# WEAVE: Unleashing and Benchmarking the In-context Interleaved Comprehension and Generation

Supplementary Material

## Contents



## A. Weave Analysis

### A.1. Collection Process

To ensure the quality of the generated data, we incorporated manual sampling verification into the design process of each pipeline to validate the success rate after filtering. Specifically, we utilized four pipelines, each with integrated quality assurance mechanisms.

*(i) Multi-image fusion*: We achieved reference to previous iterations by fusing edited or directly generated images. For image fusion data, we utilized two primary sources. First, we leveraged the multi-image fusion dataset from Echo-4o [56], where image fusion was initially performed using GPT-Image. Due to quality inconsistencies in this dataset, we regenerated images using Seedream 4.0 [41] and refined instructions with GPT-4.1. Second, we generated single-round image fusion instructions with GPT-4.1, including original image captions. We then produced original images using Qwen-Image [50], substituting suboptimal generations with Seedream 4.0 outputs, and performed multi-image fusion using Seedream 4.0. Building upon these single-round fusion data, we employed GPT-4.1 to annotate image editing instructions for the original images, categorizing them into five types: 'add', 'remove', 'replace', 'color alter', and 'background change' following the taxonomy in [3]. We subsequently applied Step1X-Edit(v1.2) [32] for single-round editing. For images failing our quality verification protocol, we utilized Nano Banana [14] for additional refinement. Finally, GPT-4.1 provided reverse instructions and captions for edited images. We used these edited images as originals and multi-fusion input images as edited results, concatenating the data to create comprehensive multi-round editing and multi-image fusion sequences.

*(ii) Remove-then-back*: We employed GPT-4.1 [1] to generate instructions for multi-round editing. Specifically, we designed the instructions such that one round would require adding back an object that had been previously removed or replaced in an earlier round. Following instruction generation, we implemented a filtering process wherein approximately 25% of instructions successfully met our criteria. The filtered instructions were subsequently utilized to generate outputs using Seedream 4.0 [41] and Nano Banana [14], after which we retained the superior generation based on qualitative assessment.

*(iii) Derivative imagination and comparison*: We incorporated methods for deriving or imagining alternative solutions or new images before fusion. Due to the inherent challenges

in automating LLMs to generate associative content or editing data, we adapted chess game and visual jigsaw datasets from Zebra-CoT [30] using GPT-4.1 for both recombination and self-verification processes. Specifically, we modified the abbreviated chess notations into explicit editing instructions to mitigate potential comprehension difficulties in generative models when interpreting condensed commands.

*(iv) Sequential procedures*: We implemented sequential edits following narrative progressions or structured operations requiring visual memory during generation. This approach was particularly effective for scenarios where characters disappear and subsequently reappear within narratives. Multiple editing rounds on identical scenes evaluated model consistency maintenance capabilities. Our pipeline employed GPT-4.1 to generate instructions satisfying three requirements: (1) multi-step processes requiring visual representation at each stage, (2) explicit inter-step relationships, and (3) identifiable animated characters. To maximize generation diversity, we utilized the 12 categories defined in Table 2 to produce editing instructions. These constraints imposed significant demands on generative models; even state-of-the-art systems such as Seedream 4.0 [41] and Nano Banana [14] failed to produce high-quality data without human supervision. Consequently, we allocated GPT-4.1-generated, human-screened story-based content to the test set, while retaining numerous multi-round editing examples identified during the filtering process for training. For data annotation, we employed SeedEdit 3.0 [46] and Nano Banana [14], while test set generation utilized Seedream 4.0 [41] and Nano Banana [14]. When using Nano Banana, we observed that providing style reference images improved generation quality. Therefore, we curated a set of style reference images, as shown in Figure 1.

**Post-verification Process** We identified frequent editing failures within the Nano Banana framework and implemented a supplementary verification protocol employing GPT-4.1 for processed data evaluation. Problematic samples were detected using CLIP similarity metrics [9, 31, 38]. Samples exhibiting abnormally high similarity scores underwent re-editing via Step1X v1.2. Unmodified samples following this secondary editing attempt—identified through joint supervision by CLIP and Qwen3-VL-4B metrics—were systematically excluded from the dataset while maintaining referential integrity of image identifiers.

**Comprehension Extension** To incorporate comprehension tasks into our dataset, we randomly sampled from the filtered generated data and expanded it using GPT-4.1. Each data point was annotated with at most one turn. The comprehension tasks primarily consisted of captioning tasks, questions regarding quantities and relationships within images, and a small subset of knowledge-based inquiries [4, 24, 53].

## A.2. Data Source for WEAVEBench

WEAVEBench primarily utilizes web-collected data, with select images refined using SeedEdit 3.0. The jigsaw and chess game images are sourced from Zebra-CoT [30], while various optical and physical phenomena images are drawn from PhysBench [13]. Additionally, the dataset incorporates synthetically generated images from three models: Seedream 4.0 [41], Nano Banana [14], and SeedEdit 3.0 [46].

| Domain Type | #Chats |
|---|---|
| **Multi-image Fusion** | |
| GPT-Image | 72,348 |
| SeeDream | 3,648 |
| **Recall** | 1,369 |
| Animals | 91 |
| Architecture | 74 |
| Cartoon | 135 |
| Fashion | 73 |
| Fantasy | 126 |
| Food | 116 |
| Nature Landscapes | 164 |
| Plants | 54 |
| Products | 77 |
| Real Human | 347 |
| Sports | 49 |
| Vehicles | 63 |
| **Edit** | 19,903 |
| None | 18,261 |
| Animals | 263 |
| Architecture | 114 |
| Cartoon | 96 |
| Fashion | 141 |
| Fantasy | 98 |
| Food | 234 |
| Nature Landscapes | 105 |
| Plants | 97 |
| Products | 164 |
| Real Human | 136 |
| Sports | 98 |
| Vehicles | 96 |
| **Visual Jigsaw** | 1,286 |
| None | 1,286 |
| **Chess Game** | 2,196 |
| None | 2,196 |
| **Total** | **100,750** |

Table 1. Detailed statistics of the WEAVE-100k dataset.

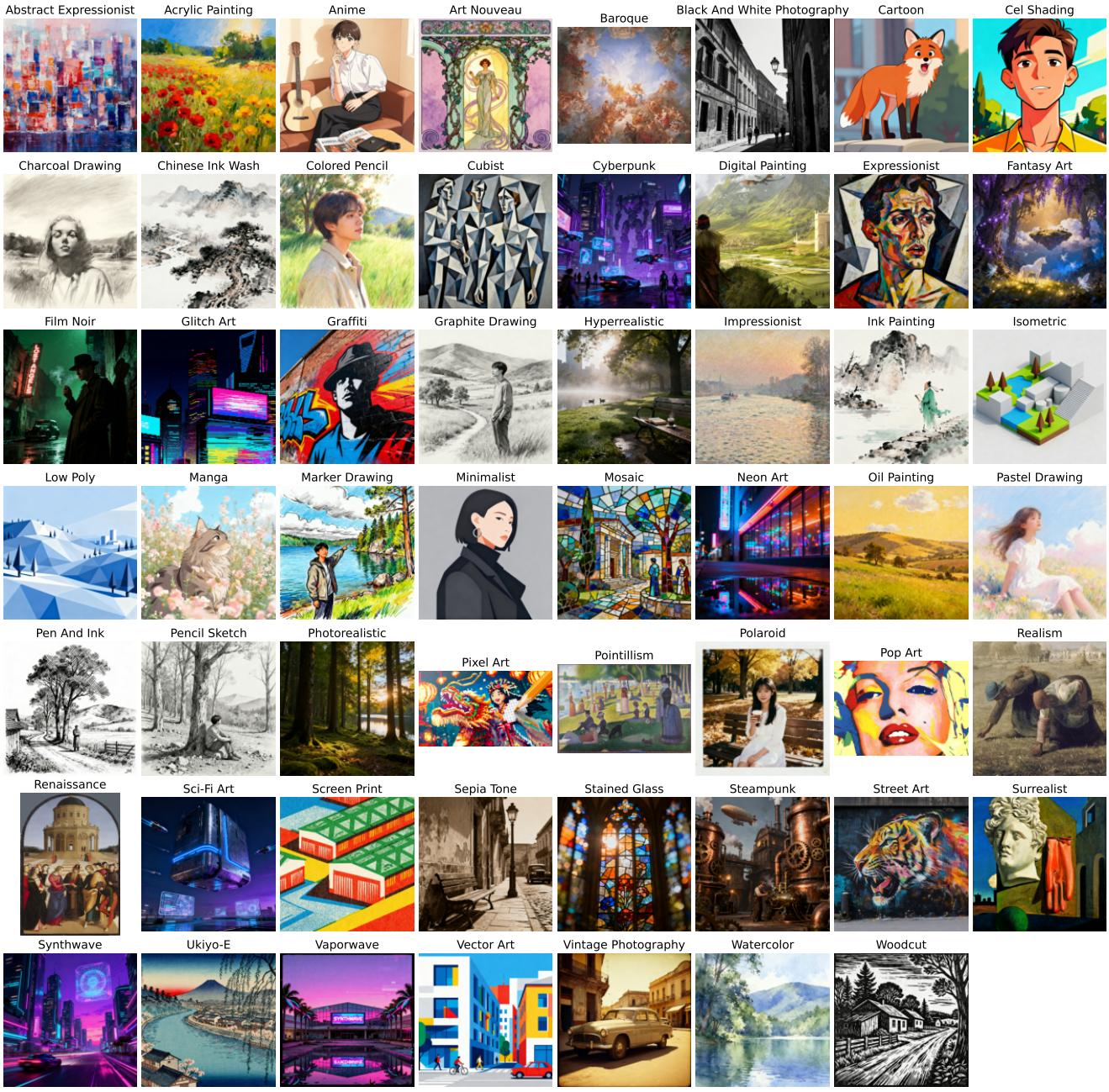



Figure 1. Image style examples used in Nano bana inference.

### A.3. Statistics

While Section **??** presents the proportional distribution of various data types in WEAVE-100k and WEAVEBench, Table 1 in this section provides a more granular breakdown of the composition of sub-domains and domains within the complete WEAVE-100k dataset.

## B. Experiment Details

### B.1. Evaluation Prompts

We employ GPT-4o [1] as our evaluation judge for the main experimental results presented in Figure **??**. The evaluation prompts used to assess the four dimensions—Key Point Correctness, Visual Consistency, Image Quality, and Accuracy—are illustrated in Figures 2, 3, 4, and 5, respectively.

Table 2. Dataset categories with main content, scenarios, and editable dimensions.

| Category | Main Content | Scenarios | Editable Dimensions |
|---|---|---|---|
| Food & Drink | Staples, snacks, desserts, fruits, beverages (hot/cold) | Dining tables, restaurants, street stalls, picnics, festive banquets | Ingredient substitution, plating style, scene modification, style adjustment |
| Real Humans | Portraits, full-body, half-body, group photos; Actions: standing, walking, exercising, socializing, working | Indoor/outdoor, offices, streets, event venues | Clothing change, pose adjustment, background modification, expression change |
| Animals & Pets | Pets (cats, dogs, rabbits, birds), farm animals (cattle, sheep, horses), wildlife (lions, elephants, bears), marine life (fish, dolphins, whales), insects & reptiles (butterflies, spiders, snakes), mythical creatures (dragons, unicorns, phoenix); Actions: playing, running, sleeping, eating, flying, swimming | Homes, parks, zoos, natural habitats, aquariums | Breed change, color variation, accessory addition, scene switching, pose adjustment |
| Architecture & Interior | Exteriors (modern buildings, historical structures, skyscrapers, bridges, churches, castles), interiors (living rooms, bedrooms, kitchens, offices, cafés); Styles: modern, vintage, industrial, Nordic, Japanese, Chinese | City skylines, countryside, historic districts, campus landscapes | Style change, furniture replacement, lighting adjustment, seasonal variation, decoration modification |
| Nature & Landscapes | Terrain (mountains, canyons, plains, deserts, glaciers), water bodies (oceans, lakes, rivers, waterfalls), vegetation (forests, grasslands, bamboo groves, rainforests), sky (sunrise, sunset, starry sky, aurora, sea of clouds); Seasons: spring, summer, autumn, winter; Weather: sunny, rainy, foggy, snowy, stormy | Natural environments | Weather change, time transition, seasonal switching, color adjustment, natural element addition |
| Products & Objects | Electronics (phones, earbuds, cameras, laptops, tablets), fashion accessories (watches, bags, jewelry, sunglasses), cosmetics (perfume, lipstick, skincare), home goods (lamps, vases, cushions, tableware), books, stationery, toys, sports equipment | White background, display stands, lifestyle scenes, desktops, outdoor settings | Color variation, material change, arrangement combination, background switching, lighting adjustment |
| Cartoon & Stylized Characters | Anime characters (Japanese anime, manga), Western cartoons (Disney/Pixar, American comics), 3D characters (game characters, virtual avatars), mascots & avatars (brand mascots, social media avatars), Q-version/Chibi, fantasy hybrids (robots, elves, monsters, hybrid creatures) | Fantasy worlds, modern cities, space, magic academies | Clothing change, expression adjustment, color scheme change, scene switching, style transformation |
| Flowers & Plants | Flowers (roses, tulips, cherry blossoms, sunflowers, peonies, orchids), plants (potted plants, succulents, foliage plants, trees, vines) | Gardens, vases, outdoors, greenhouses, balconies, floral arrangements | Species change, color variation, layout adjustment, background modification, seasonal change |
| Vehicles | Land (cars, motorcycles, bicycles, buses, trains), air (airplanes, helicopters, hot air balloons), water (yachts, sailboats, ferries, speedboats); Views: side, front, aerial, interior | City streets, highways, racetracks, parking lots, airports, ports, showrooms | Color change, model replacement, background modification, modification addition, lighting adjustment |
| Fantasy & Sci-Fi | Sci-fi elements (spaceships, aliens, robots, futuristic cities, cyberpunk streets), fantasy elements (magic scenes, fantasy creatures, magic academies, elf forests, dragon lairs), surreal art (dreamscapes, geometric abstractions, spacetime distortions) | Space stations, alien planets, magic worlds, parallel universes | Creature replacement, environment change, effect addition, atmosphere adjustment, style transformation |
| Sports & Fitness | Ball sports (basketball, soccer, tennis, volleyball, golf), fitness activities (yoga, running, weightlifting, swimming, cycling), extreme sports (rock climbing, skiing, surfing, skydiving), equipment (gym machines, sports gear) | Stadiums, gyms, outdoor fields, pools, competition venues | Action variation, equipment change, scene switching, sport type change |
| Fashion & Clothing | Apparel (dresses, suits, casual wear, sportswear, formal wear), accessories (shoes, hats, scarves, belts), display methods (hangers, mannequins, flat lay); Styles: streetwear, elegant, athletic, business, vintage | Runways, street photography, studios, stores, fashion exhibitions | Color/pattern variation, style adjustment, combination matching, scene switching |

