

# Wan-Weaver: Interleaved Multi-modal Generation via Decoupled Training

## Supplementary Material

### Contents

<b>A Societal Impact</b>	<b>1</b>
<b>B Details of Our Benchmark: WeaverBench</b>	<b>1</b>
B.1. Category Definition . . . . .	1
B.2. Evaluation Metrics . . . . .	2
B.3. Prompt for GPT-based Scoring . . . . .	4
B.4. Examples . . . . .	4
<b>C Additional Analysis</b>	<b>4</b>
<b>D More Quantitative Results</b>	<b>6</b>
<b>E More Qualitative Results</b>	<b>6</b>
E.1. Text-to-Image Generation . . . . .	6
E.2. Text-Image-to-Image Generation . . . . .	6
E.3. Interleaved Text-image Generation . . . . .	6
<b>F. Limitations</b>	<b>6</b>

More showcases and visual comparisons are available at <https://doubiiu.github.io/projects/WanWeaver>.

### A. Societal Impact

*Wan-Weaver*, an innovative interleaved text–image generation technology, has the potential to substantially advance human–AI interaction by enabling richer, more natural, and more expressive communication. A unified model capable of producing coherent mixed-modality content can support a wide range of beneficial applications, including educational content creation, visual tutoring systems, accessibility tools that jointly describe and illustrate concepts, and collaborative design workflows where users iteratively steer both text and images. These capabilities may also reduce the entry barrier for creative industries by assisting individuals without professional illustration or design experience in producing high-quality visual materials.

However, as with other large-scale generative models, interleaved generation systems carry inherent societal risks. Because they are trained on large, heterogeneous datasets, such models may inherit and amplify biases present in the underlying corpora, potentially leading to stereotypical or culturally insensitive generations across both textual and visual modalities. The ability to generate tightly interwoven text–image content also increases the risk of producing persuasive misinformation or synthetic narratives that

appear more credible due to coherent visual–textual alignment. Moreover, without appropriate safeguards, the model may generate harmful, inappropriate, or misleading visual content when prompted adversarially.

Our project is research-oriented and focuses on the scientific exploration of unified interleaved generation models. Nonetheless, we emphasize the importance of responsible deployment: future practical systems should incorporate robust content filters, watermarking, auditing mechanisms, and human-in-the-loop controls to mitigate potential misuse. Continued study of bias in multi-modal generation remains crucial as such technologies evolve.

### B. Details of Our Benchmark: WeaverBench

#### B.1. Category Definition

In this section, we present the definitions of 15 distinct categories within the WeaverBench, an evaluation framework designed for comprehensive performance assessment across various domains. Notably, a special ‘multi-language’ category also exists, which is integrated across 14 content-oriented categories. The precise definitions ensure that the benchmarks are relevant and reflective of real-world tasks. These categories are selected to encompass a wide range of familiar and frequently encountered challenges, ensuring an invaluable tool for evaluating models across diverse contexts.

**Encyclopedic Knowledge.** This category includes prompts related to articles or information found in encyclopedias, encompassing a broad spectrum of fundamental knowledge, concepts, and factual data. It emphasizes the model’s ability to accurately retrieve, synthesize, and present encyclopedic content.

**News Media.** Focused on the domain of news and media, this category involves tasks such as creating news reports, offering timely commentary, media analysis, and understanding of social dynamics. Models are tested on their capability to generate content that is informative and adheres to the stylistic norms of professional journalism.

**Travel Guide.** Involving destination recommendations, travel tips, itinerary planning, and cultural background information, this category assesses the ability to generate informative and engaging content for travel enthusiasts.

**Daily Life.** This category encompasses everyday tasks and challenges, including vlogs, household management, wellness advice, financial guidance, and interpersonal relationship tips. It measures a model’s proficiency in simulating real-world day-to-day scenarios and providing practical solutions.

Table 1. Distribution of data samples across the 15 categories in the WeaverBench benchmark. (The number before ‘+’ represents the amount for text-only input, and the number after ‘+’ represents the amount for image-text input.)

Category #Data Samples	Encyclopedic Knowledge 35 (16+19)	News Media 35 (16+19)	Travel Guide 36 (17+19)	Daily Life 35 (16+19)	Food Cooking 35 (16+19)
Category #Data Samples	Education Tutorial 36 (17+19)	Art Creation 35 (16+19)	Fashion Beauty 36 (17+19)	Media Entertainment 34 (16+18)	Product Marketing 35 (16+19)
Category #Data Samples	Academic Research 35 (17+18)	Document Guides 36 (17+19)	Social Community 29 (15+14)	Reasoning 32 (16+16)	Multi-language 28 (28+0)

**Food Cooking.** Tasks under this category involve recipes, cooking techniques, culinary exploration, and restaurant reviews. Content generation assessed here requires models to deliver accurate, creative, and user-friendly culinary insights.

**Education Tutorial.** This category involves developing educational materials such as online courses, study guides, lecture notes, and learning strategies. The ability to generate educational content that is pedagogically sound and engaging to learners is key.

**Art Creation.** Encompassing prompts related to photography, painting, drawing, handicrafts, art techniques, and exhibition of works, this category assesses creativity. Models are evaluated on their ability to produce content that reflects artistic insight and innovation.

**Fashion Beauty.** This category pertains to fashion trends, makeup techniques, skincare advice, and personal style. It evaluates the model’s ability to provide up-to-date fashion insights and beauty tips in line with industry trends.

**Media Entertainment.** Focused on film, television, variety show reviews, recommendations, and related information, this category assesses a model’s comprehension and ability to engage audiences in the realm of entertainment media.

**Product Marketing.** Involving product introductions, campaign strategies, sales techniques, industry trends, and advertising case studies, this category is designed to test the model’s capability to simulate market-oriented content generation and promotional strategies.

**Academic Research.** This category covers academic papers, research developments, scholarly discussions, and related resources. It assesses the model’s proficiency in generating well-founded and articulate academic content that meets the rigors of scholarly work.

**Document Guides.** Involves the creation of technical documents, user manuals, operating guides, and practical learning materials. The focus here is on clarity and utility in the generation of instructive content.

**Social Community.** This category is centered on user interaction, forum discussions, social media dynamics, and community activities. Models are evaluated on their capacity to generate engaging and constructive content that fosters community interactions.

**Reasoning.** In this category, prompts require complex, multi-step reasoning to generate answers. Tasks are designed to test a model’s cognitive capabilities in logical deduction, problem-solving, and articulation of coherent arguments.

**Multi-language.** This category primarily consists of English prompts that instruct the model to generate images containing text written in other languages. This task evaluates the model’s proficiency in handling multilingual contexts and integrating textual and visual data.

Through the detailed exploration of these categories, WeaverBench provides a rigorous and holistic framework for evaluating and improving the capabilities of current state-of-the-art models. We list the number of data samples in each category in Table 1.

## B.2. Evaluation Metrics

**Prompt Adherence** measures the extent to which the generated image–text output faithfully follows the user’s instruction and fulfills the specified task. This metric captures not only topical relevance, but also the correctness and usefulness of the solution provided with respect to all explicit and implicit requirements encoded in the prompt. High Prompt Adherence implies that the system addresses every key aspect of the request, avoids digressions into irrelevant content, and produces responses that are factually accurate and operationally helpful for the intended task. Conversely, low adherence is characterized by missing or ignoring parts of the instruction, addressing a different or incomplete task, or producing fundamentally incorrect or misleading content. In practice, this metric evaluates whether the output is directly responsive to the prompt, covers the necessary subcomponents, and provides an answer that a user could plausibly adopt to accomplish the stated goal.

