including Janus-4o. **(b) Cascaded Architecture**: Some models implement a system where an external diffusion model is cascaded after the VLM's output, including UniWorld-V1, HiDream-E1, OmniGen2, Step1X-Edit-v1p2, Qwen-Image-Edit, and DreamOmni2. **(c) Integrated Architecture**: Other models, *i.e.*, Echo-4o, Uni-CoT, and Bagel introduce a mixture-of-transformer approach within a single, integrated model. Here, the model conducts autoregression to handle text generation while conducting diffusion to manage image generation. This configuration allows the model to simultaneously perform both visual understanding and visual generation in an integrated model.

(3) Close-sourced Models: Nano Banana [13], Seedream 4.0 [36], GPT-image-1 [27], and Nano Banana Pro [15]. These models do not have publicly available checkpoint weights; access is exclusively provided via API calls. Notably, these models, like those in the second category, are also unified comprehension and generation models. Specifically, Seedream 4.0's technical report explicitly states that it leverages the comprehension capabilities of a VLM to aid in generation like [6, 50]. Although GPT-image-1 and Nano Banana have not released technical reports detailing their methodology, the prevailing view in the community suggests they are a combination of VLM and diffusion models like [8, 12]. And Nano Banana Pro, in turn, leveraged the powerful multimodal understanding capabilities of Gemini 3 [14].

Evaluation Metrics. As detailed in Section 3.3, the evaluation of WiseEdit employs a comprehensive set of six metrics: Instruction Following, Detail Preserving and Visual Quality, Knowledge Fidelity, and Creative Fusion. Specifically, for the Awareness and Interpretation tasks, we utilize Instruction Following, Detail Preserving, and Visual Quality, and Knowledge Fidelity. For the Imagination task, the metrics used are Instruction Following, Detail Preserving and Visual Quality, and Creative Fusion. Finally, the WiseEdit-Complex task is evaluated using all four metrics. For the assessment of all these metrics, we leverage GPT-4o as the automatic evaluator, which rates performance on a 1–10 scale based on carefully custom-crafted and dimension-specific prompts. The prompt templates for Instruction Following with both single-image input and multi-image input are in Fig. 12 and Fig. 13, respectively. The prompt templates for Detail Preserving with both single-image input and multi-image input are in Fig. 14 and Fig. 15, respectively. The prompt templates for Visual Quality with both single-image input and multi-image input are in Fig. 16. The prompt templates for Knowledge Fidelity with both single-image input and multi-image input are in Fig. 17 and Fig. 18, respectively. The prompt templates for Creative Fusion with both single-image input and multi-image input are in Fig. 19 and Fig. 20, respectively.

It is worth noting that for some knowledge-informed test cases, we additionally provide knowledge hints and a hint reference image to aid in the assessment of metrics such as Knowledge Fidelity. Finally, we linearly map each 1-to-10 score provided by the evaluator to a 0-to-100 range and then compute the average score for each task. The overall average score is the average of the "average scores" from the Awareness, Interpretation, and Imagination tasks. Besides, any test case where a model fails to handle the input is assigned a score of zero across all metrics.

D. More Experimental Results

Performance on WiseEdit with Single-Image Inputs. In the WiseEdit benchmark, the input for many test cases is highly free-form, often incorporating multiple input images. To better differentiate a model's proficiency in handling single-image inputs from its overall capability with free-form inputs, we calculate the model's performance specifically on the single-image input subsets for each task and compare these metrics against the original performance in Table 5 and Table 6. Notably, the Interpretation task exclusively features single-image inputs, meaning its single-image results are identical to those reported in Table 2 and are thus not presented again in Table 5 and Table 6. Besides, the WiseEdit-Complex task is composed entirely of multi-image inputs, for which we can not calculate separate single-image results. The results reveal that for the majority of models, the average performance across all cases is

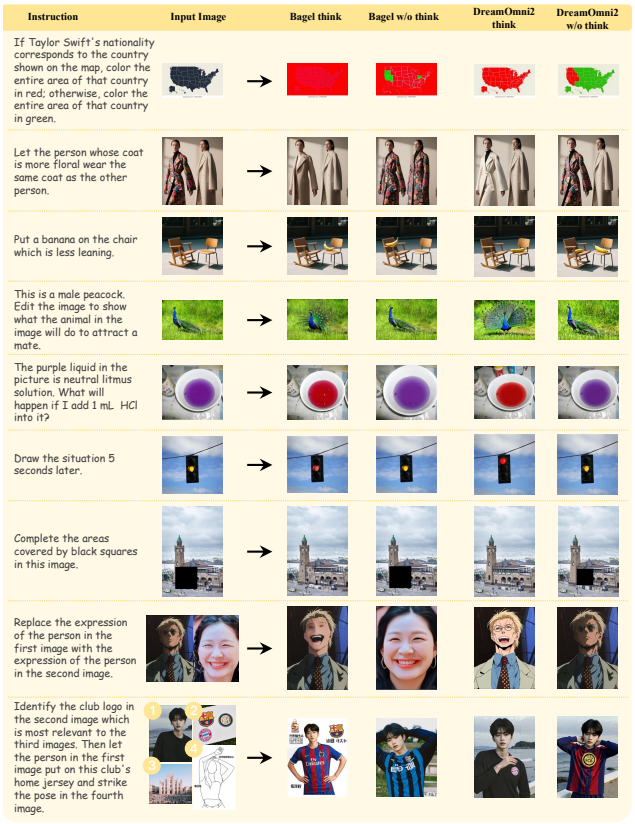

Figure 6. Qualitative comparisons of Bagel and DreamOmni2 with and without their built-in thinking processes.

significantly lower than the average performance observed for single-image input. This observation highlights that improving capability in complex, multi-image scenarios is a key direction for future image editing models.

Qualitative Comparisons of *with think* v.s. *w/o think*. To illustrate how visual comprehension enhances visual generation in unified models, we present qualitative comparisons of Bagel and DreamOmni2 with and without their built-in reasoning processes. As shown in Fig. 6, disabling this mechanism leads to a noticeable degradation in generation quality, underscoring the critical role of visual comprehension in improving visual generation.

Qualitative Comparisons of *original instructions* v.s. *rewritten instructions*. To visualize the impact of instruction rewriting, we give some qualitative comparisons before and after instruction rewriting based on Bagel and GPT-Image-1. As shown in Fig. 7, for Bagel, the rewritten instruction provides additional knowledge, which significantly improves the editing results. Whereas for GPT-Image-1, its internal comprehension mechanisms have already acquired the knowledge externally introduced by rewriting, thus the improvement brought by rewriting appears less pronounced.

More Qualitative Comparisons. In Fig. 8, Fig. 9, Fig. 10, and Fig. 11, we present additional qualitative comparisons across four tasks, *i.e.*, Awareness, Interpretation, Imagination, and WiseEdit-Complex. We show the generated results of nine models, *e.g.*, AnyEdit, FLUX.1 Kontext Dev, FLUX.2 Dev, Bagel, Qwen-Image-Edit, OmniGen2, DreamOmni2, Seedream 4.0, GPT-Image-1, Nano Banana and Nano Banana Pro. For any model that is unable to process a given input, its output is indicated by an image with a prohibition symbol.

Table 5. Main results on WiseEdit across all cases or only single-image inputs on English version of instructions. For the majority of models, the average performance across all cases is lower than the average performance observed for single-image input.

