I2I-Bench: A Comprehensive Benchmark for Image-to-Image Editing Models Appendix

Juntong Wang*, Jiarui Wang*, Huiyu Duan*, Jiaxiang Kang, Guangtao Zhai, Xiongkuo Min†
Institute of Image Communication and Network Engineering
Shanghai Jiao Tong University, Shanghai, China

1. Model Details

1.1. Evaluated Models

We evaluated 9 Single-Image (SE) editing models and 4 Multi-Image (ME) editing models, as shown in the main paper. For nano-banana, we obtained evaluation results through its official API interface. For all other models (including Qwen-Image-Edit-2509, Bagel, Step1X-Edit, UniPic-2, FLUX.1-Kontext-dev, Omnigen2, ICEdit, GoT, instruct-pix2pix, and DreamOmni2), we used their publicly available checkpoints. During inference, we uniformly adopted the default inference configurations provided in the models' official repositories or `diffusers` library. No additional parameter tuning was performed, ensuring a fair and standardized evaluation.

1.2. Evaluation Tools

Our evaluation pipeline utilizes both generalist and specialist models. For the Generalist Evaluator, all LMM-based evaluation pipelines detailed in Appendix 5 (LMM VQA-5Level, LMM VQA-2Level, and LMM Multi-Question VQA) were conducted using Qwen3-VL-8B-Instruct. For the Specialist Tools, we utilized Q-Insight, ArtiMuse, GOT-OCR2.0, Sa2VA, and DINOv3, all employed with their default public settings.

2. Human Preference Annotation Details

To rigorously validate the alignment of I2I-Bench's automated evaluation methods (especially the LMM evaluators) with genuine human perception, we conducted a large-scale human preference annotation experiment.

Annotator Guidelines. We employed a pairwise comparison format. Annotators were shown images generated by two different models (*e.g.*, Model A and Model B) for the same prompt and dimension, and were asked to select "A is better," "B is better," or "Tie." The most critical principle was that annotators were strictly instructed to judge *solely based on the single dimension being evaluated*, and to disregard all other quality aspects. For example, when evaluating "Subject Identity Fidelity," annotators were told: "You must choose the image that better preserves the subject's identity (*e.g.*, face, features). Even if the other image has better blending or fewer artifacts, you must penalize it if the subject's identity is distorted." Conversely, for "Blending Naturalness," the guidance was: "You must focus only on whether the transition of the edited region is smooth and seamless. Even if the image did not follow the instruction perfectly, you should choose it if its blending is superior." We provided detailed manuals with positive and negative examples for all 30 fine-grained dimensions to ensure a consistent understanding among all annotators.

Quality Assurance. To ensure the accuracy and consistency of the annotated data, we implemented a rigorous, multi-step quality assurance process. First, we (the authors) prepared clear definitions, criteria, and "what to look for" vs. "what to ignore" examples for all 30 dimensions. Second, before the main task, all annotators had to complete a "Pre-Labeling Trial" of approximately 30 pairwise comparison samples. Third, we reviewed these trial results and provided one-on-one feedback to annotators to clarify any misunderstandings and unify the standards. Fourth, we iterated on the guidelines, supplementing them with confusing cases found during the trial. Finally, after all annotations were complete, we (the authors) randomly sampled 20% of the total annotations from each dimension for post-labeling checks. If the error rate (disagreement with the authors) in this sample exceeded 10%, all data for that dimension was considered invalid and re-assigned to a different annotator for re-labeling. This strict training and QA process ensures our human preference data is highly reliable for alignment validation.

3. Rationale for Evaluation Methods

In I2I-Bench, we firmly contend that a single, monolithic evaluation method (*e.g.*, a "Pure LMM" score) is insufficient to capture the full spectrum of image editing quality. The 30 dimensions in our benchmark are decoupled, targeting distinct facets of quality ranging from objective technical fidelity to complex cognitive reasoning.

*Equal contribution
†Corresponding author

To achieve the highest possible alignment with human judgment, we designed a hybrid evaluation system that explicitly matches the evaluation needs of each dimension to the most appropriate tool. Our methodology is built on a "best-tool-for-the-job" principle, which can be categorized into three distinct evaluation pathways.

3.1. Specialist Models: For Objective and Perceptual Quantification

For well-defined, global image properties, mature specialist models provide the most stable, objective, and unbiased scores.

Dimensions: `Overall Image Quality`, `Aesthetic Quality`.

Method: Specialist Models (Q-Insight, ArtiMuse).

Rationale: These dimensions require assessing global, technical (*e.g.*, clarity, noise) or artistic (*e.g.*, composition, color harmony) quality. Unlike an LMM, whose judgment (as in our "Pure LMM" baseline) can be easily biased by the *semantic content* of an edit (*e.g.*, an LMM might give a high score to a "semantically correct" but blurry edit), these specialist tools are trained on specific, large-scale datasets (like Koniq-10k) to provide consistent quantitative scores for these precise attributes, aligning closely with human perception of these specific factors.

3.2. Hybrid LMM-Specialist Pipelines: For High-Fidelity and Content-Specific Tasks

This hybrid approach is a core innovation of I2I-Bench. It combines the semantic understanding of LMMs ("what to look at") with the precise quantification of specialist tools ("how to measure it"), overcoming the limitations of using either one alone.

3.2.1. Hybrid OCR: For Textual Accuracy

Dimensions: `Text Content & Style Accuracy`.

Method: Hybrid LMM VQA-5Level + Specialist OCR (GOT-OCR2.0).

Rationale: We do not use a "Pure LMM" (baseline) for this dimension due to its poor performance in precise OCR. LMMs frequently "hallucinate"—misreading, omitting, or inventing text. This dimension involves two distinct subtasks: (1) **Content:** Is the spelling correct? (2) **Style:** Are the font, position, and color correct?

- For **Content (1)**, a binary, objective task, the specialist OCR model (GOT-OCR2.0) provides a "ground truth" score for content accuracy (S_c).
- For **Style (2)**, a subjective, semantic judgment, the LMM is the ideal tool (via LMM VQA-5Level, yielding S_s).

Our hybrid pipeline (Eq. 5) uses the objective OCR score to "gate" the subjective LMM style score, ensuring a model does not receive a high score for generating beautifully styled but *incorrectly spelled* text.

Further Justification for the Fusion Strategy: We select the piecewise function in Eq.4 (from the main paper) based on a careful consideration of human perceptual mechanisms, rather than a simpler combination.

- **Inadequacy of Additive Fusion:** A simple additive combination (*e.g.*, $S_c + S_s$) is fundamentally unsuitable. It fails to implement a "gating" mechanism, meaning a perfect style score ($S_s = 1.0$) could erroneously compensate for completely incorrect content (*e.g.*, $S_c = 0.1$), leading to a high score for a failed edit.
- **Why Piecewise is Superior to Simple Multiplication:** As you correctly noted, a simple multiplicative fusion (*e.g.*, $S_c \times S_s$) does provide a basic gating effect. However, we found it insufficient as it fails to capture the *non-linear* nature of human perception. Our empirical observations suggest that human evaluators do not assess textual accuracy on a continuous linear scale. Instead, they tend to "bucket" the results into coarse-grained categories:
  1. Perfectly Correct ($S_c \approx 1.0$)
  2. Mostly Correct / Minor Error (*e.g.*, $S_c \geq 0.8$)
  3. Partially Correct / Significant Errors (*e.g.*, $S_c \approx 0.5$)
  4. Completely Wrong ($S_c < 0.3$)

  A simple multiplication treats the difference between $S_c = 0.9$ and $S_c = 0.8$ the same as the difference between $S_c = 0.6$ and $S_c = 0.5$. In contrast, our piecewise function is explicitly designed to model these discrete human perceptual thresholds, applying a gentle penalty for "mostly correct" results while applying a severe penalty once the accuracy drops below a "failure" threshold.
- **Disparity in Task Objectivity:** This design also accounts for the different nature of the sub-tasks. Rendering correct content (S_c) is an objective, difficult, and high-stakes task strictly measured by OCR. Rendering style (S_s) is a more subjective and, comparatively, lower-difficulty task evaluated by the LMM. The objective S_c score must therefore serve as a robust, non-negotiable filter for the subjective S_s score.
- **Empirical Validation:** The specific thresholds (*e.g.*, 0.8, 0.6) and their corresponding multipliers (0.8, 0.5, 0.1) in Eq. **??** were chosen empirically. During our development, we tested several fusion configurations (including simple multiplication) and found that this specific piecewise setup yielded the **highest Pearson correlation** with our human preference annotations, validating its superior alignment with human judgment.

3.2.2. Hybrid Feature Matching: For Fidelity and Identity Preservation

Dimensions: `Non-Edited Region Fidelity`, `Subject Identity Fidelity`, `Subject Consistency`, `Detail Fidelity/Preservation`.

Method: Hybrid LMM-Specialist (Sa2VA + DINOv3).

Rationale: We do not use a "Pure LMM" (baseline) for these dimensions due to "semantic drift." An LMM judges high-level *concepts* (*e.g.*, it might consider a *different person* in the same clothes as having high "subject identity"), not true *perceptual similarity*.

- Conversely, simple pixel-level metrics (*e.g.*, PSNR or LPIPS) fail to capture *feature-level* identity.
- Our hybrid pipeline (Eq. 2) leverages the LMM (via Sa2VA) for its strength: **semantic segmentation** (to identify *which* pixels correspond to the "subject" or "non-edited background").
- It then leverages the specialist model (DINOv3) for its strength: **feature-level comparison** (extracting and comparing feature vectors from those segmented regions). This provides a quantitative, robust score for fidelity that neither an LMM nor a simple metric could achieve.

3.3. Generalist LMM Pipelines: For Semantic and Cognitive Judgments

For dimensions where the core task is semantic understanding, logical reasoning, or subjective assessment, the LMM is the ideal, and often only, tool. We further refine this by selecting different LMM VQA structures based on the *nature* of the judgment required.