**Narrative Coordination** assesses how effectively text and images are organized and sequenced to form a coherent, comprehensible multimodal narrative. It focuses on whether images appear at appropriate points in the response to support or clarify the corresponding textual explanation, and whether the overall progression across modalities facilitates step-by-step understanding. High Narrative Coordination is characterized by a deliberate alignment between modalities, in which visual elements are introduced at con-

Table 2. Quantitative comparison on WeaverBench. PA: Prompt Adherence, NC: Narrative Coordination, CC: Content Consistency, IC: Image Consistency, CP: Completeness, Acc: Accuracy of the number of the generated images.

Method	PA	NC	CC			IC		CP	Overall	Acc
			Intra-output	Input-to-image	Input-to-text	Entity	Style			
Orthus	2.47	1.88	1.42	1.30	2.35	1.19	1.82	1.91	1.89	3.28%
Anole	4.14	3.76	3.56	3.31	4.44	3.00	3.83	3.64	3.74	6.56%
Emu3.5	7.65	7.55	7.42	7.41	7.86	7.22	7.78	7.41	7.53	48.36%
Nano Banana	8.53	8.19	<b>8.31</b>	<b>8.57</b>	8.70	<b>8.09</b>	<b>8.67</b>	8.29	8.38	66.39%
Wan-Weaver (Ours)	<b>8.71</b>	<b>8.33</b>	<u>8.26</u>	<u>8.41</u>	<b>8.83</b>	<u>7.85</u>	<u>8.41</u>	<b>8.46</b>	<b>8.43</b>	<b>93.44%</b>

ceptually critical junctures, reinforce or concretize key textual descriptions, and avoid redundancy or disruption of the reading flow. In contrast, poor coordination may manifest as images that are placed arbitrarily, lack necessary accompanying text, appear out of order relative to the described steps, or introduce confusion rather than clarity. This metric thus captures the structural quality of the multimodal presentation, emphasizing pedagogically sound timing, logical ordering, and smooth integration of textual and visual information.

**Content Consistency** evaluates the degree of semantic and stylistic alignment between images, text, and the user’s original input. It consists of three complementary aspects: (1) Intra-output Coherence, which examines whether each generated image matches the adjacent textual description in content, mood, and style (e.g., an image labeled as “a watercolor sketch of a rainy street” should visually resemble a watercolor depiction of a rainy urban scene); (2) Input-to-Image Fidelity, which measures how well the generated images respect the explicit and implicit visual constraints specified in the user’s prompt, such as character attributes, scene configurations, or requested artistic styles; and (3) Input-to-Text Fidelity, which assesses whether the generated text adheres to the intended domain, tone, and factual expectations of the task (for example, employing precise, formal exposition for scientific explanations rather than casual or poetic language). High Image–Text Consistency indicates that images and text are mutually supportive, accurately reflect the user’s constraints, and maintain a coherent semantic and stylistic relationship throughout the output.

**Image Consistency** measures the stability and coherence of visual and semantic elements across multiple steps, scenes, or turns within a multi-image or multi-stage output. This metric comprises two main dimensions: (1) Entity Consistency, which evaluates whether recurring objects, characters, or scenes maintain identifiable and stable properties across all relevant images, including aspects such as color, shape, clothing, and spatial relationships. Substantial unexplained changes in these properties reduce the user’s ability to track entities and undermine narrative coherence. (2) Visual Style Uniformity, which assesses the extent to

which the images share a consistent visual style, such as medium (e.g., illustration vs. photograph), color palette, lighting conditions, contrast, and overall aesthetic. High Multi-step Consistency implies that the output resembles a cohesive visual sequence (e.g., frames from the same story or instructional series), allowing users to easily follow longitudinal processes or narratives without being distracted by unintended variations in entity appearance or visual style.

**Completeness** evaluates whether the multimodal output covers all essential components, steps, or stages implied or explicitly required by the user’s request, and whether the provided explanations and visual supports are sufficiently detailed. This metric concerns both breadth and depth: a complete response should address all major sub-tasks, intermediate steps, and boundary conditions necessary to achieve the requested objective, while also furnishing enough textual and visual information for the user to understand and potentially reproduce the described process. High Completeness entails that no critical stages are omitted, that important concepts are not merely mentioned but adequately explained, and that images are supplied where they substantially enhance understanding of complex or abstract content. Outputs are penalized when they skip key steps, provide only superficial treatment of central issues, or fail to include visual support for pivotal multimodal tasks. A highly complete response therefore anticipates the user’s informational needs, providing a comprehensive and well-structured answer that is both conceptually and practically useful.

**Accuracy.** To evaluate the model’s ability to follow precise instructions, we explicitly specify the required number of images to be produced in certain user prompts for interleaved text–image generation. The overall distribution of the requested image counts per prompt is shown in Fig. 4 (c) in the main paper. By comparing the number of generated images with the required targets, we compute an accuracy metric that reflects the model’s fidelity to user-specified structural constraints. This metric captures the model’s robustness in adhering to quantitative generation requirements.

Table 3. Comparison across single-modality generation tasks (understanding, image generation, and editing). †: Our in-house base model with thinking mode (enabled only for understanding). ‡: Methods are using LLM rewriter.

Model	Understanding		Image Generation		Image Editing	
	MMMU	MathVista	GenEval	DPG	ImgEdit	GEEdit-EN
<b>Visual Understanding Models</b>						
LLaVA-1.5 [10]	67.8	–	–	–	–	–
LLaVA-NeXT [11]	51.1	39.6	–	–	–	–
InternVL3-38B [25]	69.7	76.3	–	–	–	–
Ovis2-34B [13]	66.7	76.1	–	–	–	–
Qwen2.5-VL-32B†	<b>75.1</b>	<b>84.7</b>	–	–	–	–
<b>Text-to-Image Generation Models</b>						
FLUX.1-dev [8]	–	–	0.66	84.0	–	–
SDXL [15]	–	–	0.55	74.7	–	–
SD3-medium [7]	–	–	0.62	84.1	–	–
<b>Image Editing Models</b>						
Instruct-P2P [1]	–	–	–	–	1.88	3.68
MagicBrush [22]	–	–	–	–	1.90	1.86
AnyEdit [21]	–	–	–	–	2.45	3.21
Step1X-Edit [12]	–	–	–	–	3.06	<u>6.70</u>
IC-Edit [23]	–	–	–	–	3.05	4.84
<b>Unified Models</b>						
Janus-Pro [3]	41.0	42.5	0.80	84.19	–	–
Emu3 [17]	31.6	–	0.66‡	80.60	–	–
UniPic [16]	–	–	0.86	85.50	<u>3.49</u>	5.83
MetaQuery-XL [14]	58.6	–	0.80‡	82.05	–	–
Show-o2 [20]	48.9	–	0.76	<u>86.14</u>	–	–
OmniGen [19]	–	–	0.68	81.16	2.96	5.06
OmniGen2 [18]	53.1	–	0.86‡	83.57	3.44	6.42
BLIP3-o [2]	58.6	–	0.84‡	81.60	–	–
Bagel [6]	55.3	73.1	<u>0.88‡</u>	85.07	3.20	6.52
UniWorld-V1 [9]	58.6	–	0.84‡	81.38	3.26	4.85
Wan-Weaver (Ours)	<u>74.9</u>	<u>84.3</u>	<b>0.89</b>	<b>87.21</b>	<b>4.31</b>	<b>7.39</b>

### B.3. Prompt for GPT-based Scoring

The system prompt designed for GPT-based evaluator is presented in Table 5, 6, 7, and 8. The prompt instructs GPT to evaluate the generated interleaved text-image content based on the key evaluation metrics mentioned in Sec. B.2 (excluding ‘Accuracy’). For each metric, the GPT-based evaluator is required to rate the results on a scale from 0 to 10. Importantly, unlike prior benchmarks that rely on GPT-based evaluators [24], we provide explicit and fine-grained criteria for every integer score from 0 to 10. This design offers more precise guidance for the evaluator model and mitigates ambiguity during the scoring process.