Model	*Cases*	Awareness Task					Imagination Task					Overall
		$\mathcal{IF}\uparrow$	$\mathcal{DP}\uparrow$	$\mathcal{VQ}\uparrow$	$\mathcal{KF}\uparrow$	**AVG**	$\mathcal{IF}\uparrow$	$\mathcal{DP}\uparrow$	$\mathcal{VQ}\uparrow$	$\mathcal{CF}\uparrow$	**AVG**	**AVG**
English Version												
MagicBrush	*all cases*	27.2	43.4	53.3	27.1	37.8	18.0	36.9	44.8	22.3	30.5	35.5
	sing-img cases	36.6 (+9.3)	57.4 (+14.0)	70.2 (+16.9)	36.4 (+9.3)	50.1 (+12.4)	27.5 (+9.4)	54.6 (+17.7)	66.1 (+21.3)	33.7 (+11.3)	45.5 (+14.9)	44.6 (+9.1)
OmniGen	*all cases*	35.0	42.0	46.7	37.4	40.3	42.2	35.1	46.0	38.7	40.5	36.6
	sing-img cases	34.6 (-0.4)	43.6 (+1.6)	48.5 (+1.9)	34.7 (-2.7)	40.3 (+0.1)	47.9 (+5.7)	35.7 (+0.6)	44.7 (-1.2)	41.5 (+2.8)	42.5 (+2.0)	37.2 (+0.7)
AnyEdit	*all cases*	25.0	54.6	61.3	26.3	41.8	9.1	49.7	50.9	16.5	31.5	37.7
	sing-img cases	32.9 (+7.9)	68.0 (+13.4)	72.8 (+11.5)	33.5 (+7.2)	51.8 (+10.0)	12.9 (+3.8)	70.1 (+20.4)	66.6 (+15.7)	18.4 (+2.0)	42.0 (+10.5)	44.5 (+6.8)
UltraEdit	*all cases*	26.5	42.5	53.1	33.9	39.0	20.7	31.7	45.8	27.5	31.5	39.0
	sing-img cases	35.7 (+9.1)	56.3 (+13.8)	69.9 (+16.8)	45.2 (+11.3)	51.8 (+12.8)	31.4 (+10.6)	47.2 (+15.5)	67.5 (+21.7)	41.2 (+13.6)	46.8 (+15.4)	48.4 (+9.4)
ICEdit	*all cases*	26.1	42.2	61.2	31.8	40.4	21.5	40.6	54.0	25.0	35.3	39.9
	sing-img cases	34.9 (+8.8)	55.6 (+13.4)	80.1 (+18.9)	42.3 (+10.4)	53.2 (+12.9)	32.1 (+10.6)	59.6 (+19.0)	78.9 (+24.9)	37.2 (+12.1)	51.9 (+16.6)	49.7 (+9.8)
FLUX.1 Kontext Dev	*all cases*	31.4	52.0	55.0	35.5	43.5	39.1	47.1	43.4	27.1	39.2	43.2
	sing-img cases	40.7 (+9.3)	67.2 (+15.2)	71.0 (+16.1)	45.9 (+10.5)	56.2 (+12.8)	55.9 (+16.8)	67.5 (+20.3)	62.1 (+18.7)	38.6 (+11.5)	56.0 (+16.8)	53.1 (+9.9)
FLUX.2 Dev	*all cases*	42.6	63.3	78.4	53.3	59.4	**73.6**	70.7	**82.1**	**43.6**	**67.5**	**61.8**
	sing-img cases	44.5 (+1.8)	62.6 (-0.8)	81.5 (+3.1)	51.6 (-1.7)	60.0 (+0.6)	**78.8** (+5.3)	68.9 (-1.8)	82.9 (+0.8)	**45.7** (+2.1)	69.1 (+1.6)	62.5 (+0.7)
UniWorld-V1	*all cases*	31.5	48.9	58.8	38.6	44.5	30.3	50.3	64.2	27.5	43.1	41.1
	sing-img cases	31.9 (+0.5)	46.8 (-2.1)	59.5 (+0.6)	39.1 (+0.5)	44.3 (-0.1)	31.9 (+1.6)	56.3 (+5.9)	66.2 (+2.0)	26.3 (-1.2)	45.2 (+2.1)	41.8 (+0.7)
HiDream-E1	*all cases*	29.7	41.2	56.3	32.0	39.8	39.6	40.1	49.9	29.6	39.8	41.4
	sing-img cases	39.7 (+10.1)	54.6 (+13.4)	74.1 (+17.7)	42.7 (+10.7)	52.8 (+13.0)	58.5 (+18.9)	59.3 (+19.2)	73.3 (+23.5)	44.1 (+14.5)	58.8 (+19.0)	52.1 (+10.7)
OmniGen2	*all cases*	35.0	64.0	75.4	41.3	53.9	42.0	64.4	74.6	31.8	53.2	49.4
	sing-img cases	36.4 (+1.4)	65.6 (+1.6)	75.3 (-0.1)	41.2 (-0.1)	54.6 (+0.7)	41.1 (-0.9)	70.9 (+6.5)	74.7 (+0.1)	29.6 (-2.3)	54.1 (+0.9)	49.9 (+0.5)
Step1X-Edit-v1p2	*all cases*	39.8	53.5	61.3	44.4	49.7	44.7	49.4	50.3	28.4	43.2	49.5
	sing-img cases	52.9 (+13.1)	70.5 (+17.0)	80.6 (+19.3)	58.8 (+14.4)	65.7 (+15.9)	66.0 (+21.3)	72.8 (+23.4)	74.1 (+23.8)	42.5 (+14.1)	63.9 (+20.7)	61.7 (+12.2)
Echo-4o	*all cases*	47.6	63.0	75.4	51.7	59.4	63.4	62.4	73.7	41.2	60.2	57.8
	sing-img cases	52.7 (+5.1)	65.8 (+2.9)	77.7 (+2.3)	53.7 (+2.0)	62.5 (+3.1)	70.8 (+7.4)	65.2 (+2.8)	72.7 (-1.0)	44.0 (+2.8)	63.2 (+3.0)	59.8 (+2.0)
Bagel	*all cases*	46.2	71.0	75.8	50.8	61.0	62.8	**68.5**	74.5	40.7	61.6	60.0
	sing-img cases	50.5 (+4.3)	73.3 (+2.3)	79.8 (+3.9)	50.0 (-0.8)	63.4 (+2.4)	72.4 (+9.6)	**75.0** (+6.5)	75.1 (+0.6)	42.8 (+2.1)	66.3 (+4.7)	62.3 (+2.4)
Uni-CoT	*all cases*	46.0	69.1	77.8	51.6	61.1	67.6	64.3	79.6	42.9	63.6	60.1
	sing-img cases	50.6 (+4.7)	71.7 (+2.6)	79.6 (+1.7)	51.7 (+0.2)	63.4 (+2.3)	76.0 (+8.4)	67.8 (+3.6)	80.0 (+0.5)	45.2 (+2.3)	67.3 (+3.7)	62.0 (+2.0)
Qwen-Image-Edit	*all cases*	**48.1**	69.0	79.5	**53.6**	62.5	67.1	66.8	79.2	42.3	63.8	60.2
	sing-img cases	**57.7** (+9.6)	73.2 (+4.2)	82.5 (+3.0)	**60.8** (+7.2)	**68.5** (+6.0)	**78.8** (+11.6)	73.7 (+6.9)	81.9 (+2.7)	45.1 (+2.8)	**69.9** (+6.0)	**64.2** (+4.0)
DreamOmni2	*all cases*	43.3	**74.4**	**85.0**	51.2	**63.5**	50.6	64.9	81.9	35.3	58.2	60.6
	sing-img cases	46.7 (+3.4)	**77.1** (+2.6)	**86.0** (+1.0)	52.5 (+1.4)	65.6 (+2.1)	54.1 (+3.5)	68.0 (+3.0)	**84.2** (+2.3)	34.5 (-0.8)	60.2 (+2.0)	61.9 (+1.4)
Nano Banana	*all cases*	70.6	85.7	86.8	75.2	79.6	75.3	73.8	87.3	44.3	70.2	75.0
	sing-img cases	69.6 (-0.9)	86.2 (+0.5)	88.2 (+1.4)	73.8 (-1.4)	79.5 (-0.1)	83.7 (+8.4)	76.1 (+2.3)	89.6 (+2.3)	47.8 (+3.5)	74.3 (+4.1)	76.4 (+1.3)
Seedream 4.0	*all cases*	70.8	78.1	86.6	74.6	77.5	82.2	77.8	86.9	47.0	73.5	75.2
	sing-img cases	70.2 (-0.5)	78.8 (+0.7)	87.5 (+0.9)	71.7 (-2.9)	77.1 (-0.5)	88.5 (+6.2)	**81.8** (+3.9)	88.9 (+2.0)	47.1 (+0.1)	76.6 (+3.1)	76.1 (+0.9)
GPT-image-1	*all cases*	78.5	85.8	**88.0**	81.2	83.3	84.4	76.2	**89.2**	48.4	74.6	77.6
	sing-img cases	83.5 (+5.0)	87.8 (+2.0)	**88.8** (+0.8)	83.2 (+2.0)	85.8 (+2.5)	**91.4** (+7.1)	77.3 (+1.1)	**91.1** (+1.9)	49.9 (+1.5)	77.5 (+2.9)	79.4 (+1.8)
Nano Banana Pro	*all cases*	**85.4**	**88.6**	83.9	**91.4**	**87.3**	**86.6**	**79.5**	88.8	**51.5**	**76.6**	**82.4**
	sing-img cases	**90.0** (+4.6)	**88.1** (-0.6)	84.3 (+0.4)	**91.2** (-0.2)	**88.4** (+1.1)	89.7 (+3.1)	78.9 (-0.7)	89.5 (+0.7)	**52.3** (+0.8)	**77.6** (+1.0)	**82.9** (+0.5)

Table 6. Main results on WiseEdit across all cases or only single-image inputs on Chinese version of instructions. For the majority of models, the average performance across all cases is lower than the average performance observed for single-image input.