3.3.1. 5-Level VQA (for Subjective, Holistic Scoring)

Dimensions: `Blending Naturalness`, `Editing Artifacts`, `Instruction Following (Macro)`.

Method: LMM VQA-5Level (Eq. 1).

Rationale: These dimensions are inherently **subjective and holistic**. There is no "binary" correct answer for `Blending Naturalness` or `Editing Artifacts`; they exist on a spectrum. The 5-level weighted score (from "excellent" to "bad") is designed to capture this nuanced, perceptual "feel." Similarly, `Instruction Following (Macro)` assesses the overall *gist* and *intent* of the edit, making it a perfect choice for a 5-level holistic judgment.

3.3.2. 2-Level VQA (for Binary, Factual Success)

Dimensions: `Object Manipulation Accuracy`, `Local Attribute Accuracy`, `Action/State Change Accuracy`, `Spatial Accuracy`, `World Knowledge & Reasoning (SE)`.

Method: LMM VQA-2Level (Eq. 4).

Rationale: This targets factual, binary (Yes/No) task success, primarily in Single-Image (SE) edits. The edit either *happened as specified* (Yes) or *it did not* (No).

- "Was the horse changed to *white*?" (Yes/No).
- "Was the *correct* mayor's statue added?" (Yes/No).
- "Was the object placed to the *left* of the target?" (Yes/No).

A 2-Level (Yes/No) VQA is the most direct and unambiguous way to measure this **knowledge-retrieval** or **instruction-execution** semantic correctness. **Crucially, we are not testing the LMM's own knowledge; we use the LMM VQA to judge if the "edited image" matches our "pre-defined correct answer set." This ensures the objectivity of the evaluation.**

3.3.3. Multi-Question VQA (for Complex, Decomposed Reasoning)

Dimensions: `Physical Plausibility`, `Composition & Interaction`.

Method: LMM Multi-Question VQA (Eq. 3).

Rationale: These are cognitively complex dimensions. Asking an LMM for a single 1-5 score (i.e., the "Pure LMM" baseline) is **unreliable**. An LMM suffers from "attention bias" when evaluating complex scenes; it may focus only on the edited subject while completely ignoring its incorrect shadow or perspective.

- Our Multi-Question VQA method acts as a **"forced attention mechanism."** It *decomposes* the complex concept into a series of simple, verifiable sub-questions (*e.g.*, "1. Is the shadow direction of the new object correct?" "2. Is the perspective of the new object consistent?").
- This forces the LMM to perform a more robust, "chain-of-thought"-like analysis across multiple facets (lighting, shadows, perspective, occlusion). The resulting score, an aggregation of "Yes" answers (Eq. 3), is far more reliable and fine-grained than a single, potentially biased, holistic judgment.

3.3.4. LMM VQA (for Multi-Image Abstract Cognition & Cross-Image Relations)

These are the most difficult, SOTA-challenging dimensions in I2I-Bench, characterized by their need for **relational understanding** and **semantic reasoning across multiple source images**. No specialist model can perform such tasks, making the LMM the only tool for evaluation. We match the VQA structure to the nature of each task:

Dimension 1: `Cross-Source Attribute/Pose Transfer`

- **Method:** LMM VQA-2Level (Binary Factual Judgment).
- **Rationale:** This task (*e.g.*, "Transfer pose from A in Image 1 to B in Image 2") is a purely **"relational instruction."** The LMM must: (1) identify the source attribute (pose) in Image 1; (2) identify the target (B) in Image 2; and (3) judge if B in the generated image has *factually adopted* A's pose. No specialist model (*e.g.*, pose-estimator) can understand the semantic command "to transfer." This is a binary (Yes/No) cross-image semantic verification, making 2-Level VQA most appropriate.

Dimension 2: `Inter-Subject Consistency`

- **Method:** LMM VQA-5Level (Subjective Spectral Judgment).
- **Rationale:** This task assesses "how harmonious subjects from different sources look when composited together." This is distinct from Sec 2.2's `Subject Consistency` (which measures fidelity to the *source*). This is a high-level, **scene-wide artistic and semantic judgment**. The LMM must evaluate if lighting, scale, and artistic style (*e.g.*, a photorealistic person vs. an anime person) are consistent *within the new scene*. This is a subjective "feel", not a binary (Yes/No) question, making the 5-Level VQA spectrum the best fit.

Dimension 3: `World Knowledge & Reasoning (ME)`

- **Method:** LMM VQA-2Level (Binary Factual Judgment).
- **Rationale:** This task, unlike its SE "knowledge retrieval" counterpart, demands **"abstract logical reasoning"** (*e.g.*, map inference, Sudoku solving, logical combinations based on multiple images). The task occurs in *logical space*, not pixel space. The LMM is the only tool that can understand "logic". In line with Sec 3.2, to ensure objectivity, we **use the LMM VQA to judge if the "edited image" satisfies our "pre-defined correct answer set"** (*e.g.*, the correct Sudoku solution, the correct city name from the map). This makes the evaluation an objective (Yes/No) check, for which 2-Level VQA is ideal.

4. Prompt Quota per I2I-Bench Evaluation Dimension

This section details the exact number of prompts used to calculate the final scores for each model across all 30 dimensions in the proposed I2I-Bench evaluation suite. The total number of prompts is 1000, split equally between 500 Single-Image Editing (SE) prompts and 500 Multi-Image Editing (ME) prompts.

Table 1. Prompt Quota for Single-Image Editing (SE) Dimensions.

English Dimension Name	Prompt Quota
Aesthetic Quality	500
Blending Naturalness	500
Editing Artifacts	500
Image Quality	500
Instruction Following (Macro)	500
Non-Edited Element Fidelity	500
Physical Plausibility	500
Composition & Interaction	350
Object Manipulation Accuracy	332
Local Attribute Accuracy	282
Spatial Accuracy	275
Subject Identity Fidelity	233
Text Content & Style Accuracy	100
World Knowledge & Reasoning	100
Action/State Change Accuracy	86

Table 2. Prompt Quota for Multi-Image Editing (ME) Dimensions.

English Dimension Name	Prompt Quota
Aesthetic Quality	500
Blending Naturalness	500
Composition & Interaction	500
Detail Fidelity/Preservation	500
Image Quality	500
Instruction Following (Macro)	500
Non-Edited Element Fidelity	500
Physical Plausibility	500
Subject Consistency	500
Subject Extraction & Composition	500
Spatial Accuracy	476
Inter-Subject Consistency	412
Text Content & Style Accuracy	200
Cross-Source Attribute/Pose Transfer	126
World Knowledge & Reasoning	100

5. Comparative Analysis with LMM4Edit

To demonstrate the superiority of our proposed evaluation suite, we conducted a comparative analysis against LMM4Edit, a recent image editing evaluation metric based on LMMs. We performed inference using LMM4Edit on the Single-Image Editing component of I2I-Bench. While LMM4Edit provides pre-trained weights corresponding to multiple dimensions, we observed that not all checkpoints were fully applicable within our testing environment due to technical inconsistencies. Consequently, we selected one of the viable weight versions to conduct the comparative experiment.

Table 3 presents a detailed comparison of Pearson's Rho correlations between LMM4Edit and our method (Ours). The results unequivocally demonstrate that I2I-Bench significantly outperforms LMM4Edit across the vast majority of evaluation dimensions. Specifically:

- **Superior Overall Alignment**: Our method achieves a remarkably high average correlation of **0.9425** (excluding nan), compared to 0.5968 for LMM4Edit. This substantial gap validates the effectiveness of our "Specialist-Generalist Hybrid" evaluation strategy in aligning with human perception.
- **Robustness in Fundamental Dimensions**: LMM4Edit exhibits critical failures in fundamental quality assessment. Notably, it shows a negative correlation (-0.4568) in *Blending-Naturalness* and a weak correlation (0.2508) in *Image-Quality*, failing to correctly penalize artifacts. In contrast, our method achieves high consistency scores of 0.8866 and 0.9033, respectively, in these dimensions.
- **Reasoning Capabilities**: In complex tasks such as *World-Knowledge-Reasoning* and *Text-Content-Style-Accuracy*, our method demonstrates overwhelming superiority due to the integration of specialized tools (OCR and VQA specialists), whereas the pure LMM-based approach of LMM4Edit struggles significantly.

In conclusion, this experiment confirms that I2I-Bench provides a far more robust, accurate, and human-aligned evaluation framework compared to existing LMM-based metrics.

Table 3. Comparison of Pearson's Rho correlation with human preference between LMM4Edit and our method (Ours) on the I2I-Bench Single-Image Editing task. Our method demonstrates significant superiority across almost all dimensions.

Dimension	**LMM4Edit (ρ)**	**Ours (ρ)**	**Gap (Δ)**
Image-Quality	0.2508	**0.9033**	+0.6525
Aesthetic-Quality	0.7846	**0.9889**	+0.2043
Blending-Naturalness	-0.4568	**0.8866**	+1.3434
Non-Edited-Element-Fidelity	0.5984	**0.9019**	+0.3035
Subject-Identity-Fidelity	0.8210	**0.9133**	+0.0923
Physical-Plausibility	0.7683	**0.8055**	+0.0372
Editing-Artifacts	0.6551	**0.9006**	+0.2455
Instruction-Following-Macro	0.6266	**0.9960**	+0.3694
Object-Manipulation-Accuracy	0.7559	**0.9787**	+0.2228
Local-Attribute-Accuracy	0.7445	**0.9877**	+0.2432
Action-State-Change-Accuracy	0.5996	**0.9839**	+0.3843
Spatial-Accuracy	**0.9655**	0.9303	-0.0352
Text-Content-Style-Accuracy	0.5873	**0.9979**	+0.4106
World-Knowledge-Reasoning	0.2885	**0.9628**	+0.6743
Composition-Interaction	0.9631	**0.9997**	+0.0366
Average Correlation	0.5968	**0.9425**	**+0.3457**

6. Evaluation Pipeline and Prompt Details

This section details the automated hybrid evaluation methods used to assess the 30 fine-grained dimensions in I2I-Bench. The "Generalist" Large Multimodal Model (LMM) used for all evaluations is Qwen3-VL-8-Instruct.