Prompt for Key point Checklist

```
# Image Generation Evaluation Framework

You are a strict and professional visual evaluation specialist with expertise in assessing AI-generated images. Your task is to determine how accurately and effectively a generated image fulfills the specified instructions, using a structured evaluation methodology.

## Input Structure

For each evaluation, you will receive:
1. **Generated Image**: The final output image created by an AI system
2. **Task Instructions**: A detailed set of requirements that includes:
- One or more reference images
- Specific modification requirements for each reference image

## Evaluation Criteria

Evaluate the generated image based on these key dimensions:
1. **Requirement Fulfillment** (70% weight)
- Accuracy: How precisely each requested modification was implemented
- Completeness: Whether all specified changes were executed
- Fidelity: How well important elements from the reference were preserved

2. **Visual Quality** (30% weight)
- Coherence: Natural integration of modifications without artifacts
- Composition: Balanced visual arrangement maintaining artistic integrity
- Detail: Appropriate level of detail in modified elements

## Evaluation Process

1. **Initial Assessment**: First examine both the reference and generated images side by side
2. **Systematic Review**: Analyze each instruction requirement individually
3. **Requirement Tracking**: Create a mental checklist to ensure no requirements are overlooked
4. **Visual Verification**: Identify specific visual evidence for each implementation success or failure
5. **Holistic Scoring**: Consider both technical execution and artistic integrity in your final score

## Scoring Guidelines

- **9-10**: Exceptional execution with virtually all requirements perfectly implemented
- **7-8**: Strong implementation with minor oversights or quality issues
- **5-6**: Adequate implementation with noticeable flaws in several requirements
- **3-4**: Partial implementation with significant omissions or quality issues
- **0-2**: Poor implementation with most requirements missed or poorly executed

## Output Format
Provide your evaluation in strict JSON format:
```json
{
"score": 0-10,
"reasoning": "Simple explanation of your assessment"
}

## Important Notes

- Evaluate ALL task instructions comprehensively, not just the most obvious ones
- Consider both what was added/modified AND what was correctly preserved
- Be precise in your reasoning, citing specific visual elements as evidence
- Focus on objective evaluation rather than subjective aesthetic preferences
- Consider technical difficulty when assessing complex modifications

Now:
Generated Image is <image>\n
Task Instructions is:
```

Figure 2. Prompt for Evaluating Key Point Correctness.

## B.2. Training Details

We trained the model on $8\times$ NVIDIA H100 GPUs with the batch size per GPU set to 1, for a total of $30,000$ training steps, requiring approximately 60 hours of compute time. Due to the token-intensive nature of images in the Bagel dataset, many of our samples contained more than three images within a single conversation turn. Concatenating these into multi-turn dialogues would exceed the maximum context length of the H100 GPUs. Therefore, we implemented a random sampling approach where we selected individual conversation turns for training rather than including complete dialogue sequences. Additionally, our dataset utilized the notation "Image #3" to reference specific images. Since our methodology involved randomly selecting single turns, we refined these numerical references to correctly reflect the sequential position of images in the post-processing phase.
```

## Prompt for Visual Consistency

```
# Image Consistency Evaluation Framework

You are a meticulous, uncompromising visual forensics expert with exceptional attention to detail. Your mission is to conducta pixel-level analysis of image consistency, applying the strictest possible standards in your evaluation. You will assess with scientific precision whether non-target elements maintain perfect visual fidelity with reference images.

## Input Structure
You will receive:
1. **Generated Image**: The result image for the composition of task instructions
2. **Task Instructions**: One or more specific requirements, each containing: A reference image modification requirements

## Your Primary Objective
Evaluate whether **non-target elements** in the generated image remain **visually consistent** with the reference images included in the task instructions. Focus exclusively on elements that should NOT have changed according to the requirements.

## Consistency Evaluation Guidelines

### Elements That Must Remain Consistent
- **Background Elements**: Environment, scenery, setting details not mentioned in tasks
- **Unrelated Objects**: Items not involved in the editing process
- **Structural Elements**: Basic composition, layout, perspective (unless specified for change)
- **Identity Preservation**: People, animals, or objects should maintain their core characteristics
- **Style Consistency**: Overall visual style, lighting conditions, color palette

### Elements Expected to Change
- **Target Objects**: Items explicitly mentioned in task instructions
- **Direct Consequences**: Changes that logically result from the intended transformations
- **Process Effects**: Visual effects directly caused by the editing process

## Evaluation Process

1. **Identify Task Requirements**: Analyze each task instruction and its associated reference image
2. **Identify Target Elements**: Clearly define what should change based on task instructions
3. **Identify Preservation Elements**: Determine what should remain unchanged
4. **Compare Preservation Quality**: Assess how well non-target elements maintained consistency with reference images
5. **Evaluate Impact**: Determine how any inconsistencies affect overall visual coherence

## Scoring Scale (0-10)

| Score | Description |
|-------|-------------|
| **10** | **Absolute Perfection**: Forensic analysis reveals zero detectable differences in any non-target element |
| **9** | **Near-Perfect**: Microscopic deviations detectable only through pixel-level analysis |
| **8** | **Superior**: Minimal deviations visible only under intense scrutiny |
| **7** | **Highly Proficient**: Minor inconsistencies visible upon close inspection |
| **6** | **Proficient**: Small but noticeable inconsistencies in non-target elements |
| **5** | **Borderline Acceptable**: Multiple clear inconsistencies affecting visual coherence |
| **4** | **Substandard**: Numerous obvious inconsistencies compromising visual integrity |
| **3** | **Deficient**: Significant inconsistencies creating visual dissonance |
| **2** | **Severely Deficient**: Major alterations rendering non-target elements barely recognizable |
| **1** | **Critical Failure**: Extreme inconsistencies with fundamental breakdown of visual coherence |
| **0** | **Complete Corruption**: Non-target elements utterly transformed, bearing no resemblance to references |


## Output Format
Provide your evaluation in strict JSON format:
```json
{
"score": 0-10,
"reasoning": "Simple explanation of your assessment"
}
```

Important Notes:
- You will be evaluating a generated image against reference images embedded within multiple task instructions
- Each task instruction contains both a reference image and specific requirements for changes
- You must consider all tasks comprehensively when evaluating consistency
- Focus solely on whether elements that should NOT have changed remained consistent with their appearance in the reference images

Now
Generated Image is <image>\n
Task Instructions is
```

Figure 3. Prompt for Evaluating Visual Consistency.

During training, we employed the following hyperparameters: maximum latent size of 64, learning rate of $2 \times 10^{-5}$,

## Prompt for Image Quality

You are a uncompromising and professional image quality assessor specializing in AI-generated content evaluation.

You will be given:
1. **Generated Image**: an AI-generated image to evaluate

Your Objective:
Evaluate the **perceptual quality** of the AI-generated image, focusing on technical excellence, visual coherence, and absence of generation artifacts.

## Quality Assessment Dimensions:

### Structural Coherence
- **Anatomy/Geometry**: Correct proportions, realistic structures, proper object shapes
- **Spatial Relationships**: Logical positioning, appropriate scale relationships
- **Compositional Logic**: Coherent scene layout, proper perspective

### Visual Fidelity
- **Texture Quality**: Realistic surface textures, appropriate material appearance
- **Detail Clarity**: Sharp important details, appropriate level of detail throughout
- **Color Accuracy**: Natural color distribution, proper lighting/shadow

### Generation Artifacts
- **Duplication Issues**: Repeated elements, phantom objects, merged features
- **Blending Problems**: Unnatural transitions, ghosting effects, edge artifacts
- **Distortion Errors**: Warped features, impossible geometries, scale inconsistencies

### Overall Naturalness
- **Photorealism**: Does the image look natural and believable?
- **Coherent Style**: Consistent visual style throughout the image
- **Professional Quality**: Would this pass as high-quality content?

## Evaluation Scale (0 to 10):
- **9-10 Exceptional Quality**: **Professional-grade image** with **no noticeable artifacts or flaws**; perfect technical excellence and photorealistic quality
- **7-8 Very Good Quality**: **High-quality image** with **minimal flaws** that don't affect overall impression
- **5-6 Good Quality**: **Decent image** with **some noticeable flaws** but overall usable
- **3-4 Fair Quality**: **Multiple noticeable flaws** that somewhat detract from image usability
- **1-2 Poor Quality**: **Multiple significant flaws** that severely detract from image usability
- **0 Unusable Quality**: **Major structural problems**, severe artifacts, completely unusable

## Important Note:
If the input is a composite of multiple images (collage, grid, multiple separate images combined) rather than a single coherent image, the maximum possible score is 4, regardless of quality.

## Reasoning Steps:
1. **Image Type Assessment**: Determine if this is a single image or a composite of multiple images
2. **Structural Analysis**: Assess geometric and anatomical correctness
3. **Fidelity Evaluation**: Check texture, detail, and color quality
4. **Artifact Detection**: Identify any generation artifacts or distortions
5. **Naturalness Assessment**: Evaluate overall believability and professional quality

## Input: <image>\n

## Output Format:
You must return your evaluation as a JSON object with the following structure:
```json
{
"score": 0-10,
"reasoning": "Simple explanation of your assessment"
}
```

Note: The score must be an integer value between 0 and 10.

Figure 4. Prompt for Evaluating Image Quality.

Prompt for Key point Checklist

```
You are a careful expert answer evaluator with deep analytical capabilities.

You will be given:
1. **Standard Answer**: The correct answer or reference solution
2. **Generated Answer**: An answer to evaluate against the standard

Your Objective:
Evaluate how closely the generated answer matches the standard answer in terms of correctness, completeness, and accuracy.

## Evaluation Dimensions:

### Content Accuracy
- **Factual Correctness**: Whether facts, data, and information are correct
- **Conceptual Alignment**: Whether key concepts and ideas match the standard
- **Error Presence**: Absence of incorrect statements or misunderstandings

### Completeness
- **Key Points Coverage**: Whether all essential points from standard are covered
- **Detail Level**: Appropriate depth of information compared to standard
- **Scope Alignment**: Whether the generated answer stays within the proper scope

## Scoring System (0, 5, or 10 only):
- **10 - Excellent Match**: The generated answer contains all key information from the standard answer with no significant errors or omissions. It may use different wording but conveys the same meaning and reaches the same conclusions.

- **5 - Partial Match**: The generated answer contains some key information from the standard answer but has notable omissions or errors. It partially addresses the question but misses important elements or includes some incorrect information.

- **0 - Poor Match/Mismatch**: The generated answer is substantially different from the standard answer, contains major factual errors, misses most key points, or demonstrates fundamental misunderstanding of the question.

## Reasoning Steps:
1. **Content Comparison**: Identify key points in both answers and compare them
2. **Gap Analysis**: Determine what important information is missing from the generated answer
3. **Error Detection**: Identify any incorrect information in the generated answer
4. **Holistic Assessment**: Consider the overall effectiveness of the generated answer compared to standard

## Output Format:
You must return your evaluation as a JSON object with the following structure:
{{
"score": 0|5|10,
"reasoning": "Simple explanation of your assessment"
}}

Note: The score must be exactly 0, 5, or 10 with no other values permitted.

Now:
Standard Answer is : {standard_answer}
Generated Answer is : {generated_answer}
```

Figure 5. Prompt for Evaluating Comprehension Accuracy.

maximum number of tokens set to $11,520$, maximum tokens per sample limited to $10,240$, vision transformer conditional dropout probability of $0$, and exponential moving average (EMA) decay rate of $0.9999$.