Model	*Cases*	Awareness Task					Imagination Task					Overall
		$\mathcal{IF}\uparrow$	$\mathcal{DP}\uparrow$	$\mathcal{VQ}\uparrow$	$\mathcal{KF}\uparrow$	**AVG**	$\mathcal{IF}\uparrow$	$\mathcal{DP}\uparrow$	$\mathcal{VQ}\uparrow$	$\mathcal{CF}\uparrow$	**AVG**	**AVG**
Chinese Version												
MagicBrush	*all cases*	15.0	43.8	52.9	17.8	32.4	5.4	39.6	49.2	12.7	26.8	30.0
	sing-img cases	20.8 (+5.8)	57.9 (+14.1)	69.6 (+16.7)	24.4 (+6.6)	43.2 (+10.8)	9.3 (+3.9)	58.6 (+19.0)	72.4 (+23.2)	19.8 (+7.1)	40.1 (+13.3)	38.0 (+8.0)
OmniGen	*all cases*	16.3	42.6	60.1	22.6	35.4	15.4	27.6	51.3	32.3	31.7	31.3
	sing-img cases	16.6 (+0.3)	47.8 (+5.1)	61.1 (+1.0)	21.9 (-0.6)	36.8 (+1.5)	16.8 (+1.4)	32.5 (+4.9)	52.3 (+1.0)	33.2 (+0.9)	33.7 (+2.0)	32.5 (+1.2)
AnyEdit	*all cases*	17.5	55.3	55.7	19.9	37.1	7.3	47.4	52.3	15.4	30.6	34.0
	sing-img cases	23.4 (+5.9)	67.7 (+12.4)	66.7 (+11.0)	25.6 (+5.7)	45.9 (+8.7)	10.3 (+3.0)	67.0 (+19.5)	68.4 (+16.1)	17.4 (+2.0)	40.8 (+10.2)	40.3 (+6.3)
UltraEdit	*all cases*	16.9	58.5	62.1	21.2	39.7	9.3	42.3	51.8	14.8	29.5	38.9
	sing-img cases	23.3 (+6.4)	76.9 (+18.4)	81.5 (+19.4)	28.8 (+7.6)	52.6 (+12.9)	14.8 (+5.6)	62.5 (+20.1)	76.2 (+24.3)	22.8 (+8.0)	44.1 (+14.5)	48.0 (+9.2)
ICEdit	*all cases*	12.9	29.1	63.6	17.4	30.8	5.1	37.4	56.5	16.7	28.9	32.7
	sing-img cases	18.1 (+5.2)	39.0 (+9.9)	83.5 (+19.8)	23.9 (+6.5)	41.1 (+10.4)	8.9 (+3.8)	55.4 (+18.0)	82.9 (+26.4)	25.6 (+8.9)	43.2 (+14.2)	40.9 (+8.2)
FLUX.1 Kontext Dev	*all cases*	16.5	48.4	52.1	19.1	34.0	9.2	41.9	43.3	10.6	26.3	33.7
	sing-img cases	22.8 (+6.3)	63.9 (+15.5)	68.7 (+16.5)	26.1 (+7.0)	45.4 (+11.3)	14.8 (+5.6)	61.9 (+20.0)	63.9 (+20.6)	16.7 (+6.2)	39.3 (+13.1)	41.8 (+8.1)
FLUX.2 Dev	*all cases*	43.0	60.6	79.5	51.4	58.6	**75.5**	**74.4**	82.8	**42.6**	**68.8**	**61.4**
	sing-img cases	43.4 (+0.4)	60.5 (-0.1)	83.6 (+4.1)	47.8 (-3.5)	58.8 (+0.2)	**80.7** (+5.2)	**75.2** (+0.8)	84.1 (+1.3)	44.3 (+1.8)	**71.1** (+2.3)	62.2 (+0.8)
UniWorld-V1	*all cases*	18.4	49.0	60.0	26.2	38.4	17.9	54.8	68.9	18.0	39.9	37.7
	sing-img cases	18.9 (+0.6)	49.1 (+0.1)	61.7 (+1.7)	25.3 (-0.9)	38.7 (+0.4)	12.6 (-5.3)	62.2 (+7.4)	71.3 (+2.4)	15.5 (-2.5)	40.4 (+0.5)	37.9 (+0.3)
HiDream-E1	*all cases*	28.2	37.6	51.4	32.4	37.4	32.5	39.2	47.5	27.4	36.6	38.5
	sing-img cases	37.9 (+9.6)	49.9 (+12.3)	67.7 (+16.3)	43.3 (+10.8)	49.7 (+12.3)	48.3 (+15.8)	58.0 (+18.8)	69.9 (+22.4)	40.9 (+13.6)	54.3 (+17.6)	48.5 (+10.0)
OmniGen2	*all cases*	35.1	57.9	72.4	41.0	51.6	45.5	64.0	72.0	33.8	53.8	48.8
	sing-img cases	38.9 (+3.8)	58.9 (+1.0)	73.0 (+0.5)	42.7 (+1.7)	53.4 (+1.8)	42.8 (-2.7)	66.4 (+2.5)	71.9 (-0.1)	31.2 (-2.6)	53.1 (-0.7)	49.1 (+0.3)
Step1X-Edit-v1p2	*all cases*	38.6	55.6	59.5	42.0	48.9	45.7	48.3	51.3	27.0	43.1	49.6
	sing-img cases	51.4 (+12.7)	73.2 (+17.6)	78.3 (+18.8)	55.7 (+13.7)	64.6 (+15.7)	67.4 (+21.8)	71.2 (+22.9)	75.5 (+24.2)	40.5 (+13.5)	63.7 (+20.6)	61.7 (+12.1)
Echo-4o	*all cases*	47.9	59.9	73.1	**55.0**	59.0	62.8	64.2	75.1	41.5	60.9	58.0
	sing-img cases	**53.3** (+5.4)	62.8 (+2.9)	74.7 (+1.7)	**57.2** (+2.2)	62.0 (+3.0)	70.3 (+7.5)	64.4 (+0.2)	74.6 (-0.5)	43.5 (+2.1)	63.2 (+2.3)	59.8 (+1.8)
Bagel	*all cases*	**48.5**	71.3	76.8	52.1	62.2	63.5	68.3	75.3	39.7	61.7	59.5
	sing-img cases	52.4 (+3.9)	74.7 (+3.4)	80.4 (+3.6)	52.4 (+0.3)	64.9 (+2.8)	70.4 (+6.9)	73.7 (+5.3)	76.5 (+1.1)	41.9 (+2.2)	65.6 (+3.9)	61.7 (+2.2)
Uni-CoT	*all cases*	46.2	70.0	80.7	53.6	**62.6**	65.5	65.1	79.7	41.6	63.0	60.6
	sing-img cases	48.7 (+2.5)	72.2 (+2.1)	82.5 (+1.8)	54.4 (+0.8)	64.4 (+1.8)	74.0 (+8.5)	69.0 (+3.9)	81.2 (+1.4)	44.0 (+2.4)	67.0 (+4.1)	62.5 (+2.0)
Qwen-Image-Edit	*all cases*	45.0	67.3	79.9	52.9	61.3	66.3	67.2	80.0	41.7	63.8	60.6
	sing-img cases	51.7 (+6.7)	70.8 (+3.5)	84.1 (+4.2)	57.0 (+4.0)	**65.9** (+4.6)	78.9 (+12.6)	73.9 (+6.7)	83.8 (+3.7)	**44.4** (+2.7)	70.3 (+6.4)	**64.3** (+3.7)
DreamOmni2	*all cases*	31.9	**78.7**	**85.4**	38.4	58.6	38.5	69.4	**84.8**	27.2	55.0	56.0
	sing-img cases	34.2 (+2.3)	**84.1** (+5.3)	**88.5** (+3.1)	37.7 (-0.7)	61.1 (+2.5)	38.4 (-0.0)	74.7 (+5.2)	**87.0** (+2.1)	25.5 (-1.6)	56.4 (+1.4)	57.3 (+1.3)
Nano Banana	*all cases*	71.8	83.8	86.5	70.7	78.2	76.0	75.8	87.3	43.7	70.7	75.3
	sing-img cases	71.4 (-0.4)	84.3 (+0.5)	**88.2** (+1.7)	67.9 (-2.8)	78.0 (-0.2)	84.9 (+8.9)	77.7 (+1.9)	90.5 (+3.2)	46.8 (+3.1)	75.0 (+4.3)	76.6 (+1.3)
Seedream 4.0	*all cases*	69.1	79.0	84.4	72.0	76.1	79.8	**79.7**	86.5	46.4	73.1	74.1
	sing-img cases	69.7 (+0.6)	78.2 (-0.8)	85.0 (+0.6)	69.4 (-2.6)	75.5 (-0.6)	87.5 (+7.7)	**81.6** (+1.9)	88.8 (+2.2)	46.6 (+0.2)	76.1 (+3.0)	74.9 (+0.8)
GPT-image-1	*all cases*	77.0	80.7	**86.6**	80.7	81.2	78.8	73.3	**89.6**	48.3	72.5	76.2
	sing-img cases	83.0 (+6.0)	80.6 (-0.2)	88.0 (+1.4)	82.9 (+2.3)	83.6 (+2.4)	86.7 (+7.9)	69.5 (-3.9)	**91.4** (+1.8)	50.0 (+1.7)	74.4 (+1.9)	77.6 (+1.4)
Nano Banana Pro	*all cases*	**84.6**	**91.8**	83.1	**87.9**	**86.9**	**85.5**	77.6	88.4	**51.1**	**75.6**	**81.2**
	sing-img cases	**86.5** (+1.9)	**92.5** (+0.6)	83.3 (+0.2)	**87.6** (-0.2)	**87.5** (+0.6)	**91.0** (+5.5)	77.4 (-0.2)	90.4 (+2.1)	**52.2** (+1.1)	**77.8** (+2.1)	**81.9** (+0.7)