6.1. Single-Reference (SE) Evaluation Dimensions

6.1.1. SE: Common Dimensions

These 7 dimensions assess the fundamental quality of all Single-Reference (SE) editing tasks.

1. Overall Image Quality & 2. Aesthetic Quality **Pipeline Type:** Specialist Models. **Tools:** Q-Insight, ArtiMuse. **Prompts:** N/A (Scores are obtained directly from the Specialist models).

3. Blending Naturalness **Pipeline Type:** LMM VQA-5Level. **Tools:** Qwen3-VL-8-Instruct.

System Prompt:

```
You are an expert evaluator of
image photorealism and coherence.
Your specific task is to assess
the realism of the edit itself.
Evaluate how seamlessly the
modified or added elements
integrate with the rest of the
image in terms of lighting,
shadows, perspective, and texture.
A high rating means the final image
looks natural and plausible, as
if it were a single, untouched
photograph.  Do not focus on
whether the instruction was
followed literally.  Your response
must be one of the following five
words directly:  excellent, good,
fair, poor, bad.  Do not add any
introductory phrases.
```

User Prompt (Q):

```
Please evaluate the realism and
coherence of the edit in the
'Generated Image' compared to
the 'Source Images'.  Assess how
seamlessly the edited elements
integrate with the rest of the
image in terms of lighting,
shadows, and overall plausibility.
Your response must begin with
one of the five rating words:
excellent, good, fair, poor, bad.
```

4. Generative/Editing Artifacts **Pipeline Type:** LMM VQA-5Level. **Tools:** Qwen3-VL-8-Instruct.

System Prompt:

```
You are an expert evaluator of
image editing quality.  You will
be given an 'Original Image', an
'Editing Instruction', and the
resulting 'Edited Image'.  Your
task is to assess the 'Edited
Image' for *unwanted artifacts*
introduced *during the editing
process*, not artifacts that were
in the original.
**Crucially**:  If the instruction
is stylistic (e.g., 'turn into a
painting', 'make it look like Van
Gogh'), *do not* penalize the image
for looking 'unnatural'.  Instead,
judge if the *application* of the
style is flawed (e.g., distorted,
incomplete, blotchy).
Also, heavily penalize edits that
*fail to preserve* unedited regions
(e.g., if the instruction is
'change the woman's hat', her face
and the background should remain
unchanged).
Based on the *severity* of these
*unwanted* artifacts, rate the
'Edited Image' using *only* one
of the following five words:
1.  excellent (Flawless edit.
The instruction is followed
perfectly with *zero* unwanted
artifacts.  Unedited areas are
perfectly preserved.)  2.  good
(Minor, hard-to-notice artifacts.
The edit is successful but may
have tiny imperfections upon
close inspection.)  3.  fair
(Noticeable artifacts.  The edit
is mostly successful, but there are
visible flaws like slight warping,
minor texture loss, or imperfect
blending.)  4.  poor (Significant,
obvious artifacts.  The edit is
flawed, with clear distortions,
unnatural warping, or significant
damage to unedited areas.)  5.  bad
(Extreme, severe artifacts.  The
edit is a total failure, resulting
in a grotesque, distorted, or
nonsensical image.)
```

User Prompt (Q):

```
Based on the instruction, how
severe are the *unwanted artifacts*
in the 'Edited Image'?
```

5. Instruction Following (Macro) **Pipeline Type:** LMM VQA-5Level. **Tools:** Qwen3-VL-8-Instruct.

System Prompt:

```
You are a meticulous evaluator
specializing in text-to-image
editing.  Your sole task is
to assess how accurately the
edited image reflects the given
instruction, based on the original
image.  Focus exclusively on
whether the edit described in
the instruction was performed
correctly.  You must ignore all
other factors, such as overall
image quality or any unintended
changes in areas not mentioned in
the instruction.  Your response
must be one of the following five
words directly:  excellent, good,
fair, poor, bad.  Do not add any
introductory phrases.
```

User Prompt (Q):

```
Please evaluate whether the
'Generated Image' successfully
implements the following
instruction.  Instruction:
"prompt".  Your response must begin
with one of the five rating words:
excellent, good, fair, poor, bad.
```

6. Non-Edited Region Fidelity **Pipeline Type:** Hybrid LMM-Specialist. **Tools:** LMM (Qwen3-VL-8-Instruct) + Sa2VA + DINOv3. **Description:** The LMM generates a segmentation command for the **"edited"** region. The mask is then **inverted (NOT)** to isolate the *non-edited region*.

LMM Segmentation Command Prompt (Q):

```
You are an image segmentation
assistant.  Compare <image_1>
(Original Image) and <image_2>
(Edited Image), and considering
the editing instruction:
"edit_instruction", generate a text
```

```
command to segment the edited or
modified region in the image. Your
response MUST begin with "Please
segment". This command will be
used to segment the edited image
in isolation. Therefore, ensure
the command is clear, focuses only
on the edited image, and makes no
reference to the original image, as
this would confuse the segmentation
model.
```

7. Physical Plausibility Pipeline Type: LMM Multi-Question VQA (Multi-VQA). **Tools:** Qwen3-VL-8-Instruct.

QG (Question Generation) Prompt:

```
You are an expert in physical
plausibility evaluation. Your
task is to design a set of (5
to 7) precise, specific binary
(Yes/No) questions based on the
given original image and an editing
instruction, to evaluate whether
the edited image is physically
plausible.
When designing questions, strictly
follow this thought process and
principles:
1. **Analyze Scene &
Instruction**: * **Identify
Subject & Action**: What object
(subject) is being manipulated
(added, removed, replaced,
attribute changed)? * **Analyze
Physical Context**: Observe
the original image to identify
relevant physical properties and
environmental factors: e.g.,
lighting direction, shadows,
reflective surfaces, support
relationships, rigid vs. soft
bodies.
2. **Generate Questions**:
Based on the analysis, generate
questions from the following
physical dimensions. Each must be
a closed-ended Yes/No question. *
**Optics (Shadows, Reflections)**:
If an object was added/moved, is
its new shadow consistent with
the scene's light source? If an
object was removed, is its shadow
also gone? Are reflections on
nearby surfaces correctly updated?
* **Mechanics (Deformation,
Support)**: If a heavy object is
placed on a soft surface, does the
surface show plausible deformation?
If a supporting object is removed,
does the object it supported (e.g.,
a vase on a table) defy gravity?
* **State Transition (Weather,
State)**: If the instruction
changes the weather (e.g., 'make it
winter'), are all elements (trees,
ground) consistently updated (e.g.,
covered in snow)?
3. **Output Format**: Please
output a JSON-formatted string
containing a list of questions.
The JSON object should have a key
named "physical_questions" with a
value that is a list of strings.
Example: { "physical_questions":
[ "Question 1: ...?", "Question 2:
...?", "Question 3: ...?" ] }
Now, please generate physical
plausibility evaluation questions
based on the following image
and instruction. Instruction:
"prompt".
```

QA (Question Answering) Prompt:

```
You are an expert in physical
plausibility evaluation. Please
carefully observe the provided
image and judge based on the
following question.
Question: "question"
Please answer with only one word
based on visual evidence: Yes or
No.
```

6.1.2. SE: Specific Dimensions

These 8 dimensions assess the execution quality of specific instructions.

8–11. Object Manipulation, Local Attribute, Action/State Change, Spatial Accuracy Pipeline Type: LMM VQA-2Level. **Tools:** Qwen3-VL-8-Instruct.

System Prompt:

```
You are an expert in evaluating
image editing. Your task is to
determine if the edited image
successfully implements the
given instruction compared to
the original image. Focus solely
on whether the instruction was
achieved. Answer strictly and only
with 'Yes' or 'No'.
```

User Prompt (Q):

```
Instruction: "prompt". Does
Image 2 successfully implement the
instruction compared to Image 1?
Answer Yes or No.
```

12. Text Content & Style Accuracy Pipeline Type: Multi-step Hybrid Pipeline. **Tools:** GOT-OCR2.0 + LMM (Qwen3-VL-8-Instruct). **Description:** GOT-OCR2.0 evaluates content accuracy (S_c). The LMM then uses a VQA-5Level pipeline to evaluate style and position (S_s).

LMM Style/Position Prompt (Q):

```
You are an expert evaluator for
AI-generated images, specializing
in text rendering. Your task
is to evaluate how well an image
follows the text-related **style
and position** instructions from a
user prompt.
**CRITICAL RULE:** Do NOT evaluate
the text's spelling or accuracy.
Assume the spelling is correct,
even if it is not. Your score
must ONLY reflect the non-accuracy
requirements (like font style,
color, placement, etc.).
You will be given: 1. **[User
Prompt]**: The original prompt
used to generate the image. 2.
**[Target Text]**: The specific
text string that was requested. 3.
**[Image]**: The generated image.
**Step 1: Analyze Requirements**
First, analyze the [User Prompt] to
identify the specific requirements
for: * **Text Style**: What
instructions were given for the
text's appearance (e.g., "neon",
"handwritten", "bold", "red
color", "glowing", "artistic
font")? * **Text Position**: What
instructions were given for the
text's location (e.g., "on the
sign", "in the top-left corner",
"on the t-shirt")?
**Step 2: Evaluate Image against
Requirements** Compare the
text in the [Image] against the
non-accuracy requirements you
identified. Remember to IGNORE
spelling errors.
**Step 3: Assign a Single
Score (1-5)** Provide a single,
holistic score for **Style and
Position Compliance** based on
this rubric. * [5] Excellent
Match: All specified style
and position instructions were
followed perfectly. * [4] Good
Match: All specified instructions
were followed, but with minor
deviations. * [3] Partial Match:
The core idea of *at least one*
instruction was attempted but
executed poorly, OR one major
instruction was followed while
another was missed. * [2] Poor
Match: At least one specified
instruction was clearly ignored
or failed. * [1] No Match: All
specified style and position
instructions were completely
ignored.
**Step 4: Provide Output in JSON
Format** Provide your evaluation
in a strict JSON format. Do not
include any text outside the JSON
block.
**JSON Output Format:** {
"analysis": { "style_requirement":
"...", "position_requirement":
"...", "image_observation": "...",
"reasoning": "...", "score":
[1-5] } }
```