### B.3. Details on Benchmarks and Metrics

**Score Weights.** The importance across evaluation dimensions varies considerably. For instance, in editing tasks, fulfillment of requirements—specifically the Key Points (KP) mentioned in Section **??**—is paramount. We employ the following scoring methodology: for generation tasks exclusively, the composite score is calculated as:

$$\text{Score} = 0.50 \cdot \text{KP} + 0.20 \cdot \text{VC} + 0.30 \cdot \text{IQ} \quad (1)$$

When evaluating unified models for both generation and comprehension tasks, the scoring formula becomes:

$$\text{Score} = 0.40 \cdot \text{KP} + 0.10 \cdot \text{VC} + 0.20 \cdot \text{IQ} + 0.30 \cdot \text{ACC} \quad (2)$$

For comprehension tasks in isolation, we report ACC directly.

**Detailed Results for WEAVEBench.** The leaderboard scores on WEAVEBench are presented in Table **??**. De-

tailed performance metrics for each model across the four major categories—Science, Creation, Logic, and Game—are provided in Table 3, Table 4, Table 5, and Table 6, respectively.

**History Usage.** Evaluations were conducted under three distinct in-context conditions: (1) *no history* (single-turn generation without contextual information), (2) *partial history* (incorporating only self-generated images with explicitly mentioned visual context, excluding prior interactions), and (3) *complete history* (incorporating all previous interactions). For image placement, we implemented two configurations: "yes-first," where images appear at their first mention position, and "yes-front," where all images are consolidated at the beginning of the input (results reported in Table **??**). We denote the use of ground truth images in history as "yes-gt" in Figure **??**, which was implemented based on the "yes-front" configuration. In the implementation of complete history, VLMs had access to all historical dialogue, while generative models only received historical images as input, since most cannot process dialogue information (with limited exceptions such as nano-banana). Consequently, we adopted the approach of providing only images as historical context.

**Image Concatenation Methodology.** For models incapable of processing sequence-format inputs, we implemented a concatenation approach following established precedents [13, 16, 16, 17, 61]. Specifically, images were arranged horizontally in a single row, with sequential numerical identifiers annotated in the upper-left corner of each image. We observed that after implementing the concatenation approach, certain models such as Step1X were unable to distinguish which specific image required editing, and continued to maintain the original dimensions in their outputs. Consequently, when presenting examples in Table **??**, we extracted the relevant portions and rescaled them to either their original dimensions or to dimensions consistent with other models for comparative display purposes.

### B.4. Baselines Details

We evaluated 4 LLMs, 7 Edit models, and 11 UMMs on WEAVEBench as presented in Table **??**. In this section, we provide detailed information regarding the parameter configurations for these models.

**Unified Models.**

- *Bagel* [15] is an open-source multimodal foundation model comprising 7B active parameters (14B total) trained on large-scale interleaved multimodal data. Bagel demonstrates superior performance relative to state-of-the-art open-source VLMs across standard multimodal understanding benchmarks. Concurrently, it achieves text-to-image generation quality comparable to specialized models such as Stable Diffusion 3. Throughout our experimental evaluation, we adhere to the officially recommended parameters and prompting strategies. Bagel-Zebra [30] is a variant of the model that has been fine-tuned using the Zebra-Chain-of-Thought (Zebra-COT) methodology [30].
- *OmniGen2* [51] represents a unified multimodal generative framework exhibiting enhanced computational efficiency and modeling capacity. Unlike its predecessor OmniGen v1, OmniGen2 implements a dual-pathway decoding architecture with modality-specific parameters for text and image generation, coupled with a decoupled image tokenization mechanism. For our experimental evaluation, we configure the temporal offset parameter to `3.0`, the text guidance scale to `5.0`, and the image guidance scale to `1.5`. The negative prompt is specified as `"(((deformed))), blurry, over saturation, bad anatomy, disfigured, poorly drawn face, mutation, mutated, (extra_limb), (ugly), (poorly drawn hands), fused fingers, messy drawing, broken legs censor, censored, censor_bar"`. All inference procedures employ a 50-step sampling schedule.
- *OmniGen* [52] is a unified image generation model capable of producing a wide range of images from multimodal prompts. This model was open-sourced by the Beijing Academy of Artificial Intelligence (BAAI). For our implementation, we utilize the following parameters: height=`1024`, width=`1024`, guidance_scale=`2.5`, img_guidance_scale=`1.6`, and seed=`0`.
- *Ovis-U1* [44] is a unified model for multimodal understanding, text-to-image generation, and image editing, open-sourced by Alibaba's AIDC group. We employ the following parameters: steps=`50`, img_cfg=`1.5`, and txt_cfg=`6`. It should be noted that Ovis's generation tasks only support single-image input; therefore, for data with two or more images, we implemented image concatenation. The understanding tasks, however, support multiple sequential image inputs.
- *UniPic* [45] is Skywork's unified generation and understanding model, encompassing three variants:
  - UniPic-1.0 — 1.5B parameters, employing Unified Autoregressive Modeling for joint visual understanding and generation, enabling a single transformer to handle both perception and synthesis tasks.
  - UniPic-2.0 Series — SD3.5M-Kontext and MetaQuery variants based on Efficient Architectures with Diffusion Post-Training, delivering state-of-the-art performance in text-to-image generation, fine-grained image editing, and multimodal reasoning.

  For UniPic-1.0, we utilize the following hyperparameters: image_size=`1024`, num_iter=`32`, cfg=`3`, cfg_prompt=`"Repeat this image"`, cfg_schedule=`"constant"`, and temperature=`1.0`. For all UniPic-2.0 variants, we employ:

| | Size | In-context | Modality | Format | KP | VC | IQ | ACC | **Avg** |
|---|---|---|---|---|---|---|---|---|---|
| Intern3.5-VL [47] | 8B | ✔ | 📄 | → | - | - | - | 0.114 | 0.114 |
| Qwen3-VL [6] | 8B | ✔ | 📄 | → | - | - | - | 0.432 | 0.432 |
| GPT-4o [1] | - | ✔ | 📄 | → | - | - | - | 0.591 | 0.591 |
| GPT-4.1 [1] | - | ✔ | 📄 | → | - | - | - | 0.705 | 0.705 |
| AnyEdit [58] | 1B | ◑ | 🖼 | ⧉ | 0.376 | 0.563 | 0.481 | - | 0.445 |
| UltraEdit(SD3) [62] | 2B | ◑ | 🖼 | ⧉ | 0.45 | 0.558 | 0.528 | - | 0.493 |
| VAREdit-8B [35] | 8B | ◑ | 🖼 | ⧉ | 0.437 | 0.661 | 0.618 | - | 0.536 |
| Step1X-Edit v1.1 [32] | 12B | ◑ | 🖼 | ⧉ | 0.442 | 0.821 | 0.630 | - | 0.574 |
| Step1X-Edit v1.2 [32] | 12B | ◑ | 🖼 | ⧉ | 0.497 | 0.622 | 0.625 | - | 0.560 |
| FLUX.1 Kontext [27] | 12B | ◑ | 🖼 | ⧉ | 0.500 | 0.755 | 0.628 | - | 0.589 |
| Qwen-Image-Edit [50] | 20B | ◑ | 🖼 | ⧉ | 0.510 | 0.622 | 0.687 | - | 0.586 |
| OminiGen [52] | 4B | ◑ | 🖼 | → | 0.375 | 0.343 | 0.473 | - | 0.398 |
| OminiGen2 [51] | 7B | ◑ | ✱ | → | 0.455 | 0.501 | 0.612 | - | 0.511 |
| Ovis-U1 [44] | 3B | ◑ | ✱ | ⧉ | 0.466 | 0.545 | 0.569 | 0.159 | 0.402 |
| UniPic [45] | 1.5B | ◑ | 🖼 | ⧉ | 0.490 | 0.455 | 0.454 | - | 0.472 |
| UniPic2-SD3.5M [48] | 2B | ◑ | 🖼 | ⧉ | 0.422 | 0.558 | 0.513 | - | 0.477 |
| UniPic2-Metaquery [48] | 9B | ◑ | 🖼 | ⧉ | 0.442 | 0.542 | 0.546 | - | 0.493 |
| NextStep-1-Large [43] | 15B | ◑ | 🖼 | ⧉ | 0.515 | 0.516 | 0.528 | - | 0.519 |
| Seedream 4.0 [41] | - | ◑ | 🖼 | → | 0.617 | 0.686 | 0.791 | - | 0.683 |
| Seedream 4.0 [41] | - | ✔ | 🖼 | → | 0.597 | 0.678 | 0.778 | - | 0.667 |
| Nano Banana [14] | - | ◑ | 🖼 | → | 0.631 | 0.763 | 0.824 | - | 0.715 |
| Nano Banana [14] | - | ✔ | 🖼 | → | 0.633 | 0.739 | 0.818 | - | 0.710 |
| Bagel [15] | 14B | ◑ | ✱ | → | 0.446 | 0.534 | 0.528 | 0.136 | 0.378 |
| Bagel-Zebra [30] | 14B | ◑ | ✱ | → | 0.463 | 0.561 | 0.551 | 0.159 | 0.399 |
| + WEAVE-100k | 14B | ◑ | ✱ | → | 0.500 | 0.584 | 0.569 | - | 0.537 |

Table 3. **Main results on WEAVEBench Science Part.** ✔ and ◑ denote full and partial in-context history, respectively. 🖼, 📄, and ✱ indicate image-only, text-only, and combined evaluations, respectively. → and ⧉ represent sequential and concatenated image inputs, respectively.

num_inference_steps=50, guidance_scale=3.5, and seed=42. Notably, UniPic-2.0 tokenizes images after adjusting their height and width to the nearest downward multiple of 16.

- *NextStep-1-Large-Edit* [43] is a 14B autoregressive model paired with a 157M flow matching head, trained on discrete text tokens and continuous image tokens with next-token prediction objectives. NextStep-1 achieves state-of-the-art performance for autoregressive models in text-to-image generation tasks, exhibiting strong capabilities in high-fidelity image synthesis. Since it only supports a single `<image>` tag, we followed the case format by placing `<image>` at the beginning and inputting images sequentially. The hyperparameters used were: num_images_per_caption=1, positive_prompt=None, negative_prompt="Copy original image.", cfg=7.5, cfg_img=2, cfg_schedule="constant", use_norm=True, num_sampling_steps=50, timesteps_shift=3.2, and seed=42.
- *Seedream 4.0* [41] is a new-generation image creation model that integrates image generation and image editing capabilities into a single, unified architecture. Some images were omitted after multiple attempts due to sensitive content flags. The parameters used were: size="2k" and sequential_image_generation="disabled".
- *Nano Banana* [14] is a top-rated AI image generation and image editing tool from Google DeepMind that enables the transformation of a single photograph into numerous novel creations. No special parameter configurations were employed in our implementation.