### B.4. Examples

Illustrations of the representative examples of the 15 categories are provided in Fig. 6, 7, and 8, highlighting

the breadth and diversity of testing scenarios covered by WeaverBench. For each category, we present both text-only user prompts and mixed text–image prompts to reflect different forms of real-world inputs.

To further emulate the practical usage patterns, where users may issue instructions with varying levels of specificity, we also construct prompts of different lengths and detail levels. The overall distribution of prompt lengths is reported in Fig. 4 (b) in the main paper, with concrete examples provided in Fig. 9.

## C. Additional Analysis

**Impact of Fine-tuned Planner on Visualizer.** Since our framework adopts a decoupled training strategy, the planner and visualizer are optimized in separate stages. This raises a natural question: after we initialize the planner from a pre-

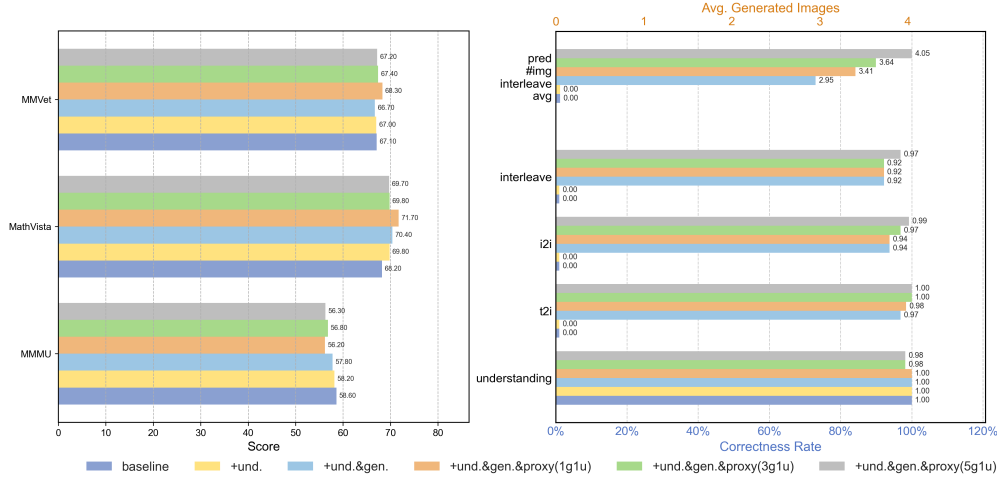


Figure 1. Impact of training on different types of data on our planner. (left) Performance on understanding metrics. (right) Token prediction statistics per task.

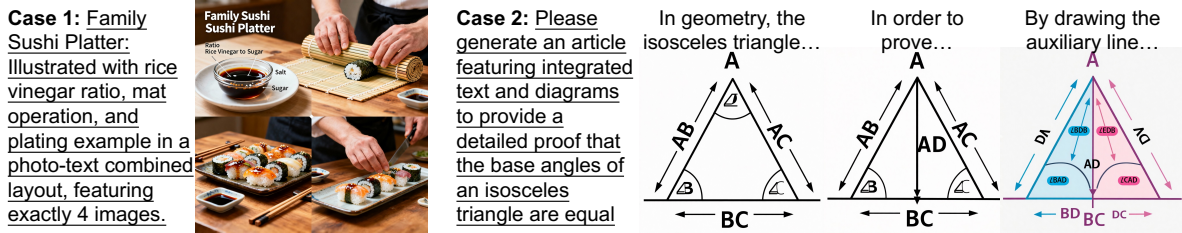


Figure 2. Visual examples of failure cases.

trained VLM and train the visualizer accordingly, will subsequent fine-tuning of the planner for planning capabilities introduce feature shifts that negatively affect image generation? To examine this, we fine-tune the visualizer again using the planner after its planning-oriented fine-tuning, and compare the resulting vision loss with that of the original visualizer trained with the pre-trained planner. As shown in Fig. 6 in the main paper, *i.e.*,  $V(T2I+SI2I+MI2I)$  vs.  $aP V(T2I+SI2I+MI2I)$ , the two loss curves are extremely close. This indicates that the planner’s features change only minimally during planning fine-tuning and do not adversely impact the visualizer’s generation quality.

**Feature Modeling in Planner.** Due to the limited space in the main paper, we provide a more detailed analysis here. To investigate how acquiring various levels of planning ability affects the underlying understanding capability, we perform planner tuning on different data compositions. The baseline corresponds to the VLM-initialized planner without any planning ability. As illustrated in Fig. 1, we construct several variants trained with different subsets of data: (1) **+und.&gen.&proxy**, which endows the planner with comprehensive planning skills across all tasks, including the generation of dense prompts for interleaved outputs; (2)

**+und.&gen.**, which preserves basic understanding while introducing generation-oriented planning; and (3) **+und.**, which preserves only the original understanding capability.

From Fig. 1 (left), we observe that the understanding performance remains highly stable across all configurations, indicating that introducing planning ability does not compromise the model’s core understanding competency. Beyond understanding accuracy, we further examine whether multi-task planning enables the planner to produce task-appropriate modality-specific patterns. Specifically, we compute the token-prediction accuracy of structural plan tokens. Given a user prompt, the planner must correctly produce the modality pattern, *e.g.*, no  $\langle BOI \rangle$  token for understanding tasks, exactly one for T2I/I2I tasks, and at least one for interleaved generation. As shown in Fig. 1 (right), we find that without any generation data (*gen.*), the model fails to emit image-generation signals. As the proportion of generation-oriented data increases, from zero *gen.* data, to 1g1u (1:1 ratio of generation to understanding), 3g1u, and 5g1u, the planner’s proficiency in generative planning improves substantially. For interleaved tasks in particular, the average number of predicted image starting tokens gradually increases as well. Balancing the overall planning reliability and understanding stability, we adopt the 5g1u ratio

as the final data composition for planner tuning.

## D. More Quantitative Results

**Single Modality Generation.** In addition to the quantitative results reported in Table 3 of the main paper, we present additional baseline performance in Table 3. The conclusions remain consistent: our method demonstrates superior performance compared with previous unified and specialized generation models.

**Detailed Results on WeaverBench.** We provide more fine-grained results of model performance on WeaverBench, serving as a complement to Table 2 in the main paper. Note that the CC and IC scores reported in the main paper correspond to the averages of their respective sub-components. In addition to the methods included previously, we further report the results of Anole [4] and Emu3.5 [5], representative models with general interleaved-generation capabilities, to offer a more comprehensive comparison. Note Emu3.5 technical report was just released on Oct.30, 2025, which is concurrent with our work. Moreover, we introduce an additional metric: Acc (accuracy), which measures whether the model generates the exact number of images specified by the user. Experimental results indicate that our approach follows user instructions much more reliably, achieving significantly higher accuracy than all competing baselines.