13. World Knowledge & Reasoning Pipeline Type: LMM VQA-2Level. **Tools:** Qwen3-VL-8-Instruct.

System Prompt:

```
You are an expert evaluator for
image editing based on multiple
reference images. Your task
is to determine if the 'Edited
Image' successfully implements the
requested change described in the
'Instruction', based *specifically*
on the 'Evaluation Criteria
(Hint)'. Compare the 'Reference
Images' and 'Edited Image'. Your
response must be *only* the single
word 'Yes' or 'No'.
'Yes' = The edit was successfully
implemented according to the
hint. 'No' = The edit was not
successfully implemented according
to the hint.
```

User Prompt (Q)**:**

```
**Instruction (Prompt):** prompt
**Evaluation Criteria (Hint):**
hint
Based on ALL reference images,
the instruction, and the specific
criteria in the hint, has the edit
been successfully implemented in
the 'Edited Image'? Answer with
only 'Yes' or 'No'.
```

14. Subject Identity Fidelity **Pipeline Type:** Hybrid LMM-Specialist. **Tools:** LMM (Qwen3-VL-8-Instruct) + Sa2VA + DINOv3. **Description:** The LMM generates a command to segment regions that **"should remain unchanged"** (*e.g.*, the face, if the instruction is 'change the shirt'). The mask is **not inverted**.

LMM Segmentation Command Prompt (Q)**:**

```
You are an expert in image editing
analysis. Given an original image,
an edited image, and the editing
instruction, identify all distinct
main subjects or regions present
in the original image that *should
remain unchanged* according to the
instruction.
Your output MUST BE ONLY a JSON
list of strings. Each string must
be a separate segmentation command
for one distinct subject/region, in
the format 'Please segment [subject
name in English]'.
For example: ["Please segment
background sky", "Please segment
mountains", "Please segment main
building"]
```

15. Composition & Interaction **Pipeline Type:** LMM Multi-Question VQA. **Tools:** Qwen3-VL-8-Instruct.

QG (Question Generation) Prompt:

```
You are an expert in visual arts
and composition evaluation.
Your task is to design a set
of (5 to 7) precise, specific
binary (Yes/No) questions based
on the given original image
and an editing instruction, to
evaluate whether the edited
image is plausible in terms of
**composition, perspective, scale,
and interaction**.
When designing questions, strictly
follow this thought process and
principles:
1. **Analyze Scene &
Instruction**: * **Identify
Subject & Action**: What
object (subject) is being
manipulated (added, removed,
replaced, attribute/pose
changed)? * **Analyze Visual
Context**: Observe the original
image to identify key visual
elements: scene perspective
(close-up, long-shot, eye-level),
key objects, spatial layout
(foreground/background), and
interaction area implied by the
instruction.
2. **Generate Questions**:
Based on the analysis, generate
questions from the following
dimensions. * **Composition &
Placement**: Is the new/moved
object in a logical position? Is
its occlusion (in front of/behind
other objects) correct? *
**Perspective & Scale**: Is the
scale of the new/modified object
consistent with other objects in
the scene? Does its perspective
```

```
match the scene's perspective? *
**Interaction & Naturalness**:
If a pose was changed, is it
anatomically natural? If objects
are interacting (e.g., hand holding
a balloon), is the contact point
believable?
3. **Output Format**: Please
output a JSON-formatted string
containing a list of questions.
The JSON object should have a key
named "composition_questions" with
a value that is a list of strings.
Now, please generate composition
and interaction plausibility
evaluation questions based on the
following image and instruction.
Instruction: "prompt".
```

QA (Question Answering) Prompt: (Reused from SE Physical Plausibility).

6.2. Multi-Reference (ME) Evaluation Dimensions

These 15 dimensions evaluate complex multi-image editing tasks.

6.2.1. ME: Common Dimensions

These 9 dimensions are shared with the SE category.

1–5. Aesthetic Quality, Blending Naturalness, Editing Artifacts, Image Quality, Instr. Following (Macro) Prompts: Reused from the corresponding SE definitions.

6. Spatial Accuracy Prompts: Reused from the SE LMM VQA-2Level definition.

7. Composition & Interaction Pipeline Type: LMM Multi-Question VQA.

QG Prompt (ME-Specific):

```
You are an expert in visual arts
and composition evaluation. Your
task is to design a set of (5 to 7)
precise, specific binary (Yes/No)
questions based on *multiple*
given source images (labeled
Figure 1, Figure 2, ...) and an
editing instruction, to evaluate
whether the *final composited
image* is plausible in terms of
**composition, perspective, scale,
and interaction**.
When designing questions, strictly
follow this thought process and
principles:
1. **Analyze Scene &
Instruction**: * **Identify
Sources & Composition**: What
elements are extracted from
which images? How are they
combined, modified, and placed?
* **Analyze Visual Context**:
Check perspective, scale, and
interaction. Does the instruction
require interaction (e.g., A sits
on B)?
2. **Generate Questions**: *
**Composition & Placement**: Are
elements placed logically (e.g.,
not floating)? Is occlusion
correct? * **Perspective &
Scale**: Is the scale of an
element from Fig 1 consistent
with the scene from Fig 2? Do
all elements share a consistent
perspective? * **Interaction
& Naturalness**: If a pose was
changed to interact (e.g., sit on a
chair), is the final pose natural?
Is the physical contact believable?
3. **Output Format**: Please
output a JSON-formatted string...
The JSON object should have a key
named "composition_questions"...
Now, please generate... questions
based on the following *multiple*
images and the instruction.
Instruction: "prompt".
```

QA Prompt: (Reused from SE Physical Plausibility).

8. Physical Plausibility Pipeline Type: LMM Multi-Question VQA.

QG Prompt (ME-Specific):

```
You are an expert in physical
plausibility evaluation. Your
task is to design a set of (5 to 7)
precise, specific binary (Yes/No)
questions based on *multiple* given
source images (labeled Figure 1,
Figure 2, ...) and an editing
instruction, to evaluate whether
the *final composited image* is
```

```
physically plausible.
When designing questions, strictly
follow this thought process and
principles:
1.  **Analyze Scene &
Instruction**:  * **Identify
Sources & Composition**:  What
elements are extracted and how
are they combined?  * **Analyze
Physical Context**:  Are the
lighting, shadows, and physics
consistent *between* elements from
different sources?  * **Analyze
Interaction**:  Do elements
interact?  Do these interactions
obey physical laws (support,
occlusion, deformation)?
2.  **Generate Questions**:  *
**Optics (Lighting/Shadows)**:
Are the shadows and lighting
on all combined elements
consistent with a single, unified
light source?  * **Mechanics
(Support/Deformation)**:  If an
element from Fig 1 is placed on
a soft element from Fig 2, does
the surface plausibly deform?  Are
support structures logical?  *
**State Consistency**:  If the
instruction changes the global
state (e.g., 'make it rain'), does
this state apply consistently to
all elements from all sources?
3.  **Output Format**:  Please
output a JSON-formatted string...
The JSON object should have a key
named "physical_questions"...
Now, please generate...  questions
based on the following *multiple*
images and the instruction.
Instruction:  "prompt".
```

QA Prompt: (Reused from SE Physical Plausibility).

9. Non-Edited Element Fidelity **Pipeline Type:** Hybrid LMM-Specialist.

LMM Segmentation Command Prompt (Q):

```
You are an expert in analyzing
complex image editing instructions.
Given multiple source images, an
edited image, and the instruction,
your task is to identify which
specific *source subjects* (e.g.,
'Luffy from Figure 2', 'Conan
from Figure 3') are explicitly
instructed to be preserved
*without changes* (e.g., 'keep
their original poses', 'remain
unchanged').
For each such non-edited
subject you find, output a
JSON object containing:  1.
'source_index':  The 0-based
index of the source image where
this subject originates.  2.
'segmentation_prompt':  A short
segmentation command in the format
'Please segment [subject name in
English]'.
Your output MUST BE ONLY a
JSON list of these objects.
Example:  ["source_index":  1,
"segmentation_prompt":  "Please
segment Luffy", "source_index":  2,
"segmentation_prompt":  "Please
segment Conan"] If the instruction
modifies *all* subjects in some way
(e.g., 'put all in new clothes'),
output an empty list `[]`.  Do not
add any text before or after the
JSON list.
```

6.2.2. ME: Specific Dimensions

These 6 dimensions are specific to multi-reference tasks.

1. Cross-Source Attribute/Pose Transfer **Pipeline Type:** LMM Multi-Question VQA.

QG Prompt:

```
You are an expert in "Attribute and
Pose Transfer" evaluation.  Your
task is to design a set of (5 to 7)
precise, specific binary (Yes/No)
questions based on *multiple* given
source images (labeled Figure 1,
Figure 2, ...)  and an editing
instruction, to evaluate whether
the *final composited image*
has **accurately and with high
quality** completed the attribute
or pose transfer.
When designing questions, strictly
follow this thought process and
```

```
principles:
1. **Analyze Instruction**: *
**Identify Source & Target**:
What attribute/pose is extracted
from which subject (e.g., clothes
from A in Fig 1)? * **Identify
Recipient**: What subject is the
attribute/pose applied to (e.g., B
in Fig 2)?
2. **Generate Questions**: *
**Attribute Transfer**: Is the
transferred attribute (e.g.,
clothing, color) accurately and
completely replicated on the target
subject? Is the target subject’s
identity (e.g., face, body shape)
preserved? Does the new attribute
fit the target’s pose naturally? *
**Pose Transfer**: Is the new pose
an exact match to the source pose?
Is the target subject’s identity
preserved while performing the new
pose? Is the new pose anatomically
plausible for the target subject?
3. **Output Format**: Please
output a JSON-formatted string...
The JSON object should have a key
named “transfer_questions”...
Now, please generate... questions
based on the following *multiple*
images and the instruction.
Instruction: “prompt”.
```

QA Prompt: (Reused from SE Physical Plausibility).