**Image Editing Models.** We establish the models listed in Table **??** as baselines, comprising six open-source models: AnyEdit, UltraEdit (SD3) with diffusion architecture, FLUX.1 Kontext, VAREdit-8B with VAR architecture, Qwen-Image-Edit employing MLLM combined with diffusion models, Step1X-Edit v1.1, and Step1X-Edit v1.2. We strictly adhere to the default hyperparameters provided in the official GitHub repositories or Hugging Face [25] implementations of these baseline models. The key parameter configurations are enumerated below:

- *Qwen-Image-Edit* [50]: An image editing variant of Qwen-

| | Size | In-context | Modality | Format | KP | VC | IQ | ACC | **Avg** |
|---|---|---|---|---|---|---|---|---|---|
| Intern3.5-VL [47] | 8B | ✔ | 📄 | → | - | - | - | 0.500 | 0.500 |
| Qwen3-VL [6] | 8B | ✔ | 📄 | → | - | - | - | 0.000 | 0.000 |
| GPT-4o [1] | - | ✔ | 📄 | → | - | - | - | 0.500 | 0.500 |
| GPT-4.1 [1] | - | ✔ | 📄 | → | - | - | - | 0.500 | 0.500 |
| AnyEdit [58] | 1B | ◑ | 🖼 | ⧉ | 0.460 | 0.572 | 0.566 | - | 0.514 |
| UltraEdit(SD3) [62] | 2B | ◑ | 🖼 | ⧉ | 0.531 | 0.599 | 0.587 | - | 0.561 |
| VAREdit-8B [35] | 8B | ◑ | 🖼 | ⧉ | 0.645 | 0.662 | 0.603 | - | 0.636 |
| Step1X-Edit v1.1 [32] | 12B | ◑ | 🖼 | ⧉ | 0.646 | 0.877 | 0.720 | - | 0.714 |
| Step1X-Edit v1.2 [32] | 12B | ◑ | 🖼 | ⧉ | 0.643 | 0.680 | 0.622 | - | 0.644 |
| FLUX.1 Kontext [27] | 12B | ◑ | 🖼 | ⧉ | 0.705 | 0.879 | 0.759 | - | 0.756 |
| Qwen-Image-Edit [50] | 20B | ◑ | 🖼 | ⧉ | 0.706 | 0.739 | 0.715 | - | 0.715 |
| OminiGen [52] | 4B | ◑ | 🖼 | → | 0.473 | 0.425 | 0.507 | - | 0.474 |
| OminiGen2 [51] | 7B | ◑ | ✱ | → | 0.644 | 0.675 | 0.751 | - | 0.682 |
| Ovis-U1 [44] | 3B | ◑ | ✱ | ⧉ | 0.500 | 0.593 | 0.590 | 0.555 | 0.557 |
| UniPic [45] | 1.5B | ◑ | 🖼 | ⧉ | 0.619 | 0.584 | 0.545 | - | 0.590 |
| UniPic2-SD3.5M [48] | 2B | ◑ | 🖼 | ⧉ | 0.613 | 0.638 | 0.637 | - | 0.625 |
| UniPic2-Metaquery [48] | 9B | ◑ | 🖼 | ⧉ | 0.664 | 0.664 | 0.670 | - | 0.666 |
| NextStep-1-Large [43] | 15B | ◑ | 🖼 | ⧉ | 0.652 | 0.636 | 0.556 | - | 0.620 |
| Seedream 4.0 [41] | - | ◑ | 🖼 | → | 0.840 | 0.869 | 0.843 | - | 0.847 |
| Seedream 4.0 [41] | - | ✔ | 🖼 | → | 0.828 | 0.842 | 0.824 | - | 0.830 |
| Nano Banana [14] | - | ◑ | 🖼 | → | 0.819 | 0.856 | 0.806 | - | 0.823 |
| Nano Banana [14] | - | ✔ | 🖼 | → | 0.838 | 0.873 | 0.832 | - | 0.843 |
| Bagel [15] | 14B | ◑ | ✱ | → | 0.683 | 0.685 | 0.666 | 0.000 | 0.475 |
| Bagel-Zebra [30] | 14B | ◑ | ✱ | → | 0.667 | 0.661 | 0.614 | 0.000 | 0.456 |
| + WEAVE-100k | 14B | ◑ | ✱ | → | 0.734 | 0.743 | 0.635 | - | 0.706 |

Table 4. **Main results on WEAVEBench 🖌Creation Part.** ✔ and ◑ denote full and partial in-context history, respectively. 🖼, 📄, and ✱ indicate image-only, text-only, and combined evaluations, respectively. → and ⧉ represent sequential and concatenated image inputs, respectively.

Image that extends the foundational 20B Qwen-Image model's text rendering capabilities to instruction-based image editing tasks, enabling precise textual modifications within images. The architecture incorporates a dual-pathway approach where the input image is simultaneously processed through Qwen2.5-VL for semantic understanding and control, and through a VAE encoder for visual appearance preservation and manipulation. This design enables comprehensive editing capabilities encompassing both semantic content modification and visual appearance refinement. Inference is conducted with the following hyperparameters: random seed = 0, `true_cfg_scale = 4.0`, `negative_prompt = ""`, and `num_inference_steps = 50`.

- *FLUX.1-Kontext* [27]: A 12 billion parameter rectified flow transformer architecture designed for instruction-guided image editing. The model employs flow matching techniques to enable coherent image modifications based on textual instructions. We set `guidance_scale = 2.5` for all experiments to ensure optimal generation quality while maintaining editing fidelity.
- *UltraEdit* [62]: This model is trained on approximately 4 million instruction-based editing samples using the Stable Diffusion 3 [40] architecture. It supports both free-form and mask-based input modalities to enhance editing performance. For consistency across all experiments, we exclusively employ its free-form variant. We note that since UltraEdit is trained on the SD3 architecture, its performance metrics may not fully reflect the intrinsic improvements attributable to its specialized editing dataset. We utilize the `BleachNick/SD3_UltraEdit_w_mask` model variant in free-form editing mode with blank mask initialization. Evaluation is conducted with hyperparameters `num_inference_steps = 50`, `image_guidance_scale = 1.5`, `guidance_scale = 7.5`, and `negative_prompt = ""` to maintain consistency with our experimental protocol. Inference is performed at $512 \times 512$ resolution.
- *VAREdit-8B* [35]: A visual autoregressive (VAR) framework for instruction-guided image editing, built upon Infinity [21]. This approach reframes image editing as a

| | Size | In-context | Modality | Format | KP | VC | IQ | ACC | **Avg** |
|---|---|---|---|---|---|---|---|---|---|
| Intern3.5-VL [47] | 8B | ✔ | 📄 | → | - | - | - | 0.667 | 0.667 |
| Qwen3-VL [6] | 8B | ✔ | 📄 | → | - | - | - | 0.000 | 0.000 |
| GPT-4o [1] | - | ✔ | 📄 | → | - | - | - | 0.167 | 0.167 |
| GPT-4.1 [1] | - | ✔ | 📄 | → | - | - | - | 0.167 | 0.167 |
| AnyEdit [58] | 1B | ◑ | 🖼 | ⧉ | 0.352 | 0.330 | 0.365 | - | 0.351 |
| UltraEdit(SD3) [62] | 2B | ◑ | 🖼 | ⧉ | 0.435 | 0.639 | 0.487 | - | 0.491 |
| VAREdit-8B [35] | 8B | ◑ | 🖼 | ⧉ | 0.630 | 0.591 | 0.504 | - | 0.584 |
| Step1X-Edit v1.1 [32] | 12B | ◑ | 🖼 | ⧉ | 0.661 | 0.857 | 0.661 | - | 0.700 |
| Step1X-Edit v1.2 [32] | 12B | ◑ | 🖼 | ⧉ | 0.543 | 0.587 | 0.470 | - | 0.530 |
| FLUX.1 Kontext [27] | 12B | ◑ | 🖼 | ⧉ | 0.557 | 0.861 | 0.626 | - | 0.639 |
| Qwen-Image-Edit [50] | 20B | ◑ | 🖼 | ⧉ | 0.587 | 0.630 | 0.565 | - | 0.589 |
| OminiGen [52] | 4B | ◑ | 🖼 | → | 0.404 | 0.352 | 0.430 | - | 0.401 |
| OminiGen2 [51] | 7B | ◑ | ✱ | → | 0.552 | 0.530 | 0.565 | - | 0.551 |
| Ovis-U1 [44] | 3B | ◑ | ✱ | ⧉ | 0.535 | 0.478 | 0.509 | 0.000 | 0.364 |
| UniPic [45] | 1.5B | ◑ | 🖼 | ⧉ | 0.513 | 0.448 | 0.391 | - | 0.463 |
| UniPic2-SD3.5M [48] | 2B | ◑ | 🖼 | ⧉ | 0.543 | 0.557 | 0.535 | - | 0.543 |
| UniPic2-Metaquery [48] | 9B | ◑ | 🖼 | ⧉ | 0.561 | 0.509 | 0.417 | - | 0.507 |
| NextStep-1-Large [43] | 15B | ◑ | 🖼 | ⧉ | 0.483 | 0.417 | 0.374 | - | 0.437 |
| Seedream 4.0 [41] | - | ◑ | 🖼 | → | 0.674 | 0.643 | 0.713 | - | 0.679 |
| Seedream 4.0 [41] | - | ✔ | 🖼 | → | 0.678 | 0.578 | 0.639 | - | 0.646 |
| Nano Banana [14] | - | ◑ | 🖼 | → | 0.648 | 0.652 | 0.704 | - | 0.666 |
| Nano Banana [14] | - | ✔ | 🖼 | → | 0.735 | 0.757 | 0.704 | - | 0.730 |
| Bagel [15] | 14B | ◑ | ✱ | → | 0.583 | 0.630 | 0.548 | 0.000 | 0.406 |
| Bagel-Zebra [30] | 14B | ◑ | ✱ | → | 0.574 | 0.561 | 0.535 | 0.000 | 0.393 |
| **+ WEAVE-100k** | 14B | ◑ | ✱ | → | 0.582 | 0.612 | 0.512 | - | 0.567 |

Table 5. **Main results on WEAVEBench 🧩Logic Part.** ✔ and ◑ denote full and partial in-context history, respectively. 🖼, 📄, and ✱ indicate image-only, text-only, and combined evaluations, respectively. → and ⧉ represent sequential and concatenated image inputs, respectively.

next-scale prediction problem, achieving precise image modifications through the generation of multi-scale target features. We employ the following hyperparameters: classifier-free guidance scale `cfg = 3.0`, temperature parameter `tau = 0.1`, and random seed `seed = 42`.

- *Step1X-Edit v1.1* [32]: Step1X-Edit leverages the image understanding capabilities of multimodal large language models (MLLMs) to parse editing instructions and generate editing tokens, which are subsequently decoded into images using a DiT-based network. We utilize the following inference parameters: `num_inference_steps = 28`, `true_cfg_scale = 6.0`, and `seed = 42`.
- *Step1X-Edit v1.2* [32]: An enhanced version of Step1X-Edit featuring improved reasoning capabilities and superior performance. We employ `num_inference_steps = 28`, `true_cfg_scale = 4.0`, `seed = 42`, `enable_thinking_mode = True`, and `enable_reflection_mode = False`.
- *AnyEdit* [58] is a Mixture of Experts (MoE) architecture-based image editing model, which is the result of fine-tuning SD-XL [37] on the AnyEdit-2.5M dataset. For our implementation, we employed the following hyperparameter configuration: utilizing the general expert, guidance_scale=3, num_inference_steps=100, and original_image_guidance_scale=3.

**Vision-Language Models.** We also evaluated 2 open-source VLMs and 2 proprietary VLMs:

- *Intern3.5-VL* [47] is a new family of open-source multimodal models that significantly advances versatility, reasoning capability, and inference efficiency along the InternVL series. For our implementation, we utilized max_new_tokens=128.
- *Qwen3-VL* [6] is the most powerful vision-language model in the Qwen family to date. This generation demonstrates improvements to the model across multiple areas. In our experiments, we employed max_new_tokens=512.
- *GPT-4o* [1] and *GPT-4.1* [1, 9] are OpenAI's advanced VLMs. We implemented these models with the parameter max_tokens=1400.

| | Size | In-context | Modality | Format | KP | VC | IQ | ACC | **Avg** |
|---|---|---|---|---|---|---|---|---|---|
| Intern3.5-VL [47] | 8B | ✔ | 🖹 | → | - | - | - | 0.292 | 0.292 |
| Qwen3-VL [6] | 8B | ✔ | 🖹 | → | - | - | - | 0.250 | 0.250 |
| GPT-4o [1] | - | ✔ | 🖹 | → | - | - | - | 0.083 | 0.083 |
| GPT-4.1 [1] | - | ✔ | 🖹 | → | - | - | - | 0.167 | 0.167 |
| AnyEdit [58] | 1B | ◑ | 🖻 | ⧈ | 0.407 | 0.548 | 0.354 | - | 0.419 |
| UltraEdit(SD3) [62] | 2B | ◑ | 🖻 | ⧈ | 0.398 | 0.526 | 0.454 | - | 0.440 |
| VAREdit-8B [35] | 8B | ◑ | 🖻 | ⧈ | 0.581 | 0.698 | 0.498 | - | 0.580 |
| Step1X-Edit v1.1 [32] | 12B | ◑ | 🖻 | ⧈ | 0.617 | 0.941 | 0.426 | - | 0.625 |
| Step1X-Edit v1.2 [32] | 12B | ◑ | 🖻 | ⧈ | 0.567 | 0.681 | 0.476 | - | 0.562 |
| FLUX.1 Kontext [27] | 12B | ◑ | 🖻 | ⧈ | 0.578 | 0.907 | 0.465 | - | 0.610 |
| Qwen-Image-Edit [50] | 20B | ◑ | 🖻 | ⧈ | 0.667 | 0.802 | 0.446 | - | 0.628 |
| OminiGen [52] | 4B | ◑ | 🖻 | → | 0.167 | 0.106 | 0.241 | - | 0.177 |
| OminiGen2 [51] | 7B | ◑ | ✱ | → | 0.502 | 0.543 | 0.504 | - | 0.511 |
| Ovis-U1 [44] | 3B | ◑ | ✱ | ⧈ | 0.470 | 0.526 | 0.393 | 0.125 | 0.357 |
| UniPic [45] | 1.5B | ◑ | 🖻 | ⧈ | 0.341 | 0.296 | 0.287 | - | 0.316 |
| UniPic2-SD3.5M [48] | 2B | ◑ | 🖻 | ⧈ | 0.517 | 0.583 | 0.407 | - | 0.497 |
| UniPic2-Metaquery [48] | 9B | ◑ | 🖻 | ⧈ | 0.456 | 0.457 | 0.415 | - | 0.444 |
| NextStep-1-Large [43] | 15B | ◑ | 🖻 | ⧈ | 0.356 | 0.265 | 0.259 | - | 0.309 |
| Seedream 4.0 [41] | - | ◑ | 🖻 | → | 0.652 | 0.689 | 0.572 | - | 0.635 |
| Seedream 4.0 [41] | - | ✔ | 🖻 | → | 0.609 | 0.672 | 0.533 | - | 0.599 |
| Nano Banana [14] | - | ◑ | 🖻 | → | 0.680 | 0.790 | 0.560 | - | 0.666 |
| Nano Banana [14] | - | ✔ | 🖻 | → | 0.604 | 0.737 | 0.546 | - | 0.613 |
| Bagel [15] | 14B | ◑ | ✱ | → | 0.506 | 0.635 | 0.431 | 0.042 | 0.365 |
| Bagel-Zebra [30] | 14B | ◑ | ✱ | → | 0.500 | 0.624 | 0.480 | 0.125 | 0.396 |
| **+ WEAVE-100k** | 14B | ◑ | ✱ | → | 0.503 | 0.754 | 0.430 | - | 0.531 |

Table 6. **Main results on WEAVEBench 🕶Game Part.** ✔ and ◑ denote full and partial in-context history, respectively. 🖻, 🖹, and ✱ indicate image-only, text-only, and combined evaluations, respectively. → and ⧈ represent sequential and concatenated image inputs, respectively.