**Win Rate Comparison on OpenING.** In addition to the GPT-scoring results shown in Table 1, we also report the win rate comparison on OpenING [24], which is tabulated in Table 4. Our model consistently achieves top-tier performance across all GPT-based evaluation metrics, surpassing all competing systems by a clear margin. It delivers the highest FDT accuracy and remains robust under different tie-breaking strategies, indicating strong pairwise preference alignment and stable comparative quality. Under IntJudge evaluation, our model maintains competitive rankings and demonstrates notably strong performance in tie-adjusted settings, highlighting its reliability and consistency across different evaluators. Overall, the results show that our method achieves state-of-the-art preference quality and general robustness in interleaved text-image generation.

## E. More Qualitative Results

### E.1. Text-to-Image Generation

As illustrated in Fig. 3, our model supports a broad spectrum of text-to-image generation capabilities, including varied aspect ratios, multiple visual styles, accurate text rendering, comic and panel layouts, charts, flow diagrams, posters, and more.

### E.2. Text-Image-to-Image Generation

Fig. 4 presents a broad range of Wan-Weaver’s text-image-to-image generation capabilities, covering object addition, deletion, replacement, element extraction, pose manipulation, text modification, style and texture transfer, novel-view synthesis, object detection and segmentation, and more.

### E.3. Interleaved Text-image Generation

Additional interleaved text–image generation examples are shown in Fig. 5. The results demonstrate that our method is capable of generating text and images that are contextually coherent. Our method also exhibits notable reasoning abilities: although the user prompt does not directly specify ‘Shanghai’, but instead refers to it as the ‘magic city’ in China, the model is nevertheless able to produce accurate textual descriptions and corresponding images. The final image represents a more challenging case, as the model must infer which references to use and what content should be generated. Our results show that the model correctly incorporates information from all four preceding images and produces a coherent collection while preserving the detailed characteristics of the original visuals. The second case in Fig. 5 shows the dense prompt generated within the `<img></img>` tags. We also provide further qualitative comparisons in an accompanying local HTML file for detailed inspection.

## F. Limitations

Despite the strong performance of our unified model for interleaved multi-modal generation, several limitations remain.

First, the current model requires users to specify a fixed resolution and aspect ratio for all generated images, or to manually predefine these settings for each image in the sequence. Such constraints limit flexibility and are not aligned with real-world usage, where image composition, aspect ratio, and level of detail should ideally be determined automatically based on the evolving narrative and visual semantics. A more user-friendly design would involve adaptive resolution planning, allowing the model to infer the appropriate size and aspect ratio from the content itself, supporting richer visual–textual storytelling and more practical downstream applications.

Second, our planner–visualizer framework inherently follows a general sequential generation process, where all previously generated content, *i.e.*, both text and images, is fed back as reference information to guide the next-step generation. Although the visualizer’s per-image denoising can be parallelized across multiple GPUs, the overall workflow remains sequential, and the amount of conditioning grows with each newly generated image. As a result, gener-

Table 4. Quantitative comparison of model win rates evaluated by GPT-4o and IntJudge. FDT: Force Dividing Tie metric. w/o Tie: Non-Tie case. w/ Tie (0) and w/ Tie Tie(.5): Count a tie as 0 and 0.5 wins for a model in a battle, respectively.

Method	GPT Evaluation				IntJudge Evaluation			
	FDT	w/o Tie	w/ Tie (0)	w/ Tie (.5)	FDT	w/o Tie	w/ Tie (0)	w/ Tie (.5)
NExT-GPT	15.70%	15.60%	15.36%	16.13%	44.86%	32.61%	2.57%	48.63%
GILL	28.27%	28.29%	27.75%	28.71%	34.73%	6.52%	0.52%	46.51%
MiniGPT-5	30.20%	29.60%	28.84%	30.12%	31.97%	9.52%	1.03%	45.64%
Show-o	32.55%	32.10%	31.46%	32.46%	39.42%	13.16%	0.90%	47.47%
Orthus	34.63%	34.20%	33.45%	34.54%	43.05%	17.50%	1.19%	47.80%
SEED-LLAMA	37.86%	37.52%	36.50%	37.86%	44.05%	17.14%	1.02%	48.04%
Emu2	41.24%	40.87%	40.38%	40.98%	42.69%	12.28%	1.20%	46.30%
Emu3	43.64%	43.26%	42.44%	43.38%	49.66%	60.00%	4.64%	50.77%
Anole	49.37%	49.18%	48.29%	49.19%	58.38%	50.00%	2.88%	50.00%
VILA-U	49.37%	49.45%	48.83%	49.46%	57.01%	70.27%	4.68%	51.35%
SEED-X	49.38%	49.00%	48.13%	49.02%	53.65%	46.34%	3.39%	49.73%
Gemini+Flux	74.09%	75.00%	73.58%	74.53%	58.03%	75.00%	5.18%	51.73%
GPT-4o+DALL-E3	83.30%	83.51%	81.91%	82.87%	63.46%	86.89%	9.27%	53.93%
Nano Banana	90.99%	91.92%	90.46%	91.25%	67.61%	94.51%	15.22%	57.17%
Wan-Weaver (Ours)	<b>91.84%</b>	<b>92.46%</b>	<b>91.31%</b>	<b>91.93%</b>	62.94%	<u>92.86%</u>	<u>11.52%</u>	<u>55.32%</u>

ation efficiency progressively decreases, and GPU memory consumption increases over time. This sequential dependency poses a bottleneck for long interleaved sequences, suggesting the need for more efficient mechanisms to reduce accumulated conditioning while preserving contextual fidelity.

Third, although the planner–visualizer design demonstrates that understanding can facilitate better generation to some extent, the reverse effect is limited: enhanced generative ability does not meaningfully improve the model’s understanding capability. Achieving such bidirectional reinforcement likely requires more specialized training strategies, tailored multi-modal datasets, or unified representation mechanisms that tightly couple understanding and generation. We leave these as future works.

In addition, we show more failure cases in Fig. 2 regarding structural layout adherence and complex geometry reasoning: 1) Structural Collapse into Grid Layout. A primary limitation involves the model’s occasional failure to maintain the requested interleaved document structure. As shown in Case 1 of Fig. 2, when prompted to generate a photo-text combined layout with multiple images, the model tends to collapse these elements into a single-canvas grid layout rather than distributing them as separate entities within the text stream. This ‘grid-image bias’ suggests that the model sometimes conflates the concept of ‘multiple images’ with a ‘composite collage’, revealing a deficiency in its ability to orchestrate fine-grained document-level formatting and sequential image placement. 2) Logical Consistency and Geometric Reasoning. Case 2 high-

lights a limitation in handling high-level abstract reasoning and symbolic grounding. Although the model attempts to generate diagrams for a geometry proof, it fails to maintain logical consistency between the visual auxiliary lines and the corresponding textual labels. The rendered text within the images often suffers from hallucinations or symbolic confusion (e.g., misaligned vertex labels and inconsistent geometric topology).