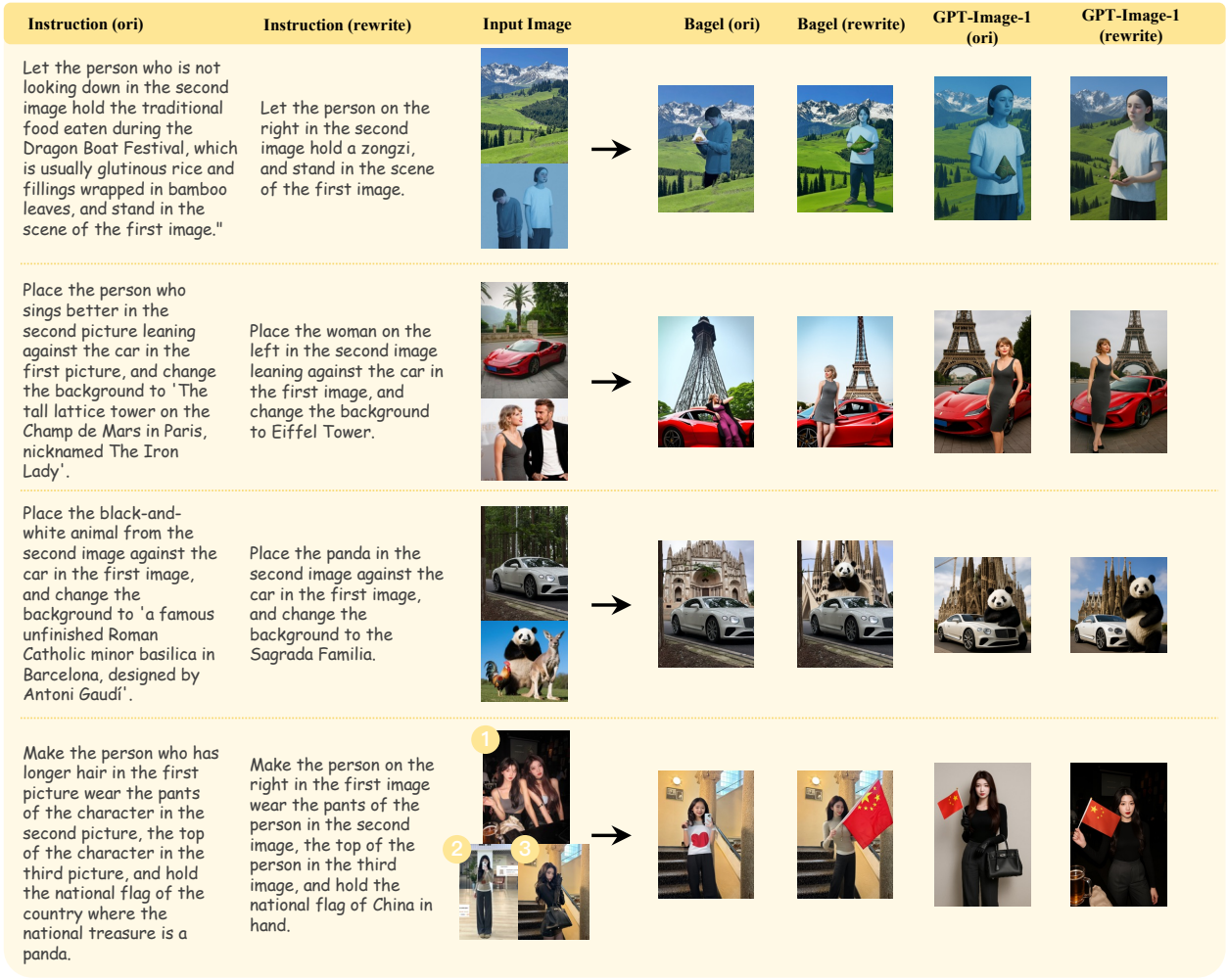

Figure 7. Qualitative comparisons before and after instruction rewriting based on Bagel and GPT-Image-1.

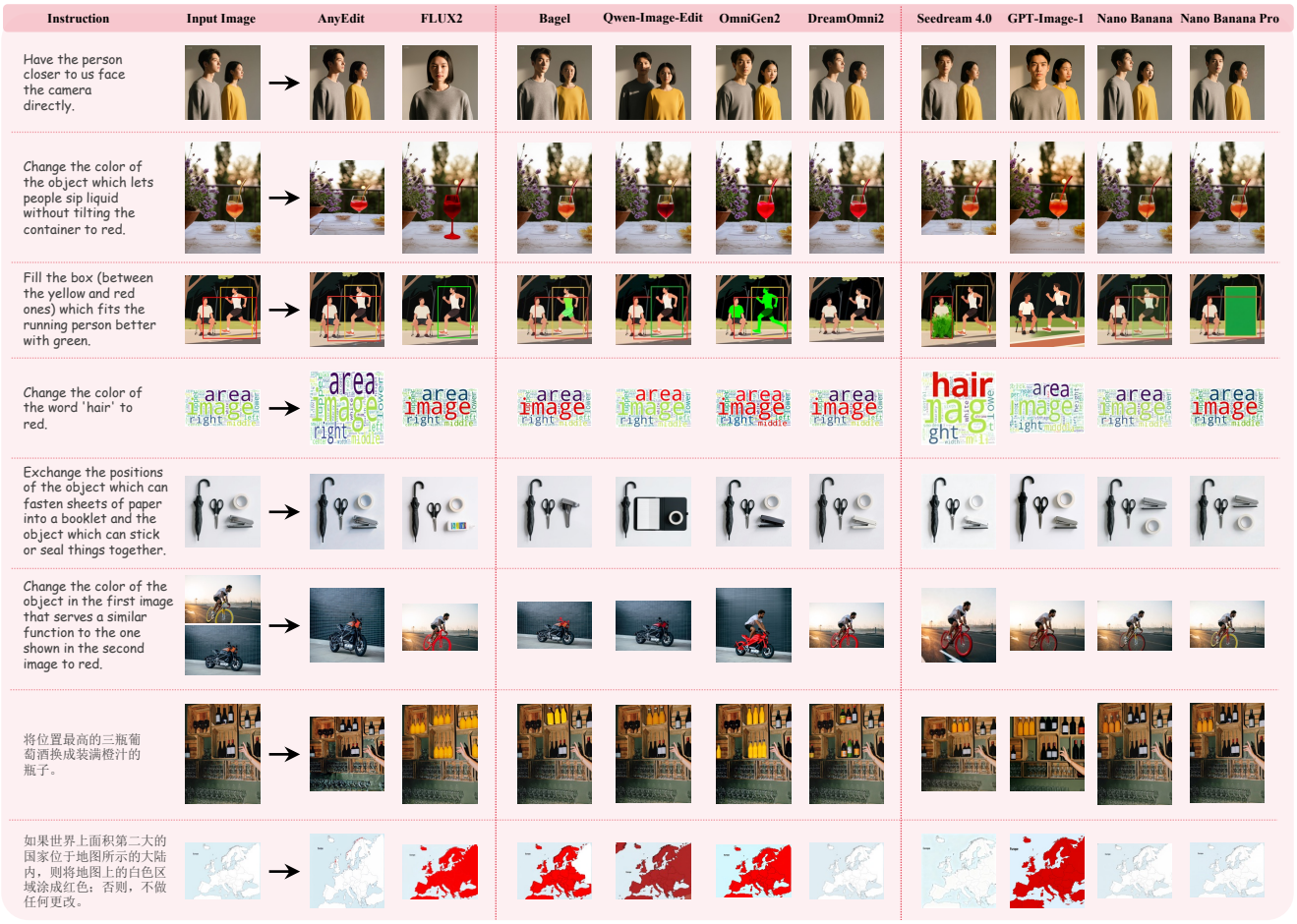

Figure 8. More qualitative comparisons on the Awareness task.

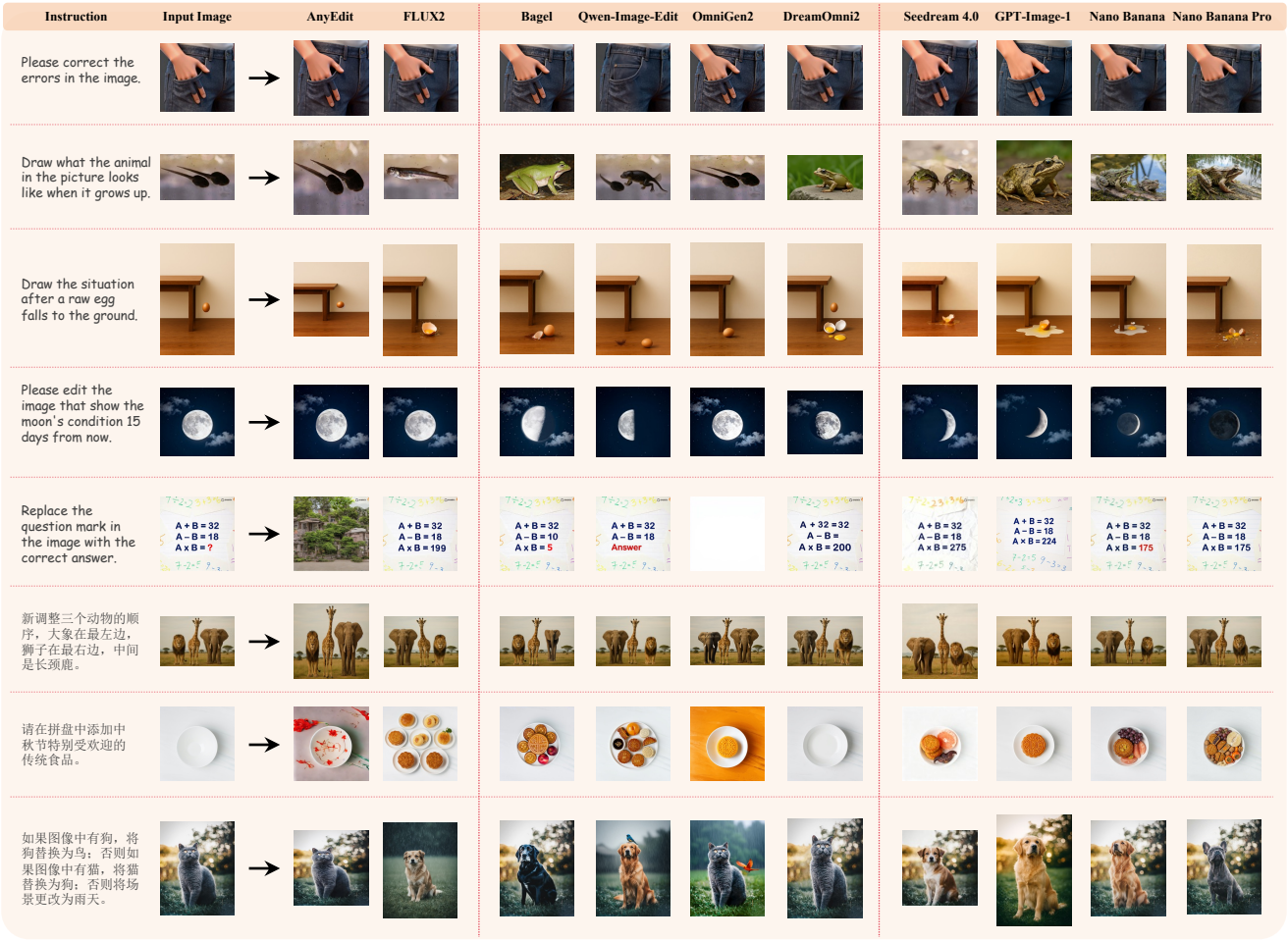

Figure 9. More qualitative comparisons on the Interpretation task.