2. Inter-Subject Consistency Pipeline Type: LMM Multi-Question VQA.

QG Prompt:

```
You are an expert in visual
consistency evaluation. Your task
is to design a set of (5 to 7)
precise, specific binary (Yes/No)
questions based on *multiple* given
source images (labeled Figure 1,
Figure 2, ...) and an editing
instruction, to evaluate whether
the elements from different source
images appear visually consistent
in the *final composited image*.
When designing questions, strictly
follow this thought process and
principles:
1. **Analyze Instruction &
Sources**: * **Identify Extracted
Elements**: What is taken from
Fig 1? From Fig 2? * **Analyze
Source Context**: What is the
lighting in Fig 1? The style in
Fig 2? The perspective in Fig 3?
* **Analyze Final Scene**: Where
are they being combined?
2. **Generate Questions**: *
**Lighting Consistency**: Do
the highlights and shadows on
the element from Fig 1 and the
element from Fig 2 look like they
are caused by the *same* light
source in the final image? *
**Scale & Perspective**: Is
the relative scale between the
element from Fig 1 and the element
from Fig 2 realistic? Do their
perspectives match the final
scene’s horizon line? * **Style
Consistency**: Do all elements
share a unified artistic style
(e.g., photographic vs. cartoon)?
Is the image quality (sharpness,
noise) consistent across elements?
3. **Output Format**: Please
output a JSON-formatted string...
The JSON object should have a key
named “consistency_questions”...
Now, please generate... questions
based on the following *multiple*
images and the instruction.
Instruction: “prompt”.
```

QA Prompt: (Reused from SE Physical Plausibility).

3. Subject Consistency and Detail Fidelity Pipeline Type: Hybrid LMM-Specialist.

LMM Segmentation Command Prompt (Q):

```
You are an expert in analyzing
image editing fidelity for
multi-reference composition. Given
multiple source images, an edited
composite image, and the editing
instruction, identify the key
visual details (like specific
accessories, textures, facial
features, fur patterns) of the
```

```
subjects *extracted from the
source images* that are critical
for preserving the subjects'
identities and should ideally
remain unchanged in the edited
image. For each identified detail,
output a JSON object containing
'source_index' (the 0-based index
of the source image where the
detail originates from the provided
list) and 'segmentation_prompt'
(a short, specific segmentation
command in the format 'Please
segment [detail name in English]').
Output ONLY a JSON list of these
objects, like `["source_index": 0,
"segmentation_prompt": "Please
segment detail1", "source_index":
1, "segmentation_prompt": "Please
segment detail2"]`. Do not add any
text before or after the JSON list.
```

4. Subject Extraction & Composition Pipeline Type: Multi-step (LMM 2-Level + Hybrid). **Description:** A 2-step process. First, $Score_{count}$ is computed. Second, this is multiplied by the $Score_{consistency}$ (from the dimension above).

LMM 2-Level (Count) Prompt (Q):

```
You are an image element counter.
Carefully observe the "Source Image
1", "Source Image 2", ... and the
"Generated Image". Also read the
"Instruction" below.
Instruction: "instruction"
Your task is to: **Judge only if
the "Generated Image" contains
all the subjects or objects
required for composition by the
"Instruction".** ... **focus only
on the quantity**...
Question: Based on the
instruction, does the
"Generated Image" contain the
**correct number** of required
subjects/objects?
Please answer with only one word...
Yes or No.
```

5. Text Content & Style Accuracy Prompts: Reused from the SE Text Content & Style Accuracy definition.

6. World Knowledge & Reasoning Prompts: Reused from the SE World Knowledge & Reasoning definition.

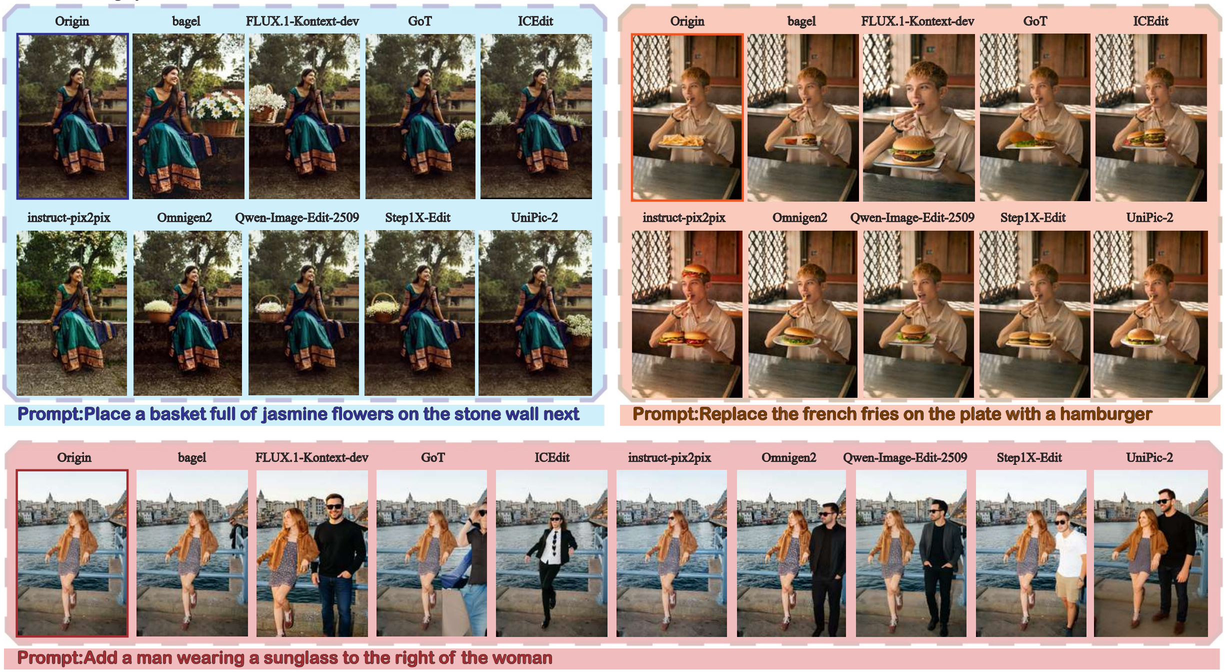
7. More Editing Example

This appendix provides visual examples for the 10 editing task categories defined in I2I-Bench (5 Single-Image Editing categories and 5 Multi-Image Editing categories).

To clearly demonstrate the specific tasks and challenges of each category, each of the following pages (Figure 1 through Figure 10) is dedicated to one category. The "full-page figure" on each page is a composite image that includes the Source Image(s) used for the example, the Prompt (instruction), and one or more representative Output Image(s).

Object Manipulation

For PEOPLE: In several image editing tasks involving people, we can see that instruct-pix2pix is basically unable to complete the task when processing people, while the other models perform more averagely.



For SCENE: In several image editing tasks related to the scene, we can see that, except for Qwen-Image-Edit-2509, the other models do not perform very well when processing the scene, and always leave editing traces when manipulating objects in the image.