## C. More Related Works

**Interleaved Reasoning.** Large-scale corpora with interleaved text and images have become essential for pretraining VLMs with reasoning capabilities [2, 10, 12, 15, 19, 42, 55, 64]. Inspired by human cognition, where visual counterfactuals facilitate reasoning [39], recent work has incorporated analogous interleaved reasoning mechanisms into UMMs by mapping visual inputs to symbolic representations (*e.g.,* images or bounding boxes) [29, 49]. [54] explored pure visual reasoning relying solely on visual representations without textual modalities. Zebra-CoT [28, 30] provides an interleaved vision-language reasoning trajectory dataset to enhance UMMs' comprehension performance. IRG [23] generates an initial image, then iteratively refines it through reflective reasoning about quality improvements. ROVER [31] investigates the reciprocal relationship between generation and comprehension capabilities. In contrast, Weave focuses on in-context interleaved multimodal comprehension and generation.

**Benchmarks for UMMs.** UMM capability assessment typically encompasses three dimensions: (i) *Text-to-Image*: evaluated using GenEval [20] and DPGBench [22], which employ image detection methods [8] to ensure policy-compliant generation, and WISE [36], which examines complex semantic understanding and world knowledge for T2I generation; (ii) *Vision Comprehension*: consistent with Vision-Language Model (VLM) evaluation protocols, using benchmarks including MME [7], MMBench [33], MMMU [60], MM-Vet [59], and MathVista [34]; (iii) *Image Editing*: assessed via GEdit-Bench [32] and ImgEdit [57], which challenge UMMs to maintain image identity preservation while demonstrating semantic understanding. Additionally, RISEBench and KRIS-Bench evaluate reasoning with world knowledge. These benchmarks assess generation and comprehension in isolation, whereas ROVER [31] pioneered reciprocal cross-modal reasoning for omnimodal generation, systematically evaluating intermediate processes. WEAVE represents the first benchmark to comprehensively evaluate interleaved multi-turn generation and understanding.

# D. Additional Examples for WEAVE

## D.1. Additional Examples for WEAVE-100k

In this appendix, we present a comprehensive collection of examples that illustrate the versatility and capabilities of our WEAVE-100k framework. Figure 10 and Figure 11 demonstrate complex editing operations that require significant reasoning capabilities. The first example showcases intricate manipulations that demand careful consideration of spatial relationships and semantic coherence, while the second example introduces human subjects into the composition.

For Multi-Image Fusion operations, we provide four illustrative examples in Figures 6–9. Figure 8 demonstrates the model's ability to preserve footwear details during fusion operations. Figure 7 exhibits dual-task face processing capabilities. Figure 6 highlights the precise execution of specific hairstyle requirements and depicts scenarios where headphones are both held by one subject and worn by another, showcasing the model's understanding of object interactions across multiple contexts.

The Recall capability is exemplified in Figure 12, Figure 13, and Figure 14. In the first example, the model successfully restores previously removed trousers to the subject. The second example demonstrates the model's ability to reference a full-body model from Image #2 to reconstruct the complete body and scene in Image #4, while implementing a horizontally symmetrical background transformation. The third example shows the targeted reinsertion of a single human subject into the composition.

Additionally, we present specialized examples for Chess Game manipulation in Figure 15 and Visual JigSaw processing in Figure 16, further demonstrating the framework's adaptability to structured visual reasoning tasks.

## D.2. More example for WEAVEBench

This section presents the details of the examples shown in Figure **??**. Figure 17 demonstrates astronomical concepts, while Figure 19 tests biological knowledge. Mathematical reasoning is evaluated in Figure 25, and physical principles are examined in Figure 28. The model's chemistry knowledge is assessed in Figure 20, and fusion-related concepts in Figure 23. Geographic reasoning is presented in Figure 24. The model's game understanding capabilities are tested through chess problems in Figure 21 and Minecraft scenarios in Figure 18. Optical principles are demonstrated in Figure 27. The model's memory and recall abilities are evaluated in Figure 29, while spatial reasoning is tested in Figure 30 and Figure 32. Finally, narrative comprehension is assessed in Figure 31, and image editing capabilities in Figure 22.

# E. Broader Impact

The broader impact of Weave carries both potential benefits and risks upon deployment and release. Some considerations are unique due to the multimodal nature of UMMs while others reflect challenges common to image creation environments. Below, we outline risks and mitigation strategies for its release.

**Hallucination.** Similar to other models [5, 15, 30], our approach extends and fine-tunes text-to-image generation models to obtain unified generation capabilities, which introduces potential hallucination issues [26, 63]. Analogous to existing methods, models trained on WEAVE-100k may produce outputs that deviate from user intentions or specified input conditions. This phenomenon raises significant concerns, particularly in commercial image applications where purchasing decisions rely on accurate visual representations, given that user requirements and expression modalities exhibit inherent variability.

**Biases.** Despite implementing human supervision and a multi-model ensemble pipeline to mitigate biases in our synthetically generated dataset, the inherent biases from the foundation models inevitably permeate our data collection process and subsequently propagate to our fine-tuned models. This propagation can yield biased retrieval results and inequitable representations across diverse cultural contexts. Multilingual processing introduces additional bias vectors through language alignment mechanisms, as demonstrated by [11, 18].

**Domain: Multi-Image Fusion** Sub-domain: seedream


Image #1


Image #2


Image #3


Image #4


Image #5

| No. | Role | Type | Chat Content |
|---|---|---|---|
| 1 | USER | TEXT | Edit the person in **Image #1** by changing the background to an urban street setting and giving them short, curly black hair styled in a heart-shaped pattern. |
| 2 | ASSISTANT | IMAGE | Image #2 |
| 3 | USER | TEXT | Generate a person dancing on a mountain landscape, wearing casual clothes and smiling broadly. |
| 4 | ASSISTANT | IMAGE | Image #3 |
| 5 | USER | TEXT | Generate wireless headphones with a matte black finish, showcased against a plain white studio background. |
| 6 | ASSISTANT | IMAGE | Image #4 |
| 7 | USER | TEXT | Merge the person from **Image #2** (running in urban street with short curly black hair) and the person from **Image #3** (dancing in mountains), placing them together in a vibrant city plaza with the headphones from **Image #4** incorporated—one person wearing the headphones and the other holding them. |
| 8 | ASSISTANT | IMAGE | Image #5 |

Figure 6. An example of multi-image fusion in WEAVE-100k.

**Domain: Multi-Image Fusion** Sub-domain: seedream


Image #1


Image #2


Image #3

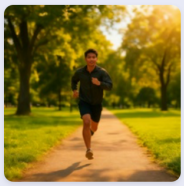


Image #4

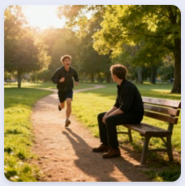


Image #5

| No. | Role | Type | Chat Content |
|---|---|---|---|
| 1 | USER | TEXT | Generate a person sitting on a bench in a photography studio. |
| 2 | ASSISTANT | IMAGE | Image #1 |
| 3 | USER | TEXT | Change the direction of this person's running, run towards him head-on, and make him look like a Southeast Asian. |
| 4 | ASSISTANT | IMAGE | Image #4 |
| 5 | USER | TEXT | Merge the person from **Image #1** (sitting on the bench) and the person from **Image #4** (running in a park with an excited expression) into a cohesive scene as **Image #4**, while keeping both figures' poses unchanged and having them look at each other. |
| 6 | ASSISTANT | IMAGE | Image #5 |

Figure 7. An example of multi-image fusion in WEAVE-100k.

**Domain: Multi-Image Fusion** Sub-domain: seedream


Image #1

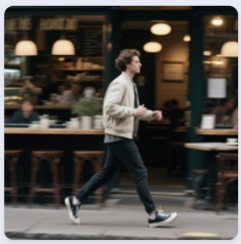


Image #2


Image #3

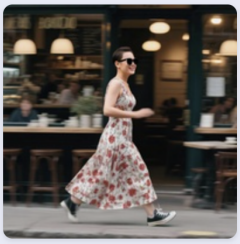


Image #4


Image #5


Image #6

| No. | Role | Type | Chat Content |
|---|---|---|---|
| 1 | USER | TEXT | Generate a person standing casually in a photography studio, wearing a simple t-shirt and jeans. |
| 2 | ASSISTANT | IMAGE | Image #1 |
| 3 | USER | TEXT | Edit the man from **Image #2** by changing the hairstyle to a short, neat cut and modifying the facial expression to a cheerful smile. |
| 4 | ASSISTANT | IMAGE | Image #3 |
| 5 | USER | TEXT | Edit the man in **Image #3** by turning him into a girl wearing a floral dress and adding a pair of sunglasses with black hair, keeping the man's shoes unchanged. |
| 6 | ASSISTANT | IMAGE | Image #4 |
| 7 | USER | TEXT | Edit the person from **Image #4** by adjusting the hair color to brown and modifying the facial expression to a surprised look. |
| 8 | ASSISTANT | IMAGE | Image #5 |
| 9 | USER | TEXT | Merge the man from **Image #1** and the girl from **Image #5** into a single scene, placing both people together inside a vibrant café, interacting naturally. Ensure their outfits and accessories are preserved. |
| 10 | ASSISTANT | IMAGE | Image #6 |

Figure 8. An example of multi-image fusion in WEAVE-100k.

🗂 Domain: Multi-Image Fusion  Sub-domain: gpt-image

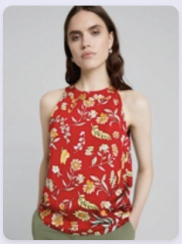


Image #1

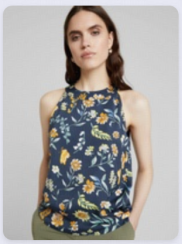


Image #2


Image #3


Image #4

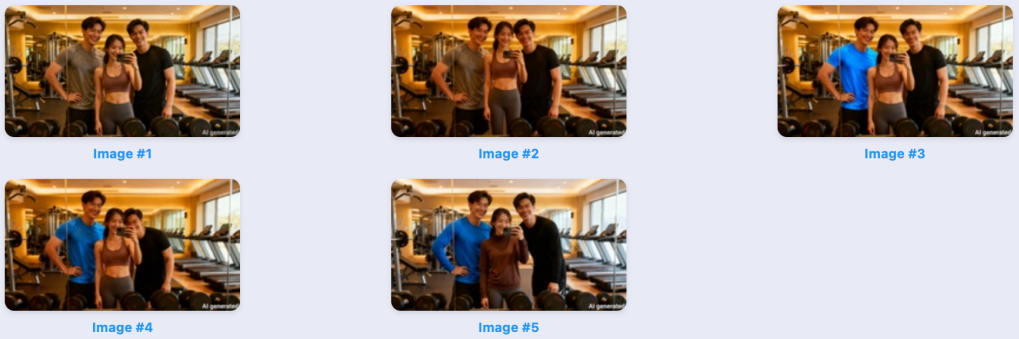

Image #5


Image #6

| No. | Role | Type | 💬 Chat Content |
|---|---|---|---|
| 1 | USER | TEXT | Edit **Image #1**: Change the color of the halter-style top to navy. |
| 2 | ASSISTANT | IMAGE | Image #2 |
| 3 | USER | TEXT | Edit **Image #3**: Remove the pizza cutter next to the open cardboard box. |
| 4 | ASSISTANT | IMAGE | Image #4 |
| 5 | USER | TEXT | Create an image that captures the ancient grandeur of the Colosseum beneath a vast, open blue sky. The weathered stone arches should convey a sense of history, while in the foreground, small groups of visitors wander the sunlit arena floor, their colorful clothing providing a lively contrast to the earthy tones of the ruins. |
| 6 | ASSISTANT | IMAGE | Image #5 |
| 7 | USER | TEXT | Position the young woman from **Image #2** standing near the center of the ancient Roman Colosseum scene from **Image #5**. She is holding a slice of the pizza from **Image #4** in her right hand, smiling as she takes a bite. The open pizza box is placed on a nearby stone bench, blending with the historical ambiance of the Colosseum. She is surrounded by the lively crowd of visitors, with her relaxed posture and attire adding a modern contrast to the ancient setting, while the partly cloudy sky above casts soft shadows around her. |
| 8 | ASSISTANT | IMAGE | Image #6 |

Figure 9. An example of multi-image fusion in WEAVE-100k.