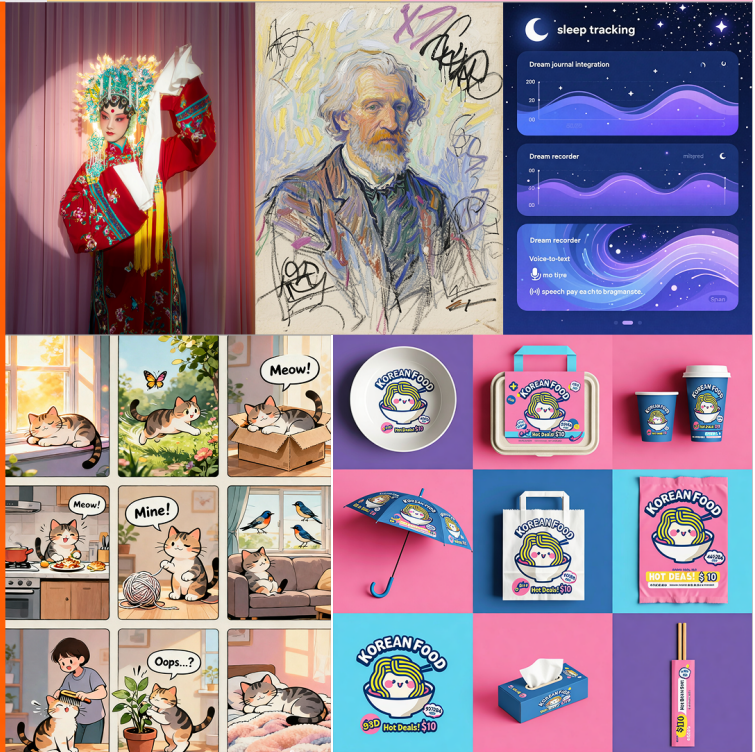
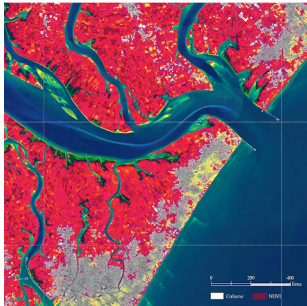
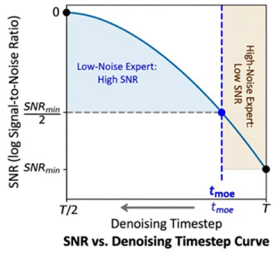


Figure 3. Showcase of Wan-Weaver in general text-to-image generation.

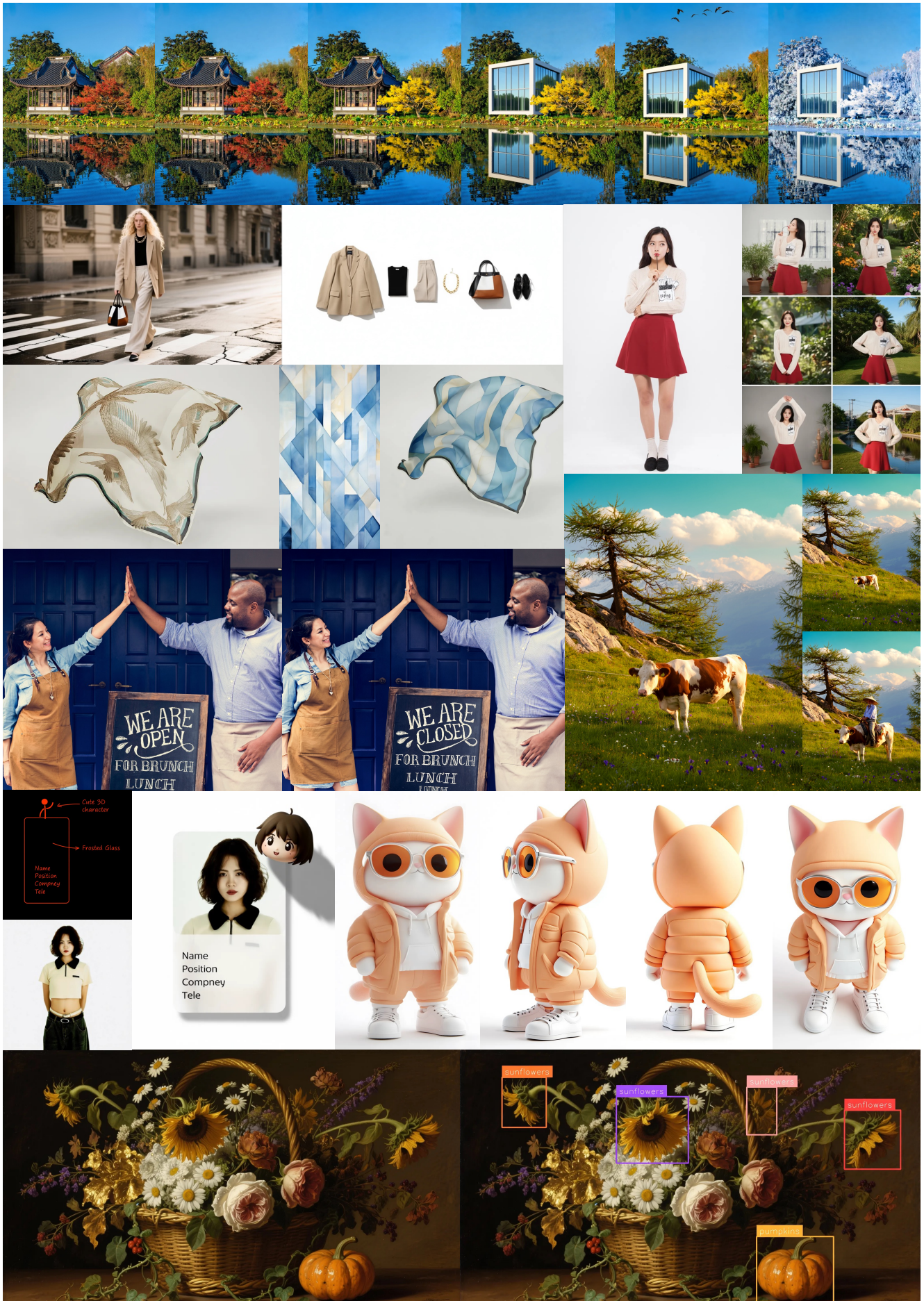


Figure 4. Showcase of Wan-Weaver in general image-to-image generation, including object addition, removal, replacement, element extraction, pose manipulation, text editing, style/texture transfer, novel view synthesis, object detection/segmentation, and more.

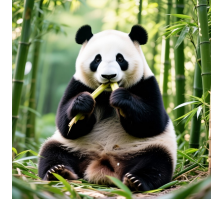
**Question 1:** Generate a Chinese travel photo-text guide featuring four tourist cities in an alternating text-and-image format. The cities are: the hometown of Lu Xun, the city known for panda culture, the capital, and the “magic city.” Each city should be accompanied by a relevant image and descriptive text. The final image should be a collage of the scenic views from all the previously mentioned cities.

China, this vast land, is home to countless travel destinations that inspire longing. From the weight of history to the vibrancy of modern life, every city has its own unique charm, waiting for us to explore and experience.

First, let us visit the hometown of Lu Xun—Shaoxing. This water town in the Jiangnan region is famous for its distinctive black-awning boats, ancient stone bridges, and rich yellow rice wine. Walking along the bluestone streets, one can almost hear the echoes of history and feel the atmosphere described in Lu Xun’s writings.



Next, we head to the city renowned for its “panda culture” — Chengdu. This is not only the hometown of giant pandas but also a paradise of cuisine and leisure. The lively atmosphere of Kuanzhai Alley, the vibrant charm of Jinli Ancient Street, and the adorable creatures at the Giant Panda Breeding Research Base all give this city its unique appeal.



Then we arrive in China’s capital — Beijing. As an ancient city with a history spanning thousands of years, Beijing is home to numerous UNESCO World Heritage Sites, such as the Forbidden City, the Great Wall, and the Temple of Heaven. It is not only a witness to history but also a crossroads of modern civilization, showcasing China’s profound cultural heritage and vibrant development.



Finally, we arrive in the magical city of Shanghai. This international metropolis features the Bund’s historic architectural complexes shining alongside the skyscrapers of Lujiazui, showcasing a perfect blend of Eastern and Western cultures. Whether it’s fashionable shopping or culinary exploration, Shanghai can fulfill all your expectations of a modern city.