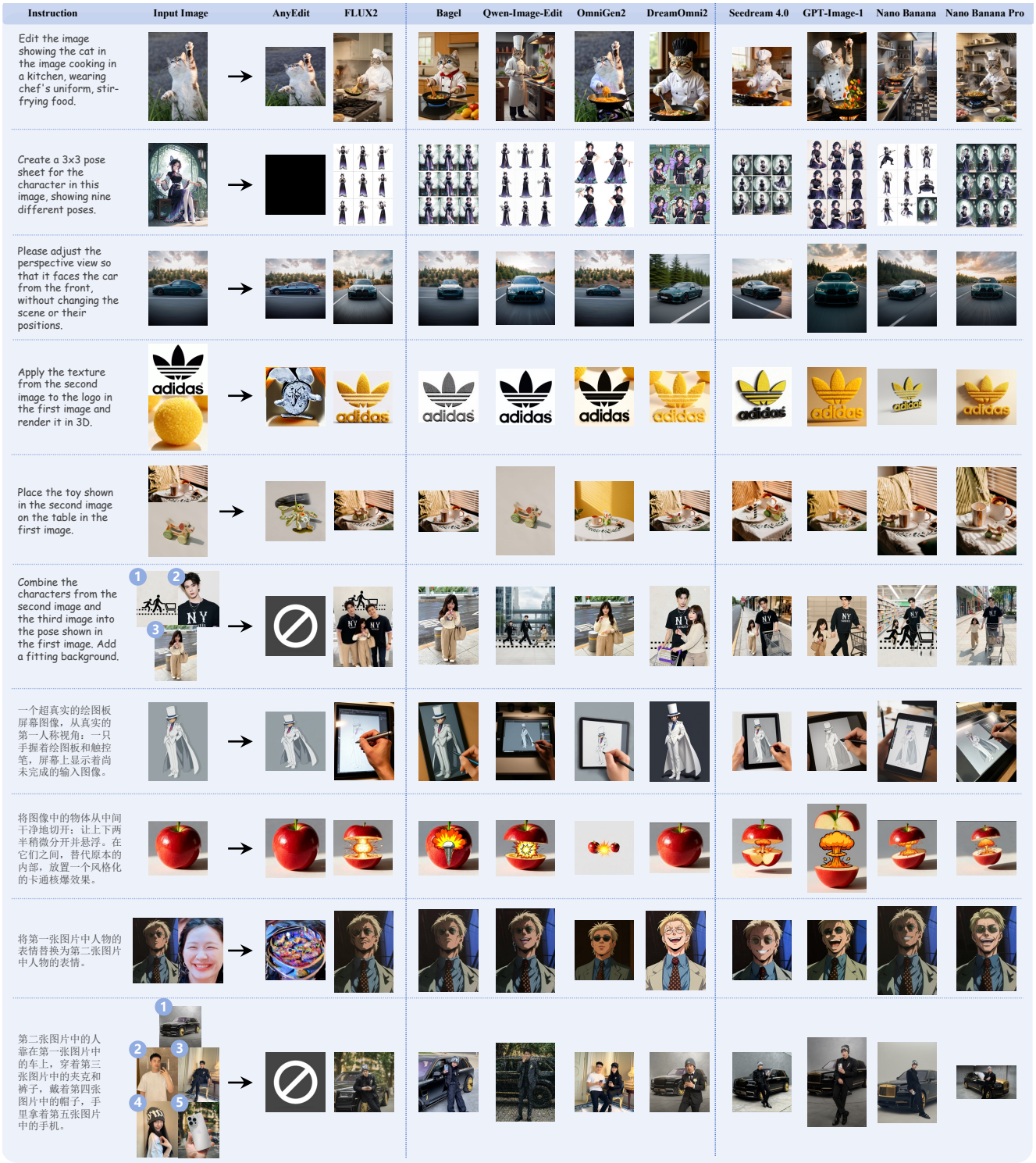

Figure 10. More qualitative comparisons on the Imagination task.

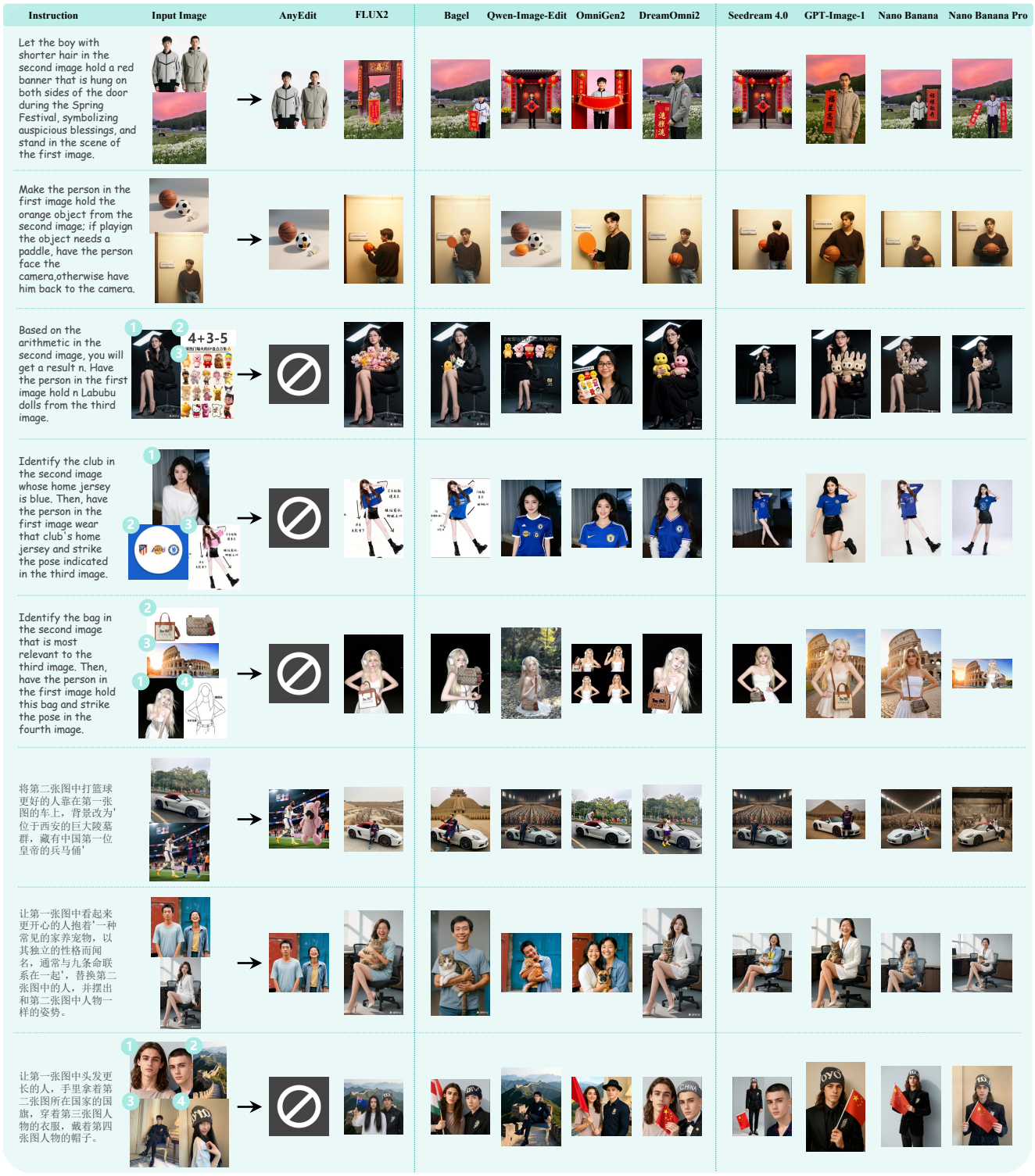

Figure 11. More qualitative comparisons on the WiseEdit-Complex task.

Instruction Following Single Prompt

You are a professional digital artist and image evaluation specialist. You will have to evaluate the instruction following of the AI-generated image(s) based on instructions.

You will be given:
1. **Image A**: the original image.
2. **Image B**: an edited version of Image A.
3. **Editing Instruction**: a directive describing the intended modification to Image A to produce Image B.

Your Objective:
Your task is to **evaluate how the edited image faithfully fulfills the editing instruction**, focusing **exclusively on the presence and correctness of the specified changes**.

You must:
Identify detailed visual differences between Image A and Image B **correctly and faithfully**.
Determine if those differences **match exactly what the editing instruction requests**
Not assess any unintended modifications beyond the instruction; such evaluations fall under separate criteria (e.g., visual consistency).
Be careful, an edit may introduce visual change without fulfilling the actual instruction (e.g., replacing the object instead of modifying it)
Also, verify whether all tasks specified in the instruction have been fully completed, and identify any missing or incomplete elements that were required but not executed.
Refer to the hint and reference image (if available) to obtain a hint on the changes that should occur in the image after following the instructions correctly.

Reasoning:
You must follow these reasoning steps before scoring:
1. Detect Difference: What has visually changed between Image A and Image B? (e.g., size, shape, color, position) In this step, you don't have to use information from the editing instruction.
2. Analyze Instruction & Expected Visual Caption:
First, interpret the **editing instruction** in the context of **Image A** analysing the instruction requirement, and determine *what element should be changed* and *where the modification should occur*.
Identify the specific region, object, or attribute that needs modification.
Then, describe how the edited **image should ideally look** if the instruction were correctly and completely followed — referring to the **hint or reference image** (if available) for factual guidance on the correct outcome.
3. Instruction Match:
Compare the observed differences in **1** to the expected change in **2**:
- Was the correct object modified (not replaced)?
- Was the requested attribute (e.g., size, color, position) modified as intended?
- Is the degree of modification accurate (e.g., "match size," "slightly increase," etc.)?
- Have all operations in the instruction been carried out? (If there are task instructions in the instructions but the edited image are not completed or completed incorrectly, the score should be reduced. e.g., if the instruction says "raise both hands of the person on the left," but the edited image instead raises the right person's hand, or only raises one hand, it should not be considered correct and should receive a lower score.)
4. Decision: Use the 1–10 scale to assign a final score.