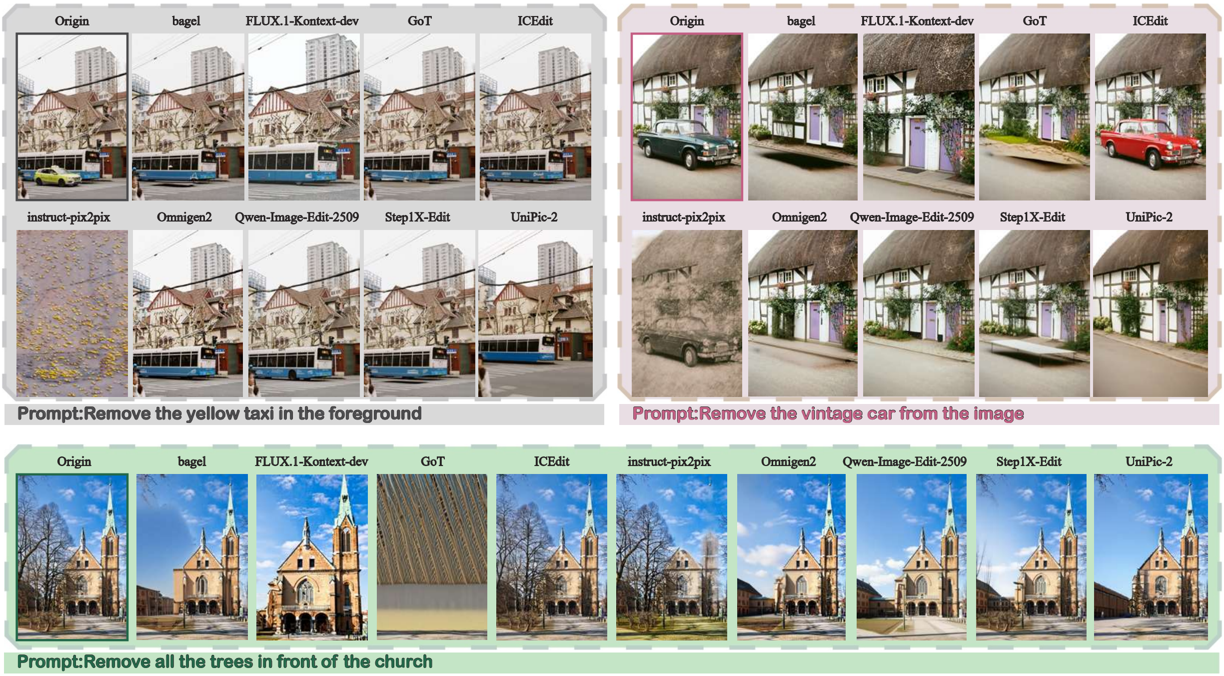
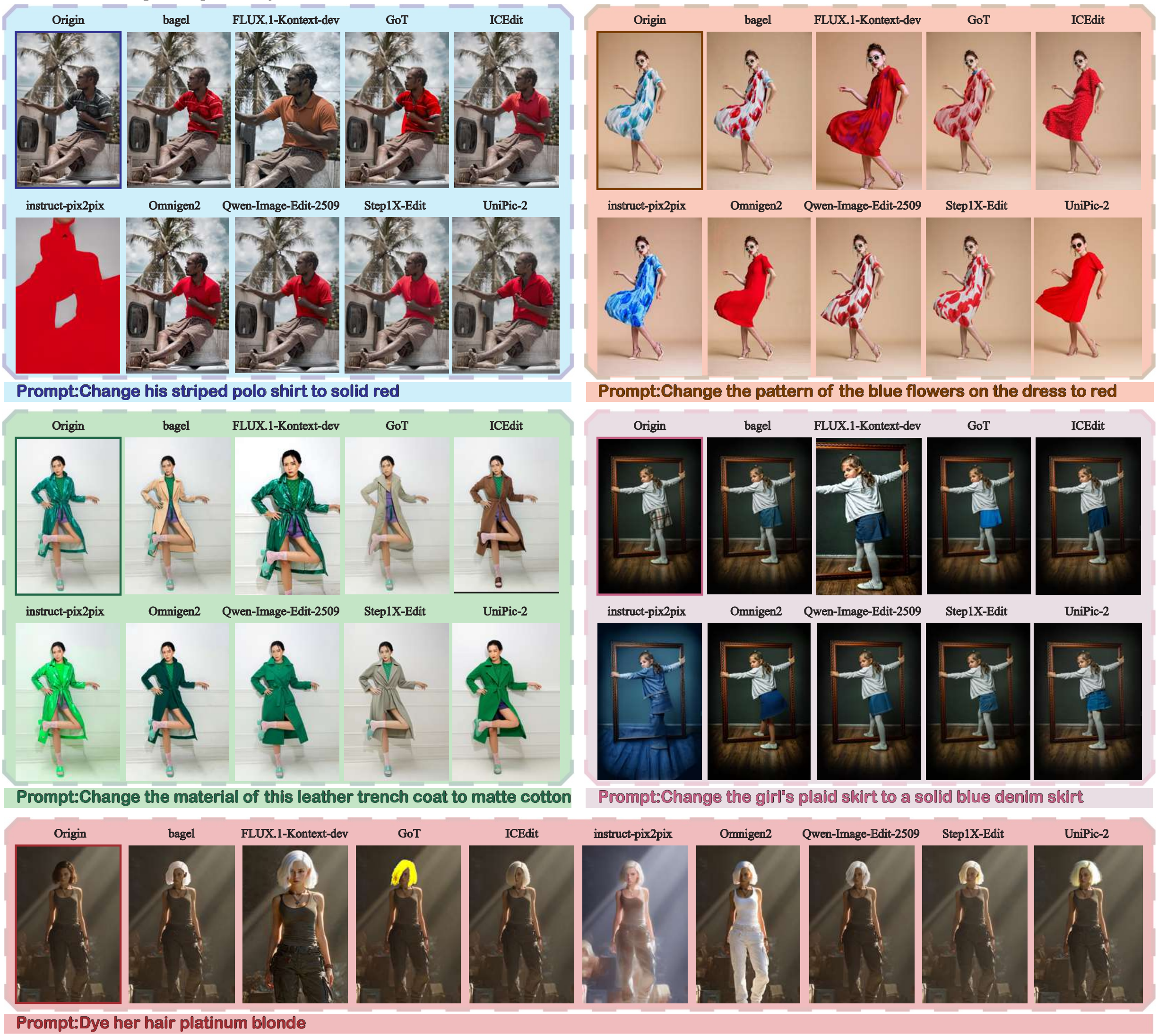


Figure 1. Visual examples for the "Object Manipulation" category. This figure shows multiple test cases from this category, including their corresponding source images, prompts, and output results.

Local Attribute Editing

For images where the subject occupies a large portion of the image, most models can effectively perform editing tasks for simple tasks. However, for tasks where the prompts are not direct, the models do not perform particularly well.

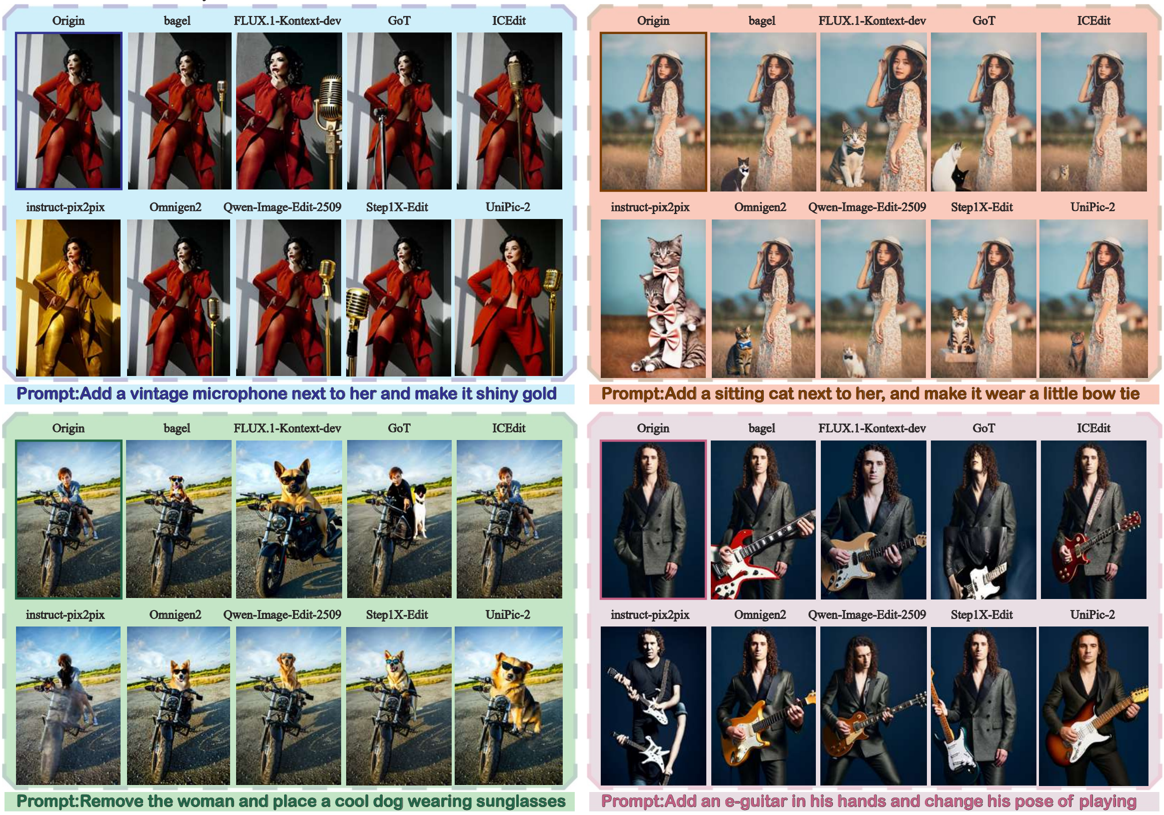


For SCENE: In several image editing tasks related to the scene, we can see that, except for Qwen-Image-Edit-2509, the other models do not perform very well when processing the scene, and always leave editing traces when manipulating objects in the image.

Figure 2. Visual examples for the "Local Attribute Editing" category. This figure shows multiple test cases from this category, including their corresponding source images, prompts, and output results.

Hybrid Editing

As we can see, in Hybrid Editing, most models perform worse than in other editing tasks, but Qwen-Image-Edit-2509 remains very stable and accurately meets the task requirements, while other models introduce many flaws.

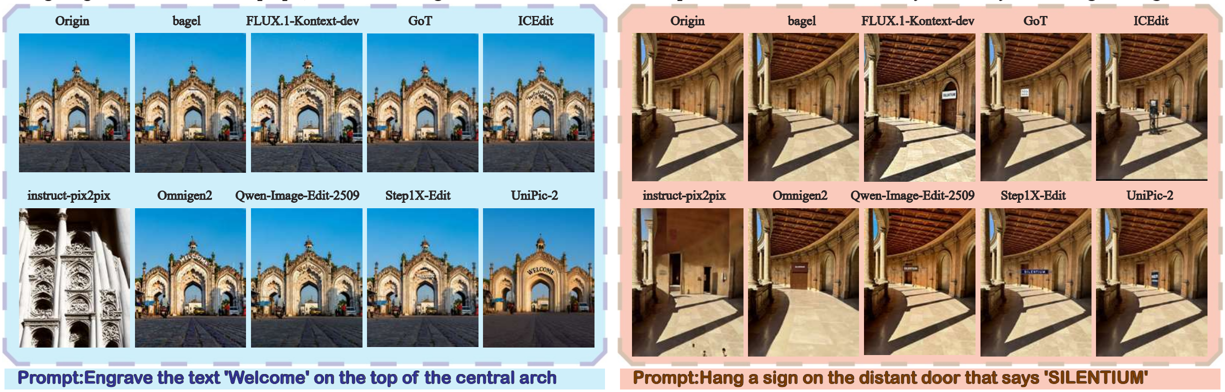


For editing scenes rather than portraits, most models perform better than the former type of task.