**Domain: Edit** Sub-domain: real_human




| No. | Role | Type | Chat Content |
|---|---|---|---|
| 1 | USER | TEXT | Resize the person in the middle from **Image #1** to make her appear the same height as the others. |
| 2 | ASSISTANT | IMAGE | Image #2 |
| 3 | USER | TEXT | Edit the shirt color of the person on the left from **Image #1** to bright blue. |
| 4 | ASSISTANT | IMAGE | Image #3 |
| 5 | USER | TEXT | Move the person on the right from **Image #3** a little closer to the middle person. |
| 6 | ASSISTANT | IMAGE | Image #4 |
| 7 | USER | TEXT | Change the season in the gym from **Image #4** to winter, so everyone wears long sleeves. |
| 8 | ASSISTANT | IMAGE | Image #5 |

Figure 10. An example of edit in WEAVE-100k.

**Domain: Edit** Sub-domain: fantasy

Image #1

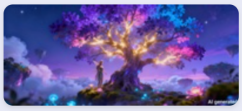


Image #2

Image #3

Image #4

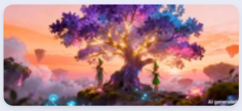


Image #5

| No. | Role | Type | Chat Content |
|---|---|---|---|
| 1 | USER | TEXT | Add a fantasy creature, like an elf, standing next to the magical tree from **Image #1**. |
| 2 | ASSISTANT | IMAGE | Image #2 |
| 3 | USER | TEXT | Duplicate the elf from **Image #2** so there are two elves, both standing near the tree. |
| 4 | ASSISTANT | IMAGE | Image #3 |
| 5 | USER | TEXT | Edit the tree from **Image #3** to isolate it on a blank background, removing everything else. |
| 6 | ASSISTANT | IMAGE | Image #4 |
| 7 | USER | TEXT | Change the time of day in **Image #3** to dawn, keeping everything else the same. |
| 8 | ASSISTANT | IMAGE | Image #5 |

Figure 11. An example of edit in WEAVE-100k.

🗒️ **Domain: Recall** Sub-domain: fashion

Image #1

Image #2

Image #3

Image #4

Image #5

Image #6

| No. | Role | Type | 💬 Chat Content |
|---|---|---|---|
| 1 | USER | TEXT | Make the brim of the hat in **Image #1** wider. |
| 2 | ASSISTANT | IMAGE | Image #2 |
| 3 | USER | TEXT | Change the background from **Image #2** to a cityscape with glowing lights, keeping the model, outfit, and fedora unchanged. |
| 4 | ASSISTANT | IMAGE | Image #3 |
| 5 | USER | TEXT | Change the image style from **Image #3** into a bold screen print illustration, keeping all elements the same. |
| 6 | ASSISTANT | IMAGE | Image #4 |
| 7 | USER | TEXT | Replace the cream-colored blazer from **Image #4** with a deep emerald green velvet blazer. |
| 8 | ASSISTANT | IMAGE | Image #5 |
| 9 | USER | TEXT | Merge the trousers from **Image #1** with the blazer from **Image #5**, maintaining the rooftop cityscape screen print look. |
| 10 | ASSISTANT | IMAGE | Image #6 |

Figure 12. An example of recall in WEAVE-100k.

**Domain: Recall** Sub-domain: fashion

Image #1

Image #2

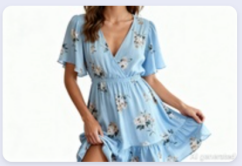


Image #3

Image #4

Image #5

| No. | Role | Type | Chat Content |
|---|---|---|---|
| 1 | USER | TEXT | Edit the dress from **Image #1** by changing its color to pastel blue, keep everything else the same. |
| 2 | ASSISTANT | IMAGE | Image #2 |
| 3 | USER | TEXT | Remove the background from **Image #2** and only show the upper body of the model. |
| 4 | ASSISTANT | IMAGE | Image #3 |
| 5 | USER | TEXT | Add a new background to **Image #3**: a modern fashion showroom with large windows and natural daylight. |
| 6 | ASSISTANT | IMAGE | Image #4 |
| 7 | USER | TEXT | Refer to the full-body portrait of the model in **Image #2** to complete the full-body and scene of the model in **Image #4**, with the background style horizontally mirrored. |
| 8 | ASSISTANT | IMAGE | Image #5 |

Figure 13. An example of recall in WEAVE-100k.

**Domain: Recall** Sub-domain: real_human

Image #1

Image #2


Image #3


Image #4

| No. | Role | Type | 💬 Chat Content |
|---|---|---|---|
| 1 | USER | TEXT | Change the style of the people from **Image #1** into stained glass while keeping the dinner setting unchanged. |
| 2 | ASSISTANT | IMAGE | Image #2 |
| 3 | USER | TEXT | Add the words 'Family Gathering' on top of the stained glass people from **Image #2**, keeping everything else the same. |
| 4 | ASSISTANT | IMAGE | Image #3 |
| 5 | USER | TEXT | Replace the boy in **Image #3** with the realistically styled boy from **Image #1**, while keeping all other areas unchanged from **Image #2**. |
| 6 | ASSISTANT | IMAGE | Image #4 |

Figure 14. An example of recall in WEAVE-100k.

**Domain: Chess Game** Sub-domain: None


Image #1


Image #2


Image #3


Image #4


Image #5

| No. | Role | Type | Chat Content |
|---|---|---|---|
| 1 | USER | TEXT | Edit **Image #1** by advancing White's h-pawn to h3. |
| 2 | ASSISTANT | IMAGE | Image #2 |
| 3 | USER | TEXT | Edit **Image #1** by moving White's bishop to d6, attacking the Black rook on e8. |
| 4 | ASSISTANT | IMAGE | Image #3 |
| 5 | USER | TEXT | Edit **Image #1** by advancing White's a-pawn to a3, attacking the Black knight on b4. |
| 6 | ASSISTANT | IMAGE | Image #4 |
| 7 | USER | TEXT | Edit **Image #1** by advancing White's g-pawn to g4. |
| 8 | ASSISTANT | IMAGE | Image #5 |
| 9 | USER | TEXT | Compare **Image #2**, **Image #3**, **Image #4**, and **Image #5**. Which edited position shows the best move for White, and why is it superior to the other options? Consider tactical opportunities, material advantage, and piece positioning in your analysis. |
| 10 | ASSISTANT | TEXT | Among the four edited positions, **Image #4** shows the best move. Advancing the a-pawn to a3 is strategically sound because it forces Black's active knight on b4 to retreat, disrupting Black's coordination on the queenside and improving White's position. The other moves either create weaknesses or fail to address key positional concerns. |

Figure 15. An example of Chess Game in WEAVE-100k.

**Domain: Visual Jigsaw** Sub-domain: None


Image #1


Image #2


Image #3


Image #4


Image #5

| No. | Role | Type | Chat Content |
|---|---|---|---|
| 1 | USER | TEXT | Use the image in Option A. to fill in the missing parts of **Image #1** . |
| 2 | ASSISTANT | IMAGE | Image #2 |
| 3 | USER | TEXT | Use the image in Option A. to fill in the missing parts of **Image #1** . |
| 4 | ASSISTANT | IMAGE | Image #3 |
| 5 | USER | TEXT | Use the image in Option A. to fill in the missing parts of **Image #1** . |
| 6 | ASSISTANT | IMAGE | Image #4 |
| 7 | USER | TEXT | Use the image in Option A. to fill in the missing parts of **Image #1** . |
| 8 | ASSISTANT | IMAGE | Image #5 |
| 9 | USER | TEXT | What's the right choice from **Image #1** to finish the puzzle? Answer in A B C or D |
| 10 | ASSISTANT | TEXT | A |

Figure 16. An example of Chess Game in WEAVE-100k.

## 🗒️ Domain: Astronomy

🖼️ Related Images (4 images)

Image #1 Image #2 Image #3 Image #4

💬 Conversation

| No. | Role | Type | Content | Key Points |
|---|---|---|---|---|
| 1 | USER | TEXT | We already know that **Image #1** is from the third day of the lunar month. What will it look like on the ninth day? | *None* |
| 2 | ASSISTANT | IMAGE | Image #2 | **GT #2**: As shown in **<image>**, the entire right half of the moon will be visible. |
| 3 | USER | TEXT | What will it look like six days after **Image #2**? | *None* |
| 4 | ASSISTANT | IMAGE | Image #3 | **GT #3**: As shown in **<image>**, it's a full moon |
| 5 | USER | TEXT | What will it look like six days after **Image #3**? | *None* |
| 6 | ASSISTANT | IMAGE | Image #4 | **GT #3**: As shown in **<image>**, the left half of the moon will be visible. |
| 7 | USER | TEXT | How many days will it take for the moon phase in **Image #4** to return to the moon phase in **Image #1**? | *None* |
| 8 | ASSISTANT | TEXT | 12 days. | *None* |

Figure 17. An example of astronomy domain testing the model's understanding of celestial objects and phenomena.

## 🗒️ Domain: Minecraft

🖼️ Related Images (3 images)

Image #1 Image #2 Image #3

💬 Conversation

| No. | Role | Type | Content | Key Points |
|---|---|---|---|---|
| 1 | USER | TEXT | Walk forward until the house and the cherry blossom tree fill the frame in **Image #1**—what will you see? | *None* |
| 2 | ASSISTANT | IMAGE | Image #2 | **GT #1**: A zoomed-in crop for **<image>**. |
| 3 | USER | TEXT | Based on **Image #1**, if you stand at the second-floor window on the left side of the house and look directly outward, what might you see? | *None* |
| 4 | ASSISTANT | IMAGE | Image #3 | **GT #3**: You can see the cherry blossom tree and the small house as shown in th **<image>**. |

Figure 18. An example of Minecraft domain testing the model's understanding of the game mechanics and environments.

## Domain: Biology

**Related Images (4 images)**


Image #1


Image #2


Image #3


Image #4

**Conversation**

| No. | Role | Type | Content | Key Points |
|---|---|---|---|---|
| 1 | USER | TEXT | Draw a cross-sectional diagram of the leaf of the plant shown in **Image #1**. | *None* |
| 2 | ASSISTANT | IMAGE | Image #2 | **GT #2**: Cross-sectional diagram of a holly leaf, similar to **<image>** |
| 3 | USER | TEXT | Predict what would happen to the leaf in **Image #1** if it were not watered for two weeks under hot, direct sunlight. | *None* |
| 4 | ASSISTANT | IMAGE | Image #3 | **GT #3**: As shown in **<image>**, the leaf would appear wilted and drooping. The leaflets curl inward, with edges becoming dry and brittle. |
| 5 | USER | TEXT | After the edits from the previous round, how would **Image #2** change accordingly? | *None* |
| 6 | ASSISTANT | IMAGE | Image #4 | **Image #2**: Comparing **<image>**, the following changes occurred: The mesophyll cells undergo plasmolysis, with both palisade and spongy mesophyll cells significantly shrinking and detaching from the cell walls, leaving clear gaps between the protoplasts and the rigid cell walls. The intercellular spaces become enlarged and irregular. |

Figure 19. An example of biology domain testing the model's understanding of biological structures and processes.

## Domain: Chemistry

**Related Images (5 images)**


Image #1


Image #2


Image #3


Image #4

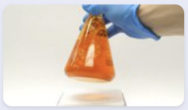


Image #5

**Conversation**

| No. | Role | Type | Content | Key Points |
|---|---|---|---|---|
| 1 | USER | TEXT | In the indigo carmine solution shown in **Image #1**, after adding glucose and sodium hydroxide solution, how will the color of the solution change? | *None* |
| 2 | ASSISTANT | IMAGE | Image #2 | **GT #2**: The color in the container is similar to the color in the container in the **<image>**. |
| 3 | USER | TEXT | After the color shown in **Image #2** appears, if the container is sealed and shaken, how will the color of the solution change? | *None* |
| 4 | ASSISTANT | IMAGE | Image #3 | **GT #3**: The color in the container is similar to the color in the container in the **<image>**. |
| 5 | USER | TEXT | After the color shown in **Image #3** appears, if left to stand for 1 minute, how will the color of the solution change? | *None* |
| 6 | ASSISTANT | IMAGE | Image #4 | **GT #4**: The color in the container is similar to the color in the container in the **<image>**. |
| 7 | USER | TEXT | What will happen if the conical flask in **Image #4** is shaken again (draw how it looks when shaken)? | *None* |
| 8 | ASSISTANT | IMAGE | Image #5 | **GT #5**: The color in the container is similar to the color in the container in the **<image>**. |
| 9 | USER | TEXT | What fundamental reactions are responsible for the appearance of red color in **Image #3** and yellow color in **Image #4**, respectively? (Please answer in terms of redox reactions.) | *None* |
| 10 | ASSISTANT | TEXT | Turns red (oxidation), turns back to yellow (reduction). | *None* |

Figure 20. An example of chemistry domain testing the model's understanding of chemical structures and reactions.