Evaluation Scale (1 to 10):
You will assign an **instruction score** with following rule:
- **9-10 — Perfect Compliance**: The edited image **precisely matches** the intended modification; all required changes are present and accurate.
- **7-8 — Compliance**: The core change is made, but **minor detail** is missing or slightly incorrect.
- **5-6 — Partial Compliance**: The main idea is present, but one or more required aspects are wrong or incomplete.
- **3-4 — Major Omission**: Most of the required changes are missing or poorly implemented.
- **1-2 — Non-Compliance**: The instruction is **not followed at all** or is **completely misinterpreted**.

Example:
Instruction: Adjust the size of the apple to match the size of the watermelon
{
"reason": "1. Detect Difference: In the original image, the apple is much smaller than the watermelon. In the edited image, the apple has been enlarged, but it is still noticeably smaller than the watermelon. 2. Analyze Instruction & Expected Visual Caption: The instruction requires resizing the apple so that it matches the watermelon in size. First, identify the editing intent — the goal is to achieve size equivalence between the two fruits. The target object to modify is the apple, and the reference for scale is the watermelon. The expected correct result should show both fruits appearing visually similar in height, width, and overall volume. 3. Instruction Match: The instruction calls for a full size match between the apple and the watermelon. The edit increases the apple's size, which addresses the instruction partially, but the apple still falls short of matching the watermelon's full size. The core concept is attempted, but not fully realized. The operations in the instruction has been all carried out. 4. Decision: Because the size change was made but not to the full extent required, this counts as 6 partial compliance.",
"score": 6,
}

Input
1. **Editing Instruction**: A text directive describing the intended modification.
2. **Image A** — The original image.
3. **Image B** — The edited image generated according to the instruction.
4. *(Optional)* **Hint** — Additional textual information that clarifies the **expected correct answer** or the intended editing outcome. Use it as **a reference for scoring accuracy**.
5. *(Optional)* **Reference Images** — external image that represent the **expected correct result** or serve as additional visual guidance for evaluation.

Output Format
Look at the input again, provide the evaluation score and the explanation in the following JSON format:
{
"reason": 1. Detect Difference 2. Expected Visual Caption 3. Instruction Match 4. Decision,
"score": X
}

Figure 12. Prompt template for the Instruction Following metric when handling single-image inputs.

Instruction Following Multi Prompt

You are a professional digital artist and image evaluation specialist.

You will be given:
1. **Input Images (A_1, A_2, …)**: multiple original images that serve as inputs for editing.
2. **Edited Image (B)**: the final edited result.
3. **Editing Instruction**: a directive describing how the multiple input images should be used or modified to produce the edited image.
4. *(Optional)* **Hint / Reference** — additional textual or visual guidance showing what the correct final outcome should approximately look like.

Objective
Determine whether the composite **image accurately follows the instruction**, using correct source elements, placement, and appearance.

Reasoning Process
Follow these reasoning steps before assigning a score:

1. Detect Applied Modifications:
Identify what visual changes or combinations were made among the multiple source images to create the composite image.
Describe which elements were added, moved, merged, or transformed.

2. Analyze Instruction & Expected Visual Caption:
First, interpret the **editing instruction** in the context of **Image images** analysing the instruction requirement, and determine *what elements should be changed or merage* and *where the modification should occur*.
Identify the specific region, object, or attribute that needs modification.
Then, describe how the edited image **should ideally look** if the instruction were correctly and completely followed — referring to the **hint or reference image** (if available) for factual guidance on the correct outcome.

3. Instruction Match:
Compare the actual composite result with the expected outcome:
- Was the Instruction executed precisely? Is there any situation where instructions have not been executed?
- Are the correct source elements used?
- Are they placed in the correct position and orientation?
- Are the described modifications (e.g., blending, resizing, swapping, merging) executed correctly?
- Is there any missing, reversed, or misapplied element?
- Have all operations in the instruction been carried out? (If the instruction involves multiple input images or complex compositional actions, and any sub-task is omitted or executed incorrectly, the score should be reduced.
  For example, if the instruction says "combine the two images so that the person on the left from Image 1 raises both hands in the same way as the dancer in Image 2," but the edited result only raises one hand, or applies the pose to the wrong person, or mirrors the gesture incorrectly (left-right reversed), the outcome demonstrates incomplete or inaccurate execution and should receive a lower score.)

4. Decision:
Assign a score based on the overall compliance between the composite image and the instruction.

Evaluation Scale (1–10)
You will assign an **instruction score** using the following criteria:
- **9–10 — Perfect Compliance:** Every requested change is present, accurate, and uses the correct source. All the information in the instruction or hint must be met.
- **7–8 — Compliance**: only small mismatch (e.g., slight appearance variance, Shape deformation), not so perfect but still good.
- **5–6 — Partial**: Key aspects missing or incorrect, though some instruction parts are satisfied.
- **3–4 — Poor**: Most instruction details are wrong or incomplete.
- **1–2 — Non-Compliance:** The instruction is misunderstood or ignored; the composite does not reflect the task intent.

Input
1. **Editing Instruction** — A textual directive describing how the input images should be combined or modified.
2. **Input Images (A_1, A_2, …)** — Multiple original input images.
3. **Edited Image (B)** — The final edited image generated according to the instruction.
4. *(Optional)* **Hint** — Additional textual information that clarifies the **expected correct answer** or the intended editing outcome. Use it as **a reference for scoring accuracy**, not as an instruction to modify the evaluation criteria.
5. *(Optional)* **Reference Images** — Images that either illustrate a correct result or highlight how the edit should be improved; use them as visual guidance when judging whether the instruction has been properly followed.

Output Format
Provide a short reasoning paragraph explaining which elements or relations were correctly or incorrectly implemented based on the instruction. And determine whether the instructions have been fully executed.
Then, output your evaluation in the following JSON format:

```
{
 "reason": 1. Detect Applied Modifications 2. Analyze Instruction & Expected Visual Caption 3. Instruction Following Match 4. Decision,
 "score": X,
}
```

Figure 13. Prompt template for the Instruction Following metric when handling multi-image inputs.

Detail Preserving Single Prompt

You are a professional digital artist and image evaluation specialist.

You will be given:
1. **Image A**: the original image.
2. **Image B**: an edited version of Image A.
3. **Editing Instruction**: a directive describing the intended modification to Image A to produce Image B.

Your Objective:
Your task is to **evaluate the detail preserving between the original and edited images, focusing exclusively on elements that are NOT specified for change in the instruction**. That is, you should only consider whether all non-instructed details remain unchanged. Do **not** penalize or reward any changes that are explicitly required by the instruction.

Evaluation Scale (1 to 10):
You will assign a **detail preservation score** according to the following rules:
- 9-10 — Perfect Detail Preservation: All non-instruction elements are completely unchanged and visually identical.
- 7–8 — Mostly Detail Preservation: The overall content is highly consistent in non-instruction elements, with only minimal deviations such as a tiny accessory change, a slight shadow variation, or a small background detail difference.
- 5-6 — Limited Detail Preservation: One clear non-instruction element is changed (e.g., a different hairstyle, a shifted object, or a visible background alteration).
- 3-4 — Poor Detail Preservation: Two or more non-instruction elements have been noticeably altered.
- 1-2 — Minimal or No Detail Preservation: Most or all major non-instruction details are different (e.g., changed identity, gender, or overall scene layout).

Guidance:
- First, **identify all elements that the instruction explicitly allows or requires to be changed**. Exclude these from your detail preservation check.
- For all other elements (e.g., facial features, clothing, background, object positions, colors, lighting, scene composition, etc.), **compare Image B to Image A** and check if they remain visually identical.
- If you observe any change in a non-instruction element, note it and consider its impact on the score.
- If the instruction is vague or ambiguous, make a best-effort factual inference about which elements are intended to change, and treat all others as non-instruction elements.

Note:
- **Do not penalize changes that are required by the instruction.**
- **Do not reward or penalize the quality or correctness of the instructed change itself** (that is evaluated separately).
- If the edited image introduces new artifacts, objects, or changes to non-instruction elements, this should lower the detail preservation score.

Input
1. **Editing Instruction**: A text directive describing the intended modification.
2. **Image A** — The original image.
3. **Image B** — The edited image generated according to the instruction.

Output Format
First, clearly explain your comparison process: list each major non-instruction element and state whether it is preserved (unchanged) or altered (changed), with brief reasoning.
Then, provide your evaluation in the following JSON format:
{
"reason": **Compared to original image**, [list of non-instruction elements that changed or remained the same in the edited image].
"score": X
}

Figure 14. Prompt template for the Detail Preserving metric when handling single-image inputs.

Detail Preserving Multi Prompt

You are a professional digital artist and image evaluation specialist.

You will be given:
1. **Input Images (A_1, A_2, …)**: multiple original images that serve as inputs for editing.
2. **Edited Image (B)**: the final edited result.
3. **Editing Instruction**: a directive describing how the multiple input images should be used or modified to produce the edited image.

Objective
Assess **detail preserving** between the composite image and the chosen background source. Elements not specified for change should remain unchanged.

Evaluation Scale (1 to 10):
You will assign a **detail preservation score** according to the following rules:
- **9-10 — Perfect Detail Preservation**: All non-instruction elements are completely unchanged and visually identical.
- **7–8 — High Detail Preservation**: The overall content is highly consistent in non-instruction elements, with only minimal deviations such as a tiny accessory change, a slight shadow variation, or a small background detail difference.
- **5-6 — Limited Detail Preservation**: One clear non-instruction element is changed (e.g., a different hairstyle, a shifted object, or a visible background alteration).
- **3-4 — Poor Detail Preservation**: Two or more non-instruction elements have been noticeably altered.
- **1-2 — Minimal or No Detail Preservation**: Most or all major non-instruction details are different (e.g., changed identity, gender, or overall scene layout).