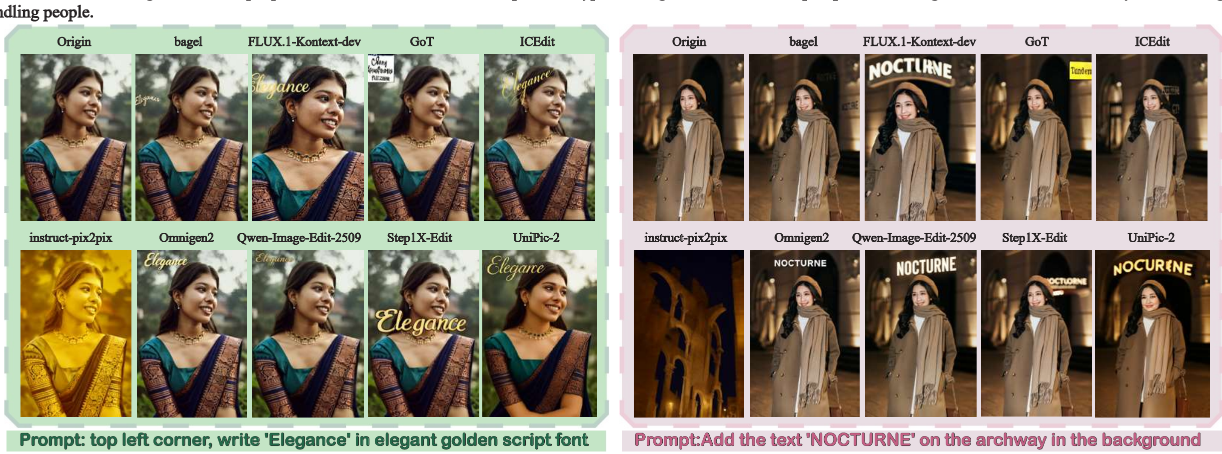
Figure 3. Visual examples for the "Hybrid Editing" category. This figure shows multiple test cases from this category, including their corresponding source images, prompts, and output results.

Text Rendering

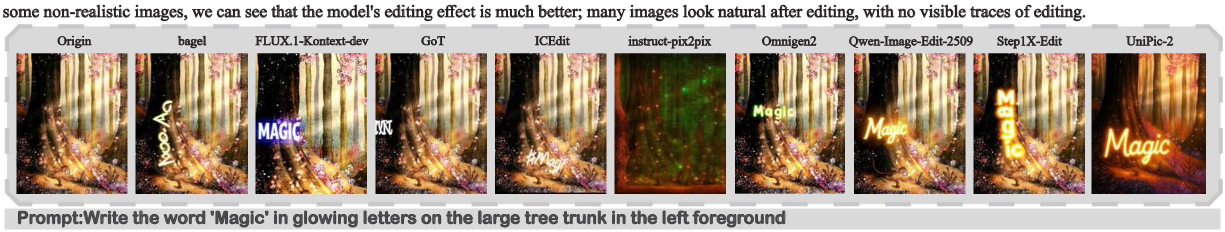
For editing images of scenes rather than people, most of the editing done on the models is rather abrupt and fails to maintain consistency with the style of the original image.



For some edits involving scenes with people, the results are better than the previous type of images, but the instruct-pix2pix and Omnigen2 models still have many shortcomings in handling people.



For some non-realistic images, we can see that the model's editing effect is much better; many images look natural after editing, with no visible traces of editing.



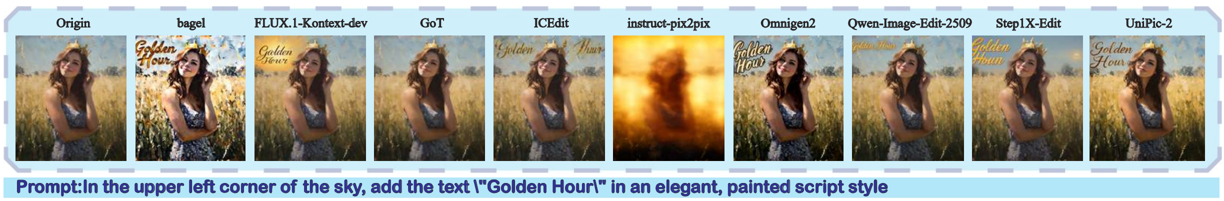


Figure 4. Visual examples for the "Text Rendering" category. This figure shows multiple test cases from this category, including their corresponding source images, prompts, and output results.

World Knowledge & Reasoning

In this chapter, our prompt is not as straightforward as the previous tasks; instead, it incorporates a lot of world knowledge and reasoning. As you can see, the image quality becomes inconsistent, with a very large variance, revealing numerous problems.

Figure 5. Visual examples for the "World Knowledge & Reasoning" category. This figure shows multiple test cases from this category, including their corresponding source images, prompts, and output results.

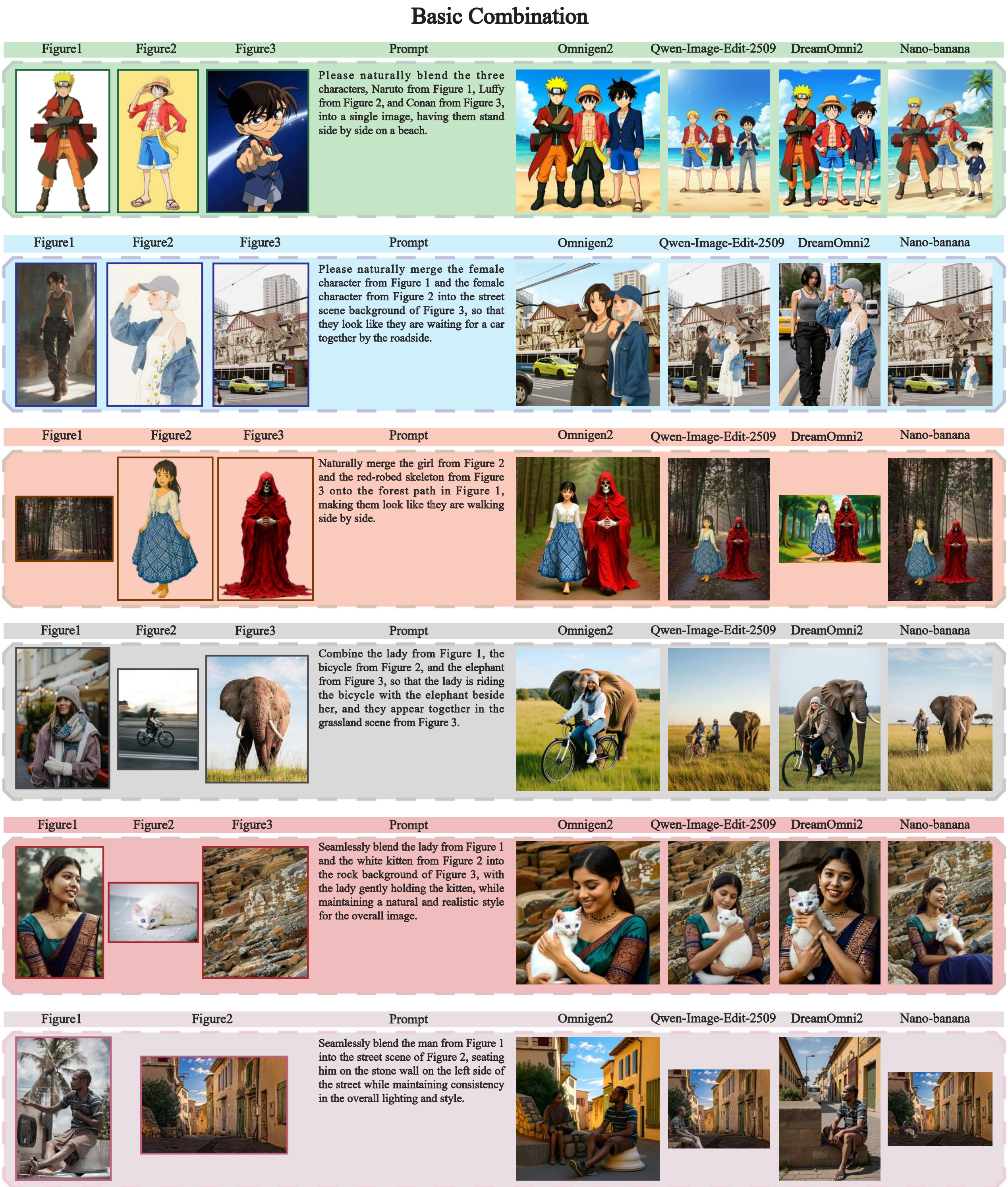


Figure 6. Visual examples for the "Basic Combination" category. This figure shows multiple test cases from this category, including their corresponding source images, prompts, and output results.