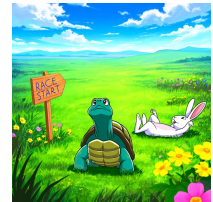
These four cities, each with its own distinctive charm, together form a brilliant chapter in Chinese tourism. Whether you are tracing the footsteps of history or feeling the pulse of modern life, they will all offer you an unforgettable travel experience.



**Question 2:** Create an article where narrative paragraphs and consistent anime-style illustrations are woven together seamlessly about “The Tortoise and the Hare”.

The story begins in a sun-drenched meadow where the Tortoise, slow but steady, announces a race against the Hare, who laughs at the idea.

<imagine>A wide, grassy meadow stretches under a bright blue sky dotted with fluffy white clouds. In the foreground, a plump, green-shelled tortoise stands upright, its wise eyes sparkling with determination. Nearby, a sleek white hare with long ears lounges lazily on a patch of clover, one paw raised in a playful wave. Wildflowers in shades of yellow, pink, and violet sway gently in the breeze, adding splashes of color across the scene. A wooden signpost at the edge of the field reads “Race Start” in cheerful hand-painted letters. In the distance, rolling hills fade into a soft haze, suggesting the vastness of the course ahead. The overall style is vibrant anime, with bold outlines and expressive character poses. </imagine>



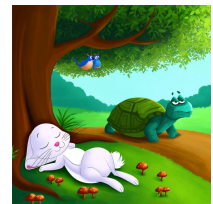
The race begins with a shout, and the Hare darts forward like a streak of lightning while the Tortoise plods along, each step measured and sure.

<imagine>A winding dirt path cuts through the meadow, leading toward a distant finish line banner fluttering in the wind. The hare is captured mid-leap, ears streaming backward, muscles taut, and eyes gleaming with confidence. Far behind, the tortoise moves steadily, its shell reflecting sunlight in gentle gradients of olive and teal. Butterflies flit above the path, their wings forming colorful blurs that echo the hare’s speed. A small crowd of woodland animals—squirrels, birds, and a curious fox—cheer from the sidelines, tiny flags waving in their paws. The sky is painted with warm afternoon light, casting long shadows that emphasize the contrast between haste and patience. The illustration balances dynamic motion lines around the hare with calm, steady shading around the tortoise, reinforcing their differing approaches. </imagine>



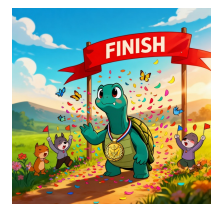
Midway through the course, the Hare, certain of victory, hops into a shady tree and drifts into a nap, while the Tortoise keeps moving, undeterred by fatigue.

<imagine>Beneath a sprawling oak tree whose branches form a natural canopy, the hare lies curled on a bed of soft moss, eyes closed in peaceful slumber. Sunbeams filter through the leaves, creating dappled patterns on its white fur and the forest floor. A few feet away, the tortoise trudges past, its face set in quiet concentration, oblivious to the sleeping rival. A pair of bluebirds perch on a low branch, whispering gossip about the race, their beaks slightly open in animated conversation. Fallen acorns and tiny mushrooms dot the ground, adding texture and detail to the woodland setting. In the background, the dirt path snakes onward, framed by ferns and wild violets. The color palette shifts to cooler greens and browns, conveying the calm of the forest and the tortoise’s steady progress. </imagine>



At last, the Tortoise reaches the finish line, greeted by cheers, while the Hare awakens to the shocking sight of the slow competitor already triumphant.

<imagine>A vibrant finish line banner reading “FINISH” in bold red letters stretches between two wooden poles, fluttering triumphantly in the breeze. The tortoise stands proudly beneath the banner, one front foot raised in a modest victory pose, eyes shining with hard-earned pride. Behind it, the hare bursts into view, panting heavily, ears drooping in disbelief as it realizes it has been outpaced. A jubilant crowd of forest animals surrounds the tortoise, tossing flower petals that swirl through the air like confetti. The sun dips toward the horizon, bathing the scene in golden light that highlights the tortoise’s shell with warm reflections. A small medal hangs around the tortoise’s neck, engraved with a laurel wreath, symbolizing perseverance. The illustration captures a moment of emotional contrast: the tortoise’s quiet joy versus the hare’s stunned surprise, all rendered in vivid anime detail. </imagine>



The moral is clear: confidence without diligence can falter, while steady effort can conquer even the swiftest rival. The meadow quiets as the sun sets, and the animals reflect on the lesson that patience and perseverance often outrun pride and haste.

Figure 5. Showcase of Wan-Weaver in interleaved text-image generation.

Social Community	Product Marketing
<p><b>Question 1:</b></p> <p>My cat master unlocked a new bizarre sleeping posture today, which is simply against the feline ergonomics! Please help me create a share with photo-comment interplay (featuring exactly 9 images in total), replying in English but ensuring all image text is in Chinese. Finally, you can add a call to encourage everyone to share the funniest and cutest photos of their pets, and see whose pet is the cutest!</p>	<p><b>Question 1:</b></p> <p>Draft a compelling, multi-section campaign piece for 'kids coding blocks', mixing step-by-step build instruction images with text, detailed views of the app screen with feature descriptions, and a vibrant showcase gallery with multiple creation examples and captions, arranged in picture-text interplay format.</p>
<p><b>Question 2:</b></p>  <p>Create multiple variations of this meme by adding different text for different social scenarios. Each version should include the finalized image and a detailed explanation of how it fits the context.</p>	<p><b>Question 2:</b></p>  <p>Based on the co-branded Jellycat × McDonald's plush I provide, please design an ad copy with an interleaved text-and-image layout that showcases the plush in various scenarios.</p>
Reasoning	News Media
<p><b>Question 1:</b></p> <p>Please explain with illustrations: why does a sphere have the smallest surface area under the same volume condition; include deformation diagrams and formula derivations, presented as synchronized text-picture segments. Use exactly six clearly labeled figures to support your explanation.</p>	<p><b>Question 1:</b></p> <p>Please generate a detailed article reporting on the 'river cleanup day.' I need it to be structured with engaging text in multiple paragraphs, visually interweaving 'before' and 'after' shoreline images to show the impact, along with a clear bar chart illustrating the total waste weight collected, in a balanced visual-text composition.</p>
<p><b>Question 2:</b></p> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;"> <p><b>Solve the quadratic equation:</b></p> <math display="block">x^2 + 5x + 3 = 0</math> </div> <p>Please solve the equation shown in the provided image using an interleaved text-and-image format.</p>	<p><b>Question 2:</b></p>  <p>Produce a dynamic news feature on Europe's green energy demands, where text and relevant images, like the one provided, seamlessly complement each other.</p>
Media Entertainment	Food Cooking
<p><b>Question 1:</b></p> <p>Please generate an illustrated analytical report on the "Importance of Soundtracks." The article should include multiple images to elaborate on a case study in detail, such as comparing "emotional curve graphs of two soundtracks for the same scene," accompanied by corresponding textual analysis in a graph-with-caption juxtaposition style.</p>	<p><b>Question 1:</b></p> <p>Please generate a well-illustrated guide article on "sous-vide steak". The article should be divided into detailed sections and include multiple images and charts for presentation. The content must include: a core temperature chart, a resting time curve, and a comparison of the pinkness when sliced at different levels of doneness, composed as a sequential text-photo hybrid.</p>
<p><b>Question 2:</b></p>  <p>Using the provided image of Stitch passionately singing, create a mixed text-and-visual review about the new live-action Lilo &amp; Stitch. Highlight early reactions, its faithful heartwarming remake qualities, and fresh creative updates, blending fun visuals with engaging commentary.</p>	<p><b>Question 2:</b></p>  <p>Using the given image, craft a durian pizza-themed piece where vivid descriptions dance with illustrative images, words and visuals woven together to make the reader taste and see each step.</p>

Figure 6. Examples of test samples from WeaverBench. For each category, we show one text-only user prompt and one mixed text-image user prompt (Part I).