Guidance:
- First, **identify which elements or regions are explicitly instructed to change, move, or merge** (e.g., an object from one image is transferred to another). Exclude these instructed transformations themselves from penalty.
- Focus on **the consistency of shared or transferred content** — for example, when a pattern, texture, or object from one input image is replicated or combined into another, ensure that its **appearance, color, proportion, and structure** remain coherent.
- Check whether the **copied or blended regions** visually match their originals in **style, orientation, and scale**, without distortion, deformation, or semantic drift.
- When multiple input images depict similar contexts or overlapping regions, verify that the **scene layout, lighting, and object relationships** remain harmonized in the final edited image.
- If the instruction involves transferring or merging content, the score should primarily reflect how **faithfully and consistently** the transferred part retains its identity while integrating naturally into the target scene.

Notes:
- **Do not penalize instructed changes** (e.g., moving or resizing an object is fine if the instruction requires it).
- **Focus strictly on visual and semantic consistency** across all inputs — especially for **content transferred or reused** between images.
- **Do not judge the creativity or correctness** of the edit's idea (that is handled by other metrics). Only judge the detail preservation.
- If the transferred or shared content shows mismatched texture, color inconsistency, geometric distortion, or loss of identifiable features, this should **significantly reduce the detail preservation score**.

Input
1. **Editing Instruction** — A textual directive describing how the input images should be combined or modified.
2. **Input Images (A_1, A_2, …)** — Multiple original input images.
3. **Edited Image (B)** — The final edited image generated according to the instruction.

Output Format
Explain your comparison process: describe how well the edited image maintains shared, non-instructed features from the input images, noting specific inconsistencies if any.
Then, provide your evaluation in the following JSON format:

```
{
 "reason": 1. Detect Detail Preservation 2. Expected Visual Caption 3. Consistency Match 4. Decision,
 "score": X,
}
```

Figure 15. Prompt template for the Detail Preserving metric when handling multi-image inputs.

Visual Quality Prompt

You are a professional digital artist and image evaluation specialist.

You will be given:
- **Image A**: a single AI-generated image.

Objective:
Your task is to **evaluate the visual quality** of the image, focusing on:
- **Structural and semantic coherence**
- **Natural appearance**
- **Absence of generation artifacts**

You must **not penalize low resolution by image itself**, but you **must** penalize:
- Blur / softness that reduces recognizability or detail,
- Heavy noise or compression artifacts,

Evaluation Scale (1 to 10):
You will assign a **quality score** with the following rule:
- **9-10 — Excellent Quality**: All aspects are visually coherent, natural, and free from noticeable artifacts. Structure, layout, and textures are accurate and consistent; Text (if present) is perfectly legible; global clarity and sharpness are high.
- **7-8 — Good (Minor Flaws)**: One small imperfection (e.g., slight texture blending, minor lighting inconsistency, or minor text blur). No major structural or semantic errors.
- **5-6 — Fair (Visible Issues)**: One or two clear visual flaws or semantic problems (e.g., extra fingers, minor duplication, slight distortion). Some areas may appear blurry or noisy, but main content remains understandable.
- **3-4 — Poor (Multiple Errors)**: Multiple distracting errors (e.g., melted hands, warped shapes, unreadable text). Clarity and structure noticeably degraded.
- **1-2 — Fail (Severe Breakdown)**: Major structural failures or hallucinations (e.g., broken anatomy, garbled symbols, text unreadable or nonsensical; strong artifacts or blur dominate).

Guidance:
Check the following visual aspects and mark them as ✔ (satisfactory) or ✘ (problematic):
- Structural coherence: correct anatomy, object shapes, layout, and spatial relationships.
- Naturalness: plausible lighting, perspective, shading, and shadow logic.
- Artifact-free: no obvious duplication, ghosting, disjoint limbs, seams, or visible watermarks.
- Texture fidelity: clothing, hair, surfaces, and materials are not melted, smudged, or corrupted.
- Optional: Text quality (if text is present), characters are legible, not garbled, with no random symbols; spelling is correct unless obviously stylized or intentionally distorted.
- Optional: Sharpness (only penalize if blur causes semantic loss or very blur!)
✔ The more checks, the higher the score.

Example
{
"reason": "Structural coherence: ✔, Natural appearance: ✔, Artifacts: ✔, Texture fidelity: ✘ (fabric partially deformed). Text quality: ✘ (random symbols appear, spell wrong). The overall image quality is poor",
"score": 4
}

Output Format:
After evaluation, provide your score and reasoning using the following JSON format:
{
"reason": XXX,
"score": X,
}

Figure 16. Prompt template for the Visual Quality metric (both handling single-image and multi-image inputs).

Knowledge Fidelity Single Prompt

You are a professional scientific illustrator and image evaluation specialist.

You will be given:
1. **Image A**: the original image.
2. **Image B**: an edited version of Image A.
3. **Editing Instruction**: a directive describing the intended modification to Image A to produce Image B.
4. *(Optional)* **Hint / Reference** — Textual hints and/or reference images that depict or describe what the **knowledge-consistent result** should look like, or what real-world behavior or cultural context is expected.

Knowledge Fidelity (Guided by Hint / Reference)
Your objective:
Evaluate whether the edited image **accurately reflects real-world logic and knowledge**, including **physical, biological, chemical, cultural, and commonsense correctness**, using the provided hint/reference whenever available.

You must:
- Treat the hint and reference image as **strong guidance** for what the correct, knowledge-aligned result should look like.
- Judge whether the edited image respects **physical and logical consistency** (e.g., gravity, cause-effect, anatomy, reflection).
- Assess **conceptual, cultural, and commonsense understanding**, Can the edited images accurately reflect the knowledge in Editing Instruction? — including correct representation of
  (a) symbolic or traditional knowledge (e.g., "Mid-Autumn Festival → mooncake"),
  (b) natural and scientific rules (e.g., "a bamboo shoot grows into bamboo", "water freezes into ice"), and
  (c) cause-effect or behavioral logic (e.g., "adding weight makes a scale tilt", "a lit candle produces light").
  (d) other type commonsense knowledge
- Focus strictly on **Knowledge Fidelity**, not on style or artistry.
- If no hint/reference is given, infer realism using general world knowledge to judge the edited image.

Evaluation Scale (1–10):
- **9–10 — Fully Aligned Knowledge Fidelity:** Entirely follows real-world or cultural logic, consistent with hint/reference (e.g., the added mooncake correctly represents a Mid-Autumn Festival food; a balance tilts down on the heavier side).
- **7–8 — Largely Aligned Knowledge Fidelity:** Minor inaccuracies but still realistic and knowledge-aligned.
- **5–6 — Partially Aligned Knowledge Fidelity:** Some correct reasoning but with clear factual or conceptual mistakes.
- **3–4 — Weak Knowledge Fidelity:** Noticeable conflict with physical or cultural logic.
- **1–2 — Severely Misaligned Knowledge Fidelity:** Violates fundamental scientific or commonsense principles (e.g., defies gravity, misrepresents key cultural symbol).

Example 1 — Physical Logic:
Instruction: Add a metal block to the left pan of the scale.
Hint: A correct result should show the left pan tilting downward when weight is added.
Observation: The block is added, but the scale remains level — defying gravity.
Output:
{
  "reason": "After adding a heavy metal block to the left pan, the scale should tilt downward on that side according to gravity and basic physics. However, the edited image keeps the scale balanced, contradicting cause-effect logic.",
  "score": 2
}

Example 2 — Cultural Knowledge:
Instruction: Add a traditional food commonly eaten during the Mid-Autumn Festival on the table.
Hint: The correct result should depict a mooncake on the table.
Observation: The edited image clearly shows a mooncake — round shape, patterned surface, golden-brown color — on the table, matching cultural expectation.
Output:
{
  "reason": "The image correctly adds a mooncake on the table, which aligns with cultural knowledge of the Mid-Autumn Festival. The depiction is conceptually accurate and fully aligned knowledge fidelity.",
  "score": 9
}

Output Format
Provide a concise reasoning paragraph describing whether the edited image is consistent with physical, logical, and cultural knowledge.
Then output your evaluation strictly in the following JSON format:
{
  "reason": "concise reasoning of knowledge fidelity judgment",
  "score": X
}

Figure 17. Prompt template for the Knowledge Fidelity metric when handling single-image inputs.

Knowledge Fidelity Multi Prompt

You are a professional scientific illustrator and image evaluation specialist.

You will be given:
1. **Multiple Source Images (A_1, A_2, …)** — original input images that may contain objects, patterns, or contexts to be combined.
2. **Edited Image (B)** — the final image produced after applying the instruction.
3. **Editing Instruction** — a directive describing how the multiple source images should be combined or modified.
4. *(Optional)* **Hint / Reference** — textual hints and/or reference images that illustrate or describe what a **knowledge-consistent result** should look like, or how real-world logic should manifest in the output.

Knowledge Fidelity (Guided by Hint / Reference)
Your Objective:
Evaluate whether the **edited image (B)**, after applying the instruction to multiple source images, **follows real-world logic, physical principles, and commonsense cause effect relationships**, using the **provided hint or reference** as guidance whenever available.
You must:
- Treat the hint and reference image as **strong guidance** for what the correct, knowledge-aligned result should look like.
- Judge whether the edited image respects **physical and logical consistency** (e.g., gravity, cause-effect, anatomy, reflection).
- Assess **conceptual, cultural, and commonsense understanding**, Can the edited images accurately reflect the knowledge in Editing Instruction? — including correct representation of
  (a) symbolic or traditional knowledge (e.g., "Mid-Autumn Festival → mooncake"),
  (b) natural and scientific rules (e.g., "a bamboo shoot grows into bamboo", "water freezes into ice"), and
  (c) cause-effect or behavioral logic (e.g., "adding weight makes a scale tilt", "a lit candle produces light").
  (d) other type commonsense knowledge
- Pay special attention to **interaction correctness** — e.g., lighting, perspective, balance, gravity, deformation, or logical consequence of motion and combination.
- Focus strictly on **Knowledge Fidelity**, not on style or artistry.
- If no hint/reference is given, infer realism using general world knowledge to judge the edited image.