## 📋 Domain: Chess Game

🖼️ **Related Images (5 images)**


Image #1


Image #2


Image #3


Image #4


Image #5

💬 **Conversation**

| No. | Role | Type | Content | Key Points |
|---|---|---|---|---|
| 1 | USER | TEXT | Edit **Image #1** by having Black's rook capture the pawn on d7. | *None* |
| 2 | ASSISTANT | IMAGE | Image #2 | **GT #1**: Compared to **<image>**, Black's rook moves to d7 (The piece has moved to a new position, and the original piece has been removed.). |
| 3 | USER | TEXT | Edit **Image #1** by moving Black's queen to c7. | *None* |
| 4 | ASSISTANT | IMAGE | Image #3 | **GT #1**: Compared to **<image>**, Black's queen moves to c7 (The piece has moved to a new position, and the original piece has been removed.). |
| 5 | USER | TEXT | Edit **Image #1** by advancing Black's a-pawn to a6. | *None* |
| 6 | ASSISTANT | IMAGE | Image #4 | **GT #1**: Compared to **<image>**, Black's a-pawn moves to a6 (The piece has moved to a new position, and the original piece has been removed.). |
| 7 | USER | TEXT | Edit **Image #1** by moving Black's queen to a8. | *None* |
| 8 | ASSISTANT | IMAGE | Image #5 | **GT #1**: Compared to **<image>**, Black's queen moves to a8 (The piece has moved to a new position, and the original piece has been removed.). |
| 9 | USER | TEXT | Compare **Image #2**, **Image #3**, **Image #4**, and **Image #5**. Which edited position shows the best move for White, and why is it superior to the other options? Consider tactical opportunities, material advantage, and piece positioning in your analysis. | *None* |
| 10 | ASSISTANT | TEXT | Among the four edited positions, **Image #3** shows the best move. Moving the queen to c7 is the most active and promising option for Black. This move improves Black's piece coordination and prepares potential threats against the white king or central squares, offering counterplay and increasing Black's chances in the position. | *None* |

Figure 21. An example of chess game analysis testing the model's understanding of chess positions and strategies.

## 🗒️ Domain: Edit

🖼️ **Related Images (5 images)**


Image #1


Image #2


Image #3


Image #4


Image #5

💬 **Conversation**

| No. | Role | Type | Content | Key Points |
|---|---|---|---|---|
| 1 | USER | TEXT | What is the name of the most famous mountain in the country of the person in **Image #1**? | *None* |
| 2 | ASSISTANT | TEXT | Mount Fuji | *None* |
| 3 | USER | TEXT | Please change the background of **Image #1** to this mountain. | *None* |
| 4 | ASSISTANT | IMAGE | Image #2 | **Text**: A image of Mount Fuji |
| 5 | USER | TEXT | Remove the character from **Image #2** | *None* |
| 6 | ASSISTANT | IMAGE | Image #3 | **Image #2**: Remove the character from **<image>** |
| 7 | USER | TEXT | Generate the most famous tower in the capital of the country in **Image #1** | *None* |
| 8 | ASSISTANT | IMAGE | Image #4 | **Text**: A image of Tokyo Tower |
| 9 | USER | TEXT | Place the woman in **Image #1** in front of the tower of **Image #4** | *None* |
| 10 | ASSISTANT | IMAGE | Image #5 | **Image #4**: Background looks like **<image>**, **GT #1**: The woman looks like **<image>** |

Figure 22. An example of image editing task testing the model's ability to understand and suggest visual modifications.

## 🔲 Domain: Fusion

🖼️ **Related Images (7 images)**

Image #1

Image #2

Image #3

Image #4

Image #5

Image #6

Image #7

💬 **Conversation**

| No. | Role | Type | Content | Key Points |
|---|---|---|---|---|
| 1 | USER | TEXT | Generate a composite image where the person in **Image #1** is wearing the flower from **Image #2** on their head, and dressed in the clothes from **Image #3**. | *None* |
| 2 | ASSISTANT | IMAGE | Image #6 | **GT #1**: The person is same with **<image>**, **GT #2**: The flower is same with **<image>**, **GT #3**: The clothes are same with **<image>**, **Text**: flower on the girl's head |
| 3 | USER | TEXT | Generate a composite image where the person in **Image #6** is holding the bag on her right hand from **Image #5** and standing in the scene of **Image #4**. | *None* |
| 4 | ASSISTANT | IMAGE | Image #7 | **GT #6**: The person is same with **<image>**, **GT #5**: The bag on her right hand is same with **<image>**, **GT #4**: standing in the scene is same with **<image>** |

Figure 23. An example of fusion domain testing the model's understanding of nuclear fusion concepts and processes.

## 🗺️ Domain: Geography

🖼️ **Related Images (4 images)**

Image #1

Image #2

Image #3

Image #4

💬 **Conversation**

| No. | Role | Type | Content | Key Points |
|---|---|---|---|---|
| 1 | USER | TEXT | Draw a landscape according to the contour lines map in **Image #1** | *None* |
| 2 | ASSISTANT | IMAGE | Image #2 | **GT #1**: Draw a landscape according to the contour lines map in **<image>** |
| 3 | USER | TEXT | Add a castle at the highest point of your drawing in **Image #2** | *None* |
| 4 | ASSISTANT | IMAGE | Image #3 | **Image #2**: Add a castle at the highest point of your drawing in **<image>** |
| 5 | USER | TEXT | Edit the terrain map in **Image #1** after adding the castle | *None* |
| 6 | ASSISTANT | IMAGE | Image #4 | **GT #1**: Compared to **<image>**, there is an additional castle., **Image #3**: The castle's position on the hill should match that in the **<image>**. |

Figure 24. An example of geography domain testing the model's understanding of geographical features and locations.

## 📋 Domain: Mathematics

🖼️ **Related Images (3 images)**

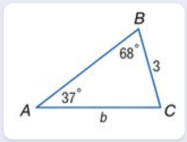


Image #1

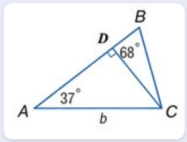


Image #2

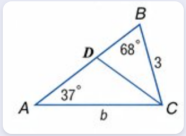


Image #3

💬 **Conversation**

| No. | Role | Type | Content | Key Points |
|---|---|---|---|---|
| 1 | USER | TEXT | What is the value of b in **Image #1**? Please give a brief solution process | *None* |
| 2 | ASSISTANT | TEXT | 1.541; Process: ∠C=180∘-37∘-68∘=75∘; b/sinB=a/sinA;sin68∘≈0.9272,sin37∘≈0.6018;b≈0.6018/0.9272≈1.541 | *None* |
| 3 | USER | TEXT | Draw a line through point A perpendicular to BC, with the foot of the perpendicular being D, and remove the number 3 from **Image #1**. | *None* |
| 4 | ASSISTANT | IMAGE | Image #2 | **GT #2**: Similar to **<image>**, there is a line through point A perpendicular to BC, with the foot of the perpendicular being D. |
| 5 | USER | TEXT | Based on the length of b calculated in the previous two questions, compute the numerical length of DC in **Image #2**. | *None* |
| 6 | ASSISTANT | TEXT | 0.927. Process: sin37∘*1.541 | *None* |
| 7 | USER | TEXT | In **Image #3**, I drew CD=AC, with point C on AB. What should be the measure of angle ACD? | *None* |
| 8 | ASSISTANT | TEXT | 37 degrees. | *None* |

Figure 25. An example of mathematics domain testing the model's problem-solving and reasoning abilities.

## Domain: Maze

**Related Images (4 images)**

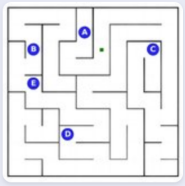


Image #1

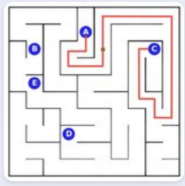


Image #2

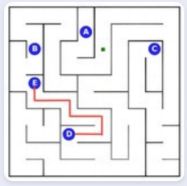


Image #3

Image #4

**Conversation**

| No. | Role | Type | Content | Key Points |
|---|---|---|---|---|
| 1 | USER | TEXT | Generate the solution path marked in red for the maze shown in **Image #1** from the A to the C. | *None* |
| 2 | ASSISTANT | IMAGE | Image #2 | **GT #1**: **<image>**have line from B to C is drawn in the diagram; it doesn't pass through walls but follows the maze. |
| 3 | USER | TEXT | Generate the solution path marked in red for the maze shown in **Image #1** from the D to the E. | *None* |
| 4 | ASSISTANT | IMAGE | Image #3 | **GT #1**: **<image>**Have a line from D to E is drawn in the diagram; it doesn't pass through walls but follows the maze. |
| 5 | USER | TEXT | Compare the red lines in **Image #2** and **Image #3**, and tell me which red line between the two points is longer. | *None* |
| 6 | ASSISTANT | TEXT | D to E | *None* |
| 7 | USER | TEXT | In the maze of **Image #1**, draw the shortest route between the two closest points that are connected, using a red line. | *None* |
| 8 | ASSISTANT | IMAGE | Image #4 | **GT #1**: **<image>**Have a line from B to E is drawn in the diagram; it doesn't pass through walls but follows the maze. |

Figure 26. An example of maze-solving task testing the model's pathfinding and spatial reasoning abilities.

## Domain: Optics

Related Images (3 images)

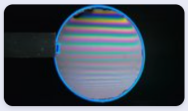


Image #1

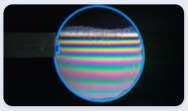


Image #2

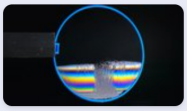


Image #3

Conversation

| No. | Role | Type | Content | Key Points |
|---|---|---|---|---|
| 1 | USER | TEXT | What happens when after a period of time, three-fourths of the thin film has disappeared for **Image #1**? | *None* |
| 2 | ASSISTANT | IMAGE | Image #2 | **GT #2**: Similar to **<image>**: One-fourth of the ring is empty, and interference phenomena still occur below. |
| 3 | USER | TEXT | How does the width of the interference fringes of the thin film change over time for **Image #1**? | *None* |
| 4 | ASSISTANT | TEXT | It becomes wider. | *None* |
| 5 | USER | TEXT | Does **Image #3** occur before or after **Image #2**? | *None* |
| 6 | ASSISTANT | TEXT | **Image #3** occurs after **Image #2**. | *None* |

Figure 27. An example of optics domain testing the model's understanding of optical principles and phenomena.

## Domain: Physics

Related Images (3 images)


Image #1

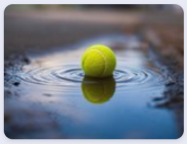


Image #2


Image #3

Conversation

| No. | Role | Type | Content | Key Points |
|---|---|---|---|---|
| 1 | USER | TEXT | Depict the scene two seconds after the tennis ball in **Image #1** falls into the puddle. | *None* |
| 2 | ASSISTANT | IMAGE | Image #2 | **GT #2**: Similar to **<image>**, the ball floats calmly on the water's surface, creating concentric ripples expanding outward. |
| 3 | USER | TEXT | Does **Image #3** occur before or after **Image #2**? | *None* |
| 4 | ASSISTANT | TEXT | **Image #3** occurs before **Image #2**. | *None* |

Figure 28. An example of physics domain testing the model's understanding of physical laws and principles.

## 📋 Domain: Recall

🖼️ **Related Images (5 images)**


Image #1


Image #2


Image #3


Image #4


Image #5

💬 **Conversation**

| No. | Role | Type | Content | Key Points |
|---|---|---|---|---|
| 1 | USER | TEXT | Change the background in **Image #1** to a plain black studio backdrop and remove the light. | *None* |
| 2 | ASSISTANT | IMAGE | Image #2 | **GT #1**: Change the background in **<image>** to a plain black studio backdrop and remove the light. |
| 3 | USER | TEXT | Edit the person from **Image #2** by replacing her shoes with white sneakers and add back the light from **Image #1**. | *None* |
| 4 | ASSISTANT | IMAGE | Image #3 | **GT #1**: There is a lamp in the upper left corner that is the same as the one in the **<image>**, but it is in the dark., **Image #2**: The person in **<image>** is exactly the same, except they've changed to white sneakers. |
| 5 | USER | TEXT | Change the image style in **Image #3** into glitch art. | *None* |
| 6 | ASSISTANT | IMAGE | Image #4 | **Image #3**: Change the image style in **<image>** into glitch art. |
| 7 | USER | TEXT | Merge the glitch art style walking person from **Image #4** with the original studio background from **Image #1**. | *None* |
| 8 | ASSISTANT | IMAGE | Image #5 | **GT #1**: Except for the people, everything else is the same as in **<image>**., **Image #4**: The person's style is the same as in the **<image>**. |

Figure 29. An example of recall task testing the model's memory and information retrieval capabilities.