<p><b>Fashion Beauty</b></p> <p><b>Question 1:</b></p> <p>Generate an image-text alternating guide for "Post-workout fresh makeup" and reply in English but all characters in the images are in Chinese. The guide should present 5 step-by-step sections, each with detailed text instructions and a corresponding visual illustration or photo. Additionally, integrate a visual "sweat durability bar" or infographic to rate the makeup's resilience.</p>	<p><b>Encyclopedic Knowledge</b></p> <p><b>Question 1:</b></p> <p>Write an illustrated introduction to "Ancient Silk Weaving Techniques": Each section respectively discusses "Reeling - Warp and Weft - Loom Mounting - Patterns," with exactly four images—one for each section—integrated directly beside the corresponding text for clarity.</p>
<p><b>Question 2:</b></p>  <p>Using the given image, craft a durian pizza-themed piece where vivid descriptions dance with illustrative images, words and visuals woven together to make the reader taste and see each step.</p>	<p><b>Question 2:</b></p>  <p>Create an encyclopedia-style entry about the vividly patterned <i>Scaphiophryne gottlebei</i> shown in the image, blending descriptive text with interleaved visuals of its habitat, anatomy, and range, ensuring an engaging mix of factual detail and illustrative elements.</p>
<p><b>Education Tutorial</b></p> <p><b>Question 1:</b></p> <p>Generate a research article about "Acid–base titration", clearly including exactly nine images with corresponding explanations presented in a visual–verbal integrated style.</p>	<p><b>Document Guides</b></p> <p><b>Question 1:</b></p> <p>Write a blended visuals-and-text guide for "cloud storage folder sharing".</p>
<p><b>Question 2:</b></p>  <p>Please use the uploaded photo of an outdoor mosquito-repellent spray as a reference to write a comprehensive feature for a wilderness survival and camping magazine. Include multiple extended-scene images that showcase different mosquito-prevention techniques and gear, and provide a detailed description for each image to create an in-depth, image-and-text article.</p>	<p><b>Question 2:</b></p>  <p>Could you whip up a super engaging, step-by-step tutorial guide on how to achieve this effect, maybe for Photoshop or a similar tool? I'm picturing a visually rich article with clear instructions interwoven with illustrative images, guiding users through the whole process. Make sure it highlights the transformation from a regular background to the kind of transparent checkerboard shown in the picture.</p>
<p><b>Daily Life</b></p> <p><b>Question 1:</b></p> <p>Please create a comprehensive, multi-section article on laundry symbols and stain removal. The article should clearly explain various laundry symbols, each accompanied by its corresponding icon. Following that, provide a step-by-step, visually illustrated process for effective stain removal, presented in a text with periodic image blocks layout.</p>	<p><b>Art Creation</b></p> <p><b>Question 1:</b></p> <p>For a 'Collage composition' lesson, could you create an illustrated guide? I need an explanation of 'focal point,' 'balance,' and 'rhythm,' with clear examples for each, specifically with a unique collage image to demonstrate each concept, along with accompanying text that explains its application, curated in an interwoven example-and-description format.</p>
<p><b>Question 2:</b></p>  <p>I love this vase! Create a decor guide based on its style. Write an article about bringing a 'serene seaside' vibe to a living space, featuring this vase and other compatible decor pieces.</p>	<p><b>Question 2:</b></p>  <p>Create a crayon-style storybook that weaves together text and images, centered on the character(s) from the provided image.</p>

Figure 7. Examples of test samples from WeaverBench. For each category, we show one text-only user prompt and one mixed text–image user prompt (Part II).

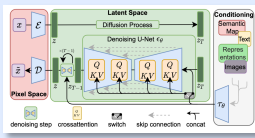

Academic Research	Travel Guide
<p><b>Question 1:</b></p> <p>Generate a detailed, multi-section analysis article on the relationship between 'Urban Greening and AQI'. Please integrate various visuals, including a scatterplot with a regression line to illustrate the core correlation, and a choropleth map to visualize geographical patterns. Each section should feature descriptive text closely linked to its associated graphics to present a comprehensive study.</p>	<p><b>Question 1:</b></p> <p>Could you put together a detailed 'Cairo essentials' piece, presented with multiple images? I'd like to see a segment explaining the Giza plateau plan, another showcasing the absolute museum must-sees, and a visually rich exploration of Cairo's street food, photo-story composition for each section.</p>
<p><b>Question 2:</b></p>  <p>Use the attached diagram showing latent space diffusion, pixel encoding, and cross-attention to craft a rich academic research-themed article blending explanatory visuals with analytical text, where imagery and narrative interlace naturally to guide understanding.</p>	<p><b>Question 2:</b></p>  <p>Using this mountain image as the backdrop, write a travel essay about hiking and exploring the Dolomites. The article should include descriptions of the mountain scenery from different perspectives and imagine several scenes from the adventure.</p>
Multi-language	
<p><b>Question 1:</b></p> <p>Please generate a complete article about "Chongqing Vertical Transport Day" in a format that alternates between images and text. The article should include: a geographical map showing a train passing through a building, accompanied by a textual explanation; a cross-sectional diagram depicting the elevation of stairways, along with relevant interpretation; and a screenshot of a queue ticket for a hot pot restaurant, supplemented by a description of the situation at that time. Provide a checklist at the end of the article, picture-narrative pattern. <i>Output English descriptions but all visible text inside the pictures should be Chinese.</i></p>	

Figure 8. Examples of test samples from WeaverBench. For each category, we show one text-only user prompt and one mixed text-image user prompt (Part III).

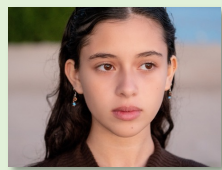


Short	Medium
 <p>Use this model to create a magazine article with interleaved images and text, showcasing different outfits and fashion elements.</p>	 <p>Analyze and report on the evolving debate around plastics, focusing on the idea that 'Plastics are greener than they seem.' Based on the provided image of a worker handling transparent plastic bottles in a production setting, create an engaging, mixed text-and-image news feature combining factual reporting, environmental impact analysis, industry perspectives, and critical commentary. Include data-driven comparisons to alternative materials, recent innovations in plastic recycling, and societal implications. Present your output in an informative, visually supported format where text paragraphs are interspersed with relevant images or infographics.</p>
Long	
 <p>Create a visually engaging, text-and-image integrated travel guide titled Guide to Rouen à Table! Gastronomy Festival, France. Use the provided image of a local charcuterie stall as a central visual reference to immerse readers in the atmosphere of this annual culinary celebration in Rouen. Describe the festival's highlights, including artisan food markets, regional specialties, cooking demonstrations, and opportunities to meet local producers. Offer destination recommendations for travelers, such as historic landmarks in Rouen, nearby day trips, and scenic spots that complement the gastronomic experience. Include practical travel tips: best times to visit, how to navigate the festival's crowded streets, language basics for interacting with vendors, and where to stay for convenient access. Integrate cultural background, explaining the heritage of Normandy cuisine, the role of charcuterie in French gastronomy, and the festival's importance in promoting local traditions. Maintain a warm, inviting tone that inspires culinary tourism while delivering actionable advice for itinerary planning.</p>	

Figure 9. Examples of short, medium, and long user prompts in WeaverBench.