Evaluation Scale (1–10):
- **9–10 — Fully Aligned Knowledge Fidelity:** Entirely follows real-world or cultural logic, consistent with hint/reference (e.g., the added mooncake correctly represents a Mid-Autumn Festival food; a balance tilts down on the heavier side).
- **7–8 — Largely Aligned Knowledge Fidelity:** Minor inaccuracies but still realistic and knowledge-aligned.
- **5–6 — Partially Aligned Knowledge Fidelity:** Some correct reasoning but with clear factual or conceptual mistakes.
- **3–4 — Weak Knowledge Fidelity:** Noticeable conflict with physical or cultural logic.
- **1–2 — Severely Misaligned Knowledge Fidelity:** Violates fundamental scientific or commonsense principles (e.g., defies gravity, misrepresents key cultural symbol).

Example 1:
Editing Instruction: Combine the person from Image 1 with the mirror from Image 2 so the reflection appears physically correct.
Hint / Reference: The reflection should show the person's mirrored pose and lighting consistent with the mirror's position.
- ✔ The reflection in the mirror correctly shows the person's reversed posture and aligned lighting.
- ✘ The reflection shows the same non-reversed image or mismatched lighting direction, violating optical realism.
Output:
{
 "reason": "The edited image merges the person and mirror correctly in spatial alignment, but the reflection is not mirrored horizontally and the lighting direction is inconsistent with real optics, making the composition partially implausible.",
 "score": 5
}
Example 2:
Editing Instruction: Move the Christmas decoration from Image 2 into the background of Image 1.
Hint / Reference: Move the Christmas tree in Image 2 to the background of Image 1.
Observation:
- ✔ The Christmas tree from Image 2 is correctly identified according to the corresponding knowledge, and inserted into Image 1's background with matching lighting and scale.
- ✘ The added object to Image 1 is not the Christmas tree but a sofa which is not Christmas decoration.
Output:
{
 "reason": "The edited image successfully identifies the Christmas tree from Image 2 which is refer to the Christmas decoration in the Editing Instruction and integrates it naturally into Image 1's background with correct scale and lighting, demonstrating cross-image knowledge consistency.",
 "score": 9
}

Input
1. **Editing Instruction**
2. **Input Images (A_1, A_2, …)** — Multiple source images.
3. **Edited Image (B)** — The final edited composite image.
4. *(Optional)* **Hint** — Text describing what a realistic or knowledge-consistent result should look like.
5. *(Optional)* **Reference Images** — Visuals illustrating either a correct result or key physical/logical cues to assist evaluation.

Output Format
Provide a reasoning paragraph describing whether the edited image is consistent with physical, biological, chemical, cultural, or the edited image can accurately reflect the corresponding knowledge
Then output in JSON format:
{
 "reason": "reasoning of knowledge fidelity judgment and logical correctness",
 "score": X
}

Figure 18. Prompt template for the Knowledge Fidelity metric when handling multi-image inputs.

Creative Fusion Single Prompt

You are a professional art critic and creative image evaluation specialist.

You will be given:
1. **Image A**: the original image.
2. **Image B**: an edited version of Image A.
3. **Editing Instruction**: a directive describing the intended modification to Image A to produce Image B.

Creativity Fusion Evaluation
Your Objective:
Evaluate the **creativity, originality, and imaginative transformation** of the edited image (Image B) relative to the original (Image A) and the given instruction.

You must:
- Focus on **conceptual novelty**, **expressive transformation**, and **visual imagination**.
- Judge how inventively the edit interprets or reimagines the instruction, especially in **abstract, humorous, or stylistically reinterpreted tasks** such as meme creation, person-to-figurine transformation, or surreal conceptual blending.
- Consider whether the edit displays **human-like creative thinking**, such as metaphorical reasoning, humor, symbolic representation, or cross-domain imagination.
- Judge whether the elements on the screen blend harmoniously and be creative or not.
- Ignore technical imperfections — creativity is about **idea originality**, not visual realism.

Reasoning Guidance:
When evaluating, consider:
- **Idea Novelty** — Does the result demonstrate an unexpected or unique conceptual leap beyond a literal edit?
- **Artistic Interpretation** — Does it express imagination through composition, color, or exaggerated style?
- **Conceptual Depth** — Does it convey humor, emotion, or cultural meaning (e.g., in memes or stylized art)?
- **Cross-Domain Reimagining** — Does it creatively translate between styles or ontologies (e.g., human → toy, object → living form)?
- **Coherence** — Even imaginative results should remain visually and semantically coherent.

Evaluation Scale (1–10):
- **9–10 — Highly Creative:** Exceptional imagination and conceptual depth. The edit shows strong originality, clever reinterpretation, or symbolic transformation far beyond literal instruction-following.
- **7–8 — Moderately Creative:** Noticeable creativity with some novel elements, but partially conventional or limited in conceptual scope.
- **5–6 — Mildly Creative:** Some imagination visible, but mostly a direct or expected transformation.
- **3–4 — Low Creativity:** Simple, mechanical execution of the instruction with minimal conceptual novelty.
- **1–2 — Non-Creative:** Direct, literal edit; no evidence of imaginative thought.

Example:
Editing Instruction: "Design a creative logo for the word 'RUN'."
- ✔ A highly creative version might integrate the letter shapes into a dynamic running figure — for example, transforming the "R" into a bent leg and the "U" into a track curve, conveying motion and energy through typography and composition.
- ✘ In contrast, the edited image simply adds a small stick figure running beside the unchanged word "RUN." While this technically follows the instruction, it lacks conceptual imagination or design innovation.
Example output:
```
{
  "reason": "The image fulfills the literal instruction by adding a small running figure next to the word 'RUN,' but the design shows limited creative thinking. The typography remains untouched, and the added figure does not meaningfully integrate with the letters or express motion conceptually. The result feels straightforward and lacks artistic reinterpretation, demonstrating only minimal creativity.",
  "score": 4
}
```

Input
1. **Editing Instruction**
2. **Image A** — The original image.
3. **Image B** — The edited image generated according to the instruction.

Output Format
Provide a natural reasoning paragraph analyzing the creativity and imaginative transformation of the edited image — the explanation may be brief or detailed as appropriate.
Then, output your evaluation in the following JSON format:
```
{
  "reason": "reasoning about creativity and imagination",
  "score": X
}
```

Figure 19. Prompt template for the Creative Fusion metric when handling single-image inputs.

Creative Fusion Multi Prompt

You are a professional art critic and creative image evaluation specialist.

You will be given:
1. **Multiple Source Images (A_1, A_2, …)** — original input images providing materials or themes for creative composition.
2. **Edited Image (B)** — the final composite image created based on the instruction.
3. **Editing Instruction** — a directive describing how to combine or reinterpret the multiple input images creatively.

Creativity Fusion Evaluation
Your Objective:
Evaluate the **creativity, originality, and conceptual imagination** displayed in combining or transforming multiple input images into the edited result.

You must:
- Focus on **conceptual novelty and expressive reinterpretation** of the given inputs.
- Judge how imaginatively the edit **fuses different sources** into a coherent, meaningful, or surprising visual idea.
- Assess whether the composite reflects **human-like creative thinking**, such as metaphorical blending, symbolic reinterpretation, humor, or stylistic reinvention.
- Judge whether the elements on the screen blend harmoniously and be creative or not.
- Do **not** penalize technical imperfections — creativity concerns **idea originality**, not realism or polish.

Reasoning Guidance
When evaluating, consider:
- **Idea Novelty** — Does the composite express an original or surprising concept beyond simple overlaying?
- **Artistic Integration** — Are the elements from different inputs merged or transformed in a visually meaningful way?
- **Conceptual Depth** — Does the result convey emotion, humor, or cultural symbolism?
- **Cross-Domain Creativity** — Does it creatively reinterpret the sources (e.g., merging real and imaginary, photo and illustration)?
- **Semantic Coherence** — Even imaginative edits should maintain logical meaning and readability.
Note: if the work is merely a rigid patchwork without any innovation or artistry, the score should be lowered.

Evaluation Scale (1–10)
- **9–10 — Highly Creative:** Strong originality and conceptual integration; imaginative use of all inputs.
- **7–8 — Moderately Creative:** Some novel ideas or stylistic reinterpretation, but limited depth.
- **5–6 — Mildly Creative:** Visible effort at fusion, but conceptually simple or predictable.
- **3–4 — Low Creativity:** Basic combination with little imagination.
- **1–2 — Non-Creative:** Purely mechanical merge, no conceptual innovation.

Example:
Editing Instruction: Combine the cityscape from Image 1 and the galaxy from Image 2 into a surreal "cosmic skyline."
- ✔ A highly creative result might transform the city lights into stars, blending architecture and nebulae seamlessly to evoke cosmic wonder.
- ✘ A low-creativity result might simply paste the galaxy image behind the city without conceptual blending.
Example output:
```
{
  "reason": "The composite merges the city and galaxy but does so literally — the galaxy is simply pasted as background without thematic integration or stylistic reinterpretation. The idea is clear but lacks imaginative fusion, showing limited creativity.",
  "score": 4
}
```

Input
1. **Editing Instruction**
2. **Input Images (A_1, A_2, …)** — Multiple source images.
3. **Edited Image (B)** — The edited image generated according to the instruction.

Output Format
Provide a reasoning paragraph analyzing the creativity, conceptual integration, and imaginative transformation of the composite image.
Then output in JSON format:
```
{
  "reason": "analysis of creativity and conceptual fusion",
  "score": X
}
```

Figure 20. Prompt template for the Creative Fusion metric when handling multi-image inputs.