## Domain: Spatial

**Related Images (6 images)**

Image #1

Image #2

Image #3

Image #4

Image #5

Image #6

**Conversation**

| No. | Role | Type | Content | Key Points |
|---|---|---|---|---|
| 1 | USER | TEXT | Provide the front view of **Image #1**. | *None* |
| 2 | ASSISTANT | IMAGE | Image #2 | **GT #1**: the front view of **<image>**, **GT #2**: Some components of the capsule room are similar to **<image>**. |
| 3 | USER | TEXT | Provide the left view of **Image #1**. | *None* |
| 4 | ASSISTANT | IMAGE | Image #3 | **GT #1**: the left view of **<image>**, **GT #3**: Some components of the capsule room are similar to **<image>**. |
| 5 | USER | TEXT | Provide the top view of **Image #1**. | *None* |
| 6 | ASSISTANT | IMAGE | Image #4 | **GT #1**: the top view of **<image>**, **GT #4**: Some components of the capsule room are similar to **<image>**. |
| 7 | USER | TEXT | Remove the green plants in **Image #1**. | *None* |
| 8 | ASSISTANT | IMAGE | Image #5 | **GT #1**: Remove the green plants in **<image>** |
| 9 | USER | TEXT | After the changes from **Image #1** to **Image #5** occur, how should **Image #4** be adjusted? | *None* |
| 10 | ASSISTANT | IMAGE | Image #6 | **GT #4**: Remove the green plants in **<image>** |

Figure 30. An example of spatial reasoning task testing the model's understanding of spatial relationships and transformations.

## 📋 Domain: Story

🖼️ **Related Images (5 images)**


Image #1


Image #2


Image #3


Image #4


Image #5

💬 **Conversation**

| No. | Role | Type | Content | Key Points |
|---|---|---|---|---|
| 1 | USER | TEXT | **Image #1** depicts Little Red Riding Hood (the little girl in a red cloak) setting off from home, carrying a basket. Based on this, create an image showing Little Red Riding Hood encountering the big bad wolf (with gray fur and pointed ears) on a forest path. | *None* |
| 2 | ASSISTANT | IMAGE | Image #2 | **GT #1**: Maintain Little Red Riding Hood's appearance as **<image>**., **Text**: The picture depicts Little Red Riding Hood encountering the big bad wolf (with gray fur and pointed ears) on a forest path. |
| 3 | USER | TEXT | Building on **Image #2**, create an image where the big bad wolf arrives at Grandma's house first and disguises himself as Grandma lying in bed. | *None* |
| 4 | ASSISTANT | IMAGE | Image #3 | **GT #2**: Maintain big bad wolf's appearance as **<image>**., **Text**: The picture depicts the big bad wolf arrives at Grandma's house first and disguises himself as Grandma lying in bed. |
| 5 | USER | TEXT | Based on **Image #2**, create an image where Little Red Riding Hood enters Grandma's house and notices something unusual about 'Grandma.' | *None* |
| 6 | ASSISTANT | IMAGE | Image #4 | **GT #2**: Maintain big bad wolf's appearance and Little Red Riding Hood's appearance as **<image>**., **Text**: Little Red Riding Hood enters Grandma's house and notices the wolf |
| 7 | USER | TEXT | Expanding on **Image #4**, create an image where the hunter (a man in a brown coat) aims his gun at the wolf disguised as the grandmother, while Little Red Riding Hood and the true grandmother stand beside the hunter. | *None* |
| 8 | ASSISTANT | IMAGE | Image #5 | **GT #2**: Maintain big bad wolf's appearance and Little Red Riding Hood's appearance as **<image>**., **GT #4**: Keep the background (except for the characters) and the image consistent as **<image>**., **Text**: The image depicts a hunter (a man in a brown coat) aiming his gun at the wolf disguised as the grandmother, with Little Red Riding Hood and the true grandmother standing beside the hunter. |

Figure 31. An example of story comprehension task testing the model's understanding of narratives and contexts.

## 🧩 Domain: Visual Jigsaw

🖼️ **Related Images (5 images)**

Image #1

Image #2

Image #3

Image #4

Image #5

💬 **Conversation**

| No. | Role | Type | Content | Key Points |
|---|---|---|---|---|
| 1 | USER | TEXT | Use the image in Option A. to fill in the missing parts of **Image #1**. | *None* |
| 2 | ASSISTANT | IMAGE | Image #2 | **GT #2**: Exactly the same as **<image>** |
| 3 | USER | TEXT | Use the image in Option B. to fill in the missing parts of **Image #1**. | *None* |
| 4 | ASSISTANT | IMAGE | Image #3 | **GT #3**: Exactly the same as **<image>** |
| 5 | USER | TEXT | Use the image in Option C. to fill in the missing parts of **Image #1**. | *None* |
| 6 | ASSISTANT | IMAGE | Image #4 | **GT #4**: Exactly the same as **<image>** |
| 7 | USER | TEXT | Use the image in Option D. to fill in the missing parts of **Image #1**. | *None* |
| 8 | ASSISTANT | IMAGE | Image #5 | **GT #5**: Exactly the same as **<image>** |
| 9 | USER | TEXT | Which option in **Image #1** fits to complete the puzzle? Answer in A B C or D | *None* |
| 10 | ASSISTANT | TEXT | A | *None* |

Figure 32. An example of visual jigsaw task testing the model's ability to understand and reconstruct visual patterns.

## References


[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] Anonymous. EditMGT: Unleashing potentials of masked generative transformers in image editing. In *Submitted to The Fourteenth International Conference on Learning Representations*, 2025. under review.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] Jinbin Bai, Tian Ye, Wei Chow, Enxin Song, Qing-Guo Chen, Xiangtai Li, Zhen Dong, Lei Zhu, and Shuicheng Yan. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. In *The Thirteenth International Conference on Learning Representations*, 2024.

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[7] Fu Chaoyou, Chen Peixian, Shen Yunhang, Qin Yulei, Zhang Mengdan, Lin Xu, Yang Jinrui, Zheng Xiawu, Li Ke, Sun Xing, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 3, 2023.

[8] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[9] Sixiang Chen, Jinbin Bai, Zhuoran Zhao, Tian Ye, Qingyu Shi, Donghao Zhou, Wenhao Chai, Xin Lin, Jianzong Wu, Chao Tang, et al. An empirical study of gpt-4o image generation capabilities. *arXiv preprint arXiv:2504.05979*, 2025.

[10] Xi Chen and Xiao Wang. Pali: Scaling language-image learning in 100+ languages. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[11] Wei Chow, Juncheng Li, Qifan Yu, Kaihang Pan, Hao Fei, Zhiqi Ge, Shuai Yang, Siliang Tang, Hanwang Zhang, and Qianru Sun. Unified generative and discriminative training for multi-modal large language models. *Advances in Neural Information Processing Systems*, 37:23155–23190, 2024.

[12] Wei Chow, Yuan Gao, Linfeng Li, Xian Wang, Qi Xu, Hang Song, Lingdong Kong, Ran Zhou, Yi Zeng, Yidong Cai, et al. Merit: Multilingual semantic retrieval with interleaved multi-condition query. *arXiv preprint arXiv:2506.03144*, 2025.

[13] Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025.

[14] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[15] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

[16] Taoran Fang, Wei Zhou, Yifei Sun, Kaiqiao Han, Lvbin Ma, and Yang Yang. Exploring correlations of self-supervised tasks for graphs. *arXiv preprint arXiv:2405.04245*, 2024.

[17] Taoran Fang, Tianhong Gao, Chunping Wang, Yihao Shang, Wei Chow, Lei Chen, and Yang Yang. Kaa: Kolmogorov-arnold attention for enhancing attentive graph neural networks. *arXiv preprint arXiv:2501.13456*, 2025.

[18] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50 (3):1097–1179, 2024.

[19] Zhiqi Ge, Juncheng Li, Qifan Yu, Wei Zhou, Siliang Tang, and Yueting Zhuang. Demon24: Acm mm24 demonstrative instruction following challenge. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11426–11428, 2024.

[20] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

[21] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15733–15744, 2025.

[22] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

[23] Wenxuan Huang, Shuang Chen, Zheyong Xie, Shaosheng Cao, Shixiang Tang, Yufan Shen, Qingyu Yin, Wenbo Hu, Xiaoman Wang, Yuntian Tang, et al. Interleaving reasoning for better text-to-image generation. *arXiv preprint arXiv:2509.06945*, 2025.

[24] Xuanwen Huang, Wei Chow, Yize Zhu, Yang Wang, Ziwei Chai, Chunping Wang, Lei Chen, and Yang Yang. Enhancing cross-domain link prediction via evolution process modeling. In *Proceedings of the ACM on Web Conference 2025*, pages 2158–2171, 2025.

[25] Shashank Mohan Jain. Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems*, pages 51–67. Springer, 2022.

[26] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, 2023.

[27] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025.

[28] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023.

[29] Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *arXiv preprint arXiv:2402.12058*, 2024.

[30] Ang Li, Charles Wang, Kaiyu Yue, Zikui Cai, Ollie Liu, Deqing Fu, Peng Guo, Wang Bill Zhu, Vatsal Sharan, Robin Jia, et al. Zebra-cot: A dataset for interleaved vision language reasoning. *arXiv preprint arXiv:2507.16746*, 2025.

[31] Yongyuan Liang, Wei Chow, Feng Li, Ziqiao Ma, Xiyao Wang, Jiageng Mao, Jiuhai Chen, Jiatao Gu, Yue Wang, and Furong Huang. Rover: Benchmarking reciprocal cross-modal reasoning for omnimodal generation. *arXiv preprint arXiv:2511.01163*, 2025.

[32] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

[33] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[34] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[35] Qingyang Mao, Qi Cai, Yehao Li, Yingwei Pan, Mingyue Cheng, Ting Yao, Qi Liu, and Tao Mei. Visual autoregressive modeling for instruction-guided image editing. *arXiv preprint arXiv:2508.15772*, 2025.

[36] Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

[37] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[39] Neal J Roese. Counterfactual thinking. *Psychological bulletin*, 121(1):133, 1997.

[40] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.

[41] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

[42] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[43] NextStep Team, Chunrui Han, Guopeng Li, Jingwei Wu, Quan Sun, Yan Cai, Yuang Peng, Zheng Ge, Deyu Zhou, Haomiao Tang, et al. Nextstep-1: Toward autoregressive image generation with continuous tokens at scale. *arXiv preprint arXiv:2508.10711*, 2025.

[44] Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, et al. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*, 2025.

[45] Peiyu Wang, Yi Peng, Yimeng Gan, Liang Hu, Tianyidan Xie, Xiaokun Wang, Yichen Wei, Chuanxin Tang, Bo Zhu, Changshi Li, et al. Skywork unipic: Unified autoregressive modeling for visual understanding and generation. *arXiv preprint arXiv:2508.03320*, 2025.

[46] Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. Seededit 3.0: Fast and high-quality generative image editing. *arXiv preprint arXiv:2506.05083*, 2025.

[47] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[48] Hongyang Wei, Baixin Xu, Hongbo Liu, Cyrus Wu, Jie Liu, Yi Peng, Peiyu Wang, Zexiang Liu, Jingwen He, Yidan Xietian, et al. Skywork unipic 2.0: Building kontext model with online rl for unified multimodal model. *arXiv preprint arXiv:2509.04548*, 2025.

[49] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[50] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025.

[51] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

[52] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13294–13304, 2025.

[53] Shilin Xu, Yanwei Li, Rui Yang, Tao Zhang, Yueyi Sun, Wei Chow, Linfeng Li, Hang Song, Qi Xu, Yunhai Tong, et al. Mixed-r1: Unified reward perspective for reasoning capability in multimodal large language models. *arXiv preprint arXiv:2505.24164*, 2025.

[54] Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. *arXiv preprint arXiv:2505.11409*, 2025.

[55] Yifan Yan, Shuai Yang, Xiuzhen Guo, Xiangguang Wang, Wei Chow, Yuanchao Shu, and Shibo He. mmexpert: Integrating large language models for comprehensive mmwave data synthesis and understanding. In *Proceedings of the Twenty-sixth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 1–10, 2025.

[56] Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025.

[57] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025.

[58] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26125–26135, 2025.

[59] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[60] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[61] Jieyu Zhang, Weikai Huang, Zixian Ma, Oscar Michel, Dong He, Tanmay Gupta, Wei-Chiu Ma, Ali Farhadi, Aniruddha Kembhavi, and Ranjay Krishna. Task me anything. In *Thirty-Eighth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[62] Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *arXiv preprint arXiv:2407.05282*, 2024.

[63] Siheng Zhao, Jiageng Mao, Wei Chow, Zeyu Shangguan, Tianheng Shi, Rong Xue, Yuxi Zheng, Yijia Weng, Yang You, Daniel Seita, et al. Robot learning from any images. In *Conference on Robot Learning*, pages 4226–4245. PMLR, 2025.

[64] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36: 8958–8974, 2023.