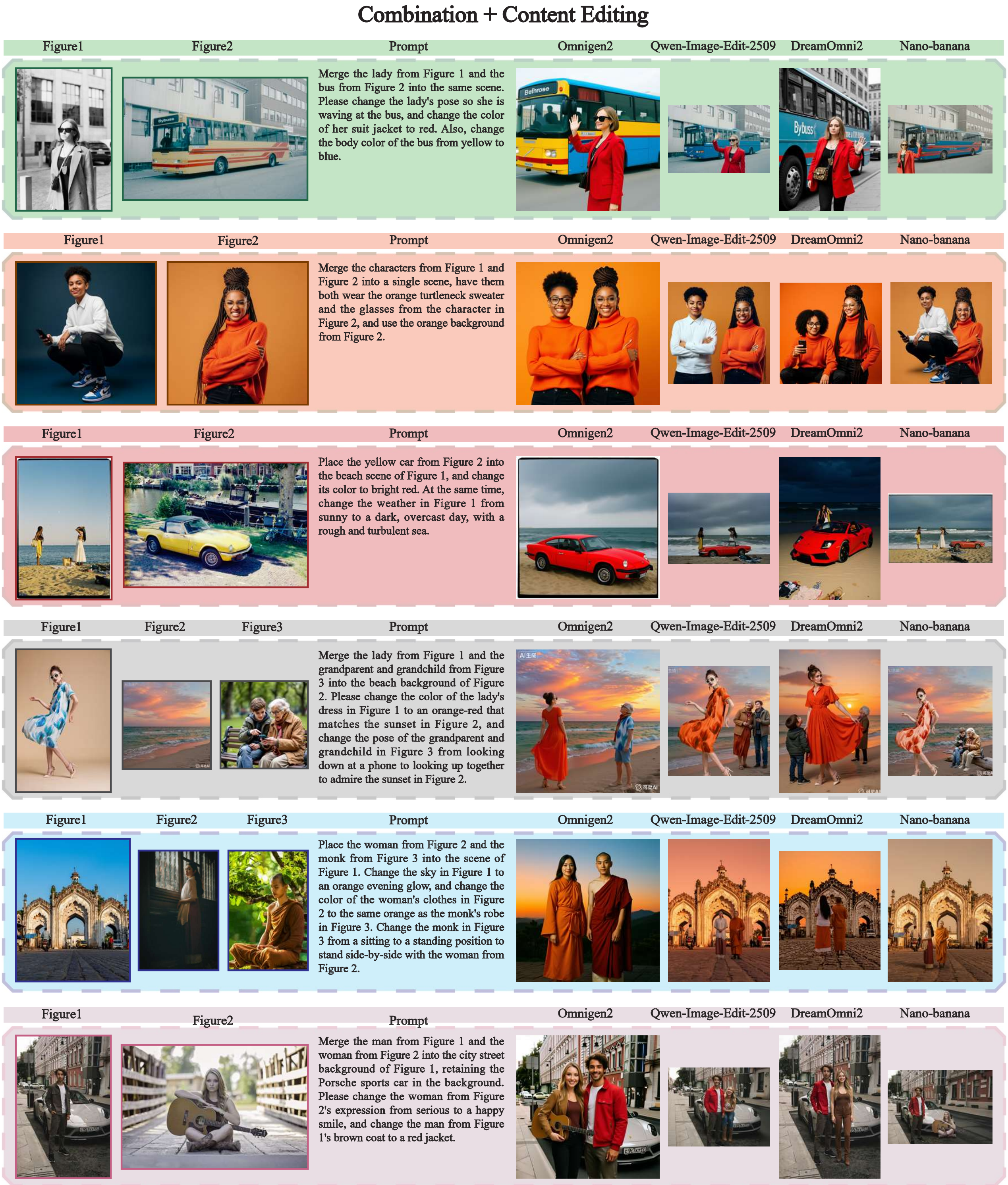


Figure 7. Visual examples for the "Combination + Content Editing" category. This figure shows multiple test cases from this category, including their corresponding source images, prompts, and output results.

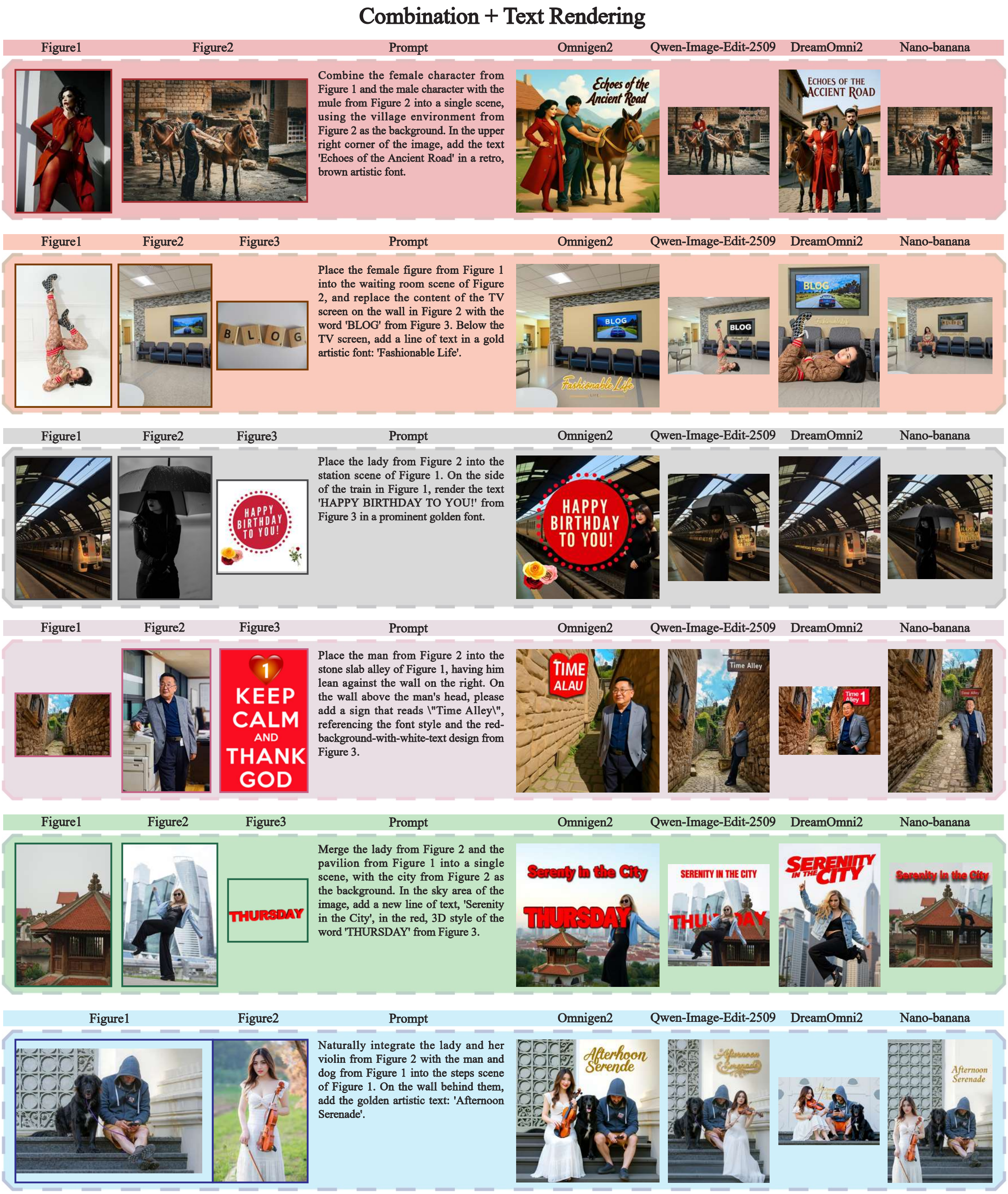


Figure 8. Visual examples for the "Combination + Text Rendering" category. This figure shows multiple test cases from this category, including their corresponding source images, prompts, and output results.

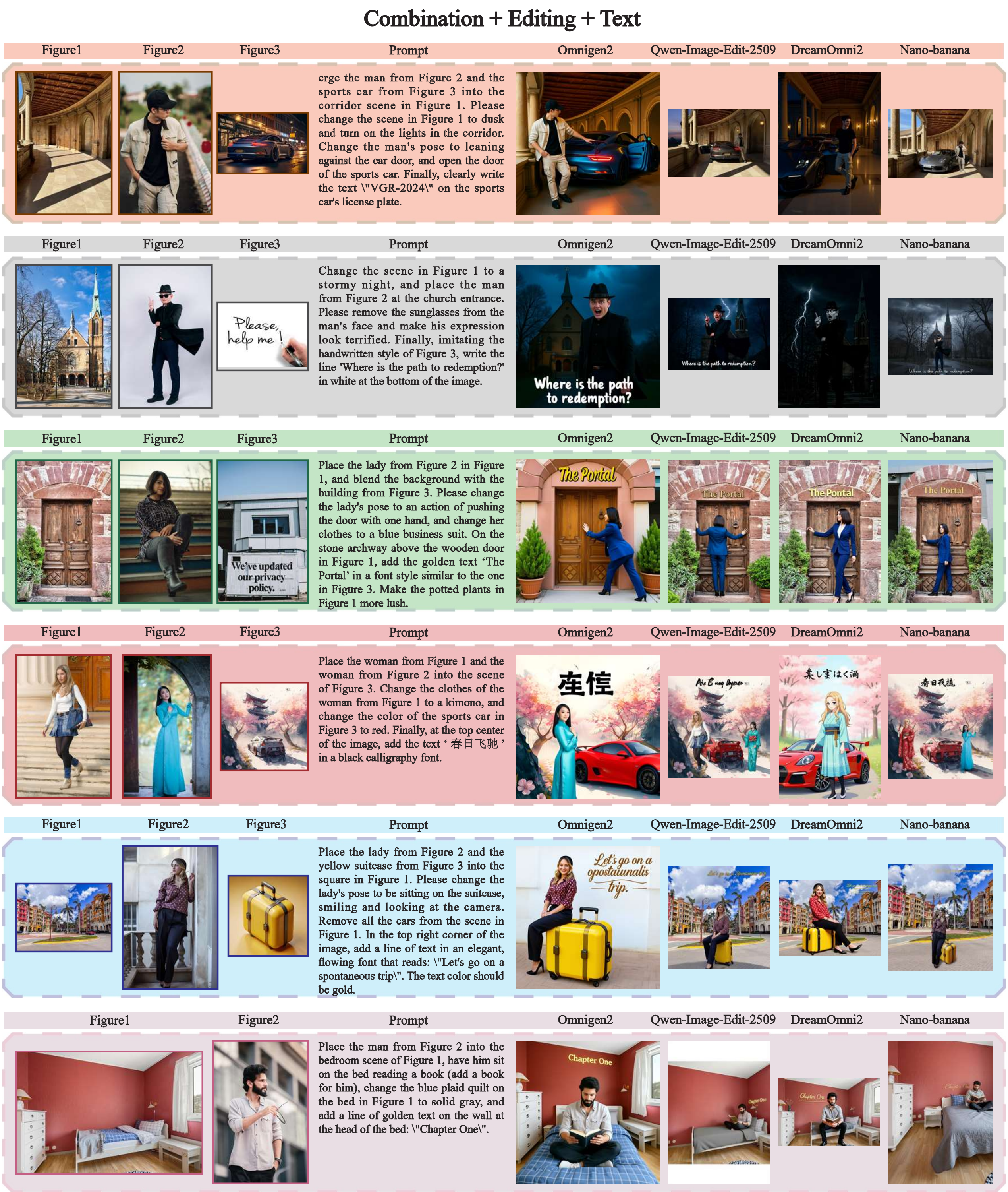


Figure 9. Visual examples for the "Combination + Editing + Text" category. This figure shows multiple test cases from this category, including their corresponding source images, prompts, and output results.

Figure 10. Visual examples for the "Combination + Reasoning" category. This figure shows multiple test cases from this category, including their corresponding source images, prompts, and output results.