###[System Prompt]

You are an expert in multimodal content evaluation. Your task is to assess interleaved image-text content generated by a model.

The model's input starts with "INPUT:" and can be a question or instruction requesting such content.

The model's output starts with "OUTPUT:" and may contain interleaved images and texts based on the input.

Image Representation:

- The image for step *i* will be enclosed by `<IMG_i>` and `</IMG_i>`.
- Images are numbered sequentially from 0 to *N* (including images in the input).
- You can directly accept image inputs for this evaluation.

—  
Evaluation Criteria – Main Categories and Subcategories

1. **Prompt Adherence:** Evaluates whether the combined image-text output directly addresses the user's request, fulfills the specified task, and provides a correct or useful solution. Penalize outputs that are off-topic, incomplete with respect to the core task, or fail to respond to key aspects of the instruction.

*Scoring Guidelines:*

- 0: *Prompt Adherence—Output is missing, irrelevant, or ignores the request entirely.*
- 1: *Prompt Adherence—Barely related; no real attempt to address the task.*
- 2: *Prompt Adherence—Addresses the wrong task or gives fundamentally incorrect content.*
- 3: *Prompt Adherence—Covers only a minor part; omits essential requirements.*
- 4: *Prompt Adherence—Partially correct but misses critical components.*
- 5: *Prompt Adherence—Meets basic intent but has noticeable omissions or errors.*
- 6: *Prompt Adherence—Mostly complete with minor flaws; solution is usable.*
- 7: *Prompt Adherence—Accurate, complete, and directly addresses the request.*
- 8: *Prompt Adherence—Fully fulfills task with relevant, useful details.*
- 9: *Prompt Adherence—Exceeds expectations by covering implicit needs precisely.*
- 10: *Prompt Adherence—Perfect execution with complete fidelity to the request.*

2. **Narrative Coordination:** Assesses whether images and text are placed at appropriate points in the response (e.g., image appears when a visual illustration is needed), and whether their sequencing supports clear understanding of the multi-step process or narrative.

*Scoring Guidelines:*

- 0: *Narrative Coordination—No structure; media and text are random or single-modality.*
- 1: *Narrative Coordination—No intentional sequencing; placement appears arbitrary.*
- 2: *Narrative Coordination—Sequence is chaotic; visuals disrupt rather than support understanding.*
- 3: *Narrative Coordination—Key steps lack needed visuals or textual context.*
- 4: *Narrative Coordination—Poor timing or placement creates confusion or gaps.*
- 5: *Narrative Coordination—Inconsistent flow; some sections poorly organized.*
- 6: *Narrative Coordination—Generally logical; small sequencing issues don't hinder understanding.*
- 7: *Narrative Coordination—Images appear where helpful; text and visuals align clearly.*
- 8: *Narrative Coordination—Thoughtful sequencing; media and text enhance each other seamlessly.*
- 9: *Narrative Coordination—Pedagogically sound structure that anticipates user comprehension.*
- 10: *Narrative Coordination—Elegant, flawless integration of text and images for maximum clarity.*

Table 5. The system prompts for obtaining detailed scores from GPT-based evaluators (Part I).

### 3. Content Consistency

- 1). Intra-output Coherence: Evaluates whether each generated image matches the semantic content, tone, and stylistic description of its accompanying or adjacent generated text (e.g., if the text says “a watercolor sketch of a rainy street,” the image should reflect that medium and mood).
- 2). Input-to-Image Fidelity: Assesses whether generated images respect explicit or implicit visual constraints from the input (e.g., character descriptions, setting, or artistic style mentioned in the prompt).
- 3). Input-to-Text Fidelity: Evaluates whether the generated text accurately reflects the intent, domain, tone, and factual expectations set by the input (e.g., a request for a scientific explanation should not yield poetic prose unless instructed).

#### *Scoring Guidelines:*

- 0: *Intra-output Coherence — Image directly contradicts its adjacent text; Input-to-Image Fidelity — Ignores all visual constraints from the input; Input-to-Text Fidelity — Text disregards input intent, domain, or task.*
- 1: *Intra-output Coherence — Image and text appear randomly paired; Input-to-Image Fidelity — Violates core visual details like character or setting; Input-to-Text Fidelity — Uses wrong tone, genre, or style entirely.*
- 2: *Intra-output Coherence — Frequent mismatches in content, mood, or style; Input-to-Image Fidelity — Misses key visual specifications from the prompt; Input-to-Text Fidelity — Misunderstands the purpose or factual expectations.*
- 3: *Intra-output Coherence — Rarely aligned; mostly inconsistent pairings; Input-to-Image Fidelity — Partially follows input but omits essential visual cues; Input-to-Text Fidelity — Occasionally relevant but largely off-topic or inaccurate.*
- 4: *Intra-output Coherence — Matches only basic or literal descriptions; Input-to-Image Fidelity — Honors at most one visual constraint (e.g., color or style); Input-to-Text Fidelity — Weakly aligned in isolated phrases only.*
- 5: *Intra-output Coherence — Moderate alignment in some image-text pairs; Input-to-Image Fidelity — Roughly respects general style or setting; Input-to-Text Fidelity — Generally on-topic but inconsistent in tone or accuracy.*
- 6: *Intra-output Coherence — Minor drifts in mood, medium, or detail; Input-to-Image Fidelity — Mostly adheres to explicit and implicit visual guidance; Input-to-Text Fidelity — Accurately reflects intent with small lapses.*
- 7: *Intra-output Coherence — Strong match in semantics, mood, and style; Input-to-Image Fidelity — Faithfully respects all key visual constraints; Input-to-Text Fidelity — Correctly captures domain, tone, and factual expectations.*
- 8: *Intra-output Coherence — Seamless harmony between image and its text; Input-to-Image Fidelity — Precisely fulfills all stated and implied visual requirements; Input-to-Text Fidelity — Text perfectly matches input instructions and context.*
- 9: *Intra-output Coherence — Near-perfect unity in meaning and aesthetic; Input-to-Image Fidelity — Fully honors nuanced visual and contextual cues; Input-to-Text Fidelity — Demonstrates deep, accurate understanding of user intent.*
- 10: *Intra-output Coherence — Image and text feel indistinguishably unified; Input-to-Image Fidelity — Flawless execution of every visual specification; Input-to-Text Fidelity — Perfectly embodies the input’s intent, tone, and detail as if user-authored.*

### 4. Image Consistency

- 1). Entity Consistency: Checks that recurring objects, characters, or scenes maintain consistent appearance, attributes, and spatial relationships across all relevant images (e.g., a red backpack in step 1 should remain red and recognizable in step 3).

Table 6. The system prompts for obtaining detailed scores from GPT-based evaluators (Part II).

2). **Visual Style Uniformity:** Evaluates consistency in artistic and photographic properties across all generated images, including media type (e.g., oil painting vs. photo), color palette, lighting, saturation, contrast, perspective, and overall aesthetic feel (e.g., consistently cartoonish or consistently photorealistic).

*Scoring Guidelines:*

- 0: *Entity Consistency* — *Recurring entities vanish or change beyond recognition; Visual Style Uniformity* — *Style shifts randomly with no consistency across images.*

- 1: *Entity Consistency* — *No effort to maintain identity; repeated items appear unrelated; Visual Style Uniformity* — *Each image uses a different medium, palette, or aesthetic.*

- 2: *Entity Consistency* — *Drastic changes make entities impossible to track; Visual Style Uniformity* — *Erratic jumps between styles break visual coherence completely.*

- 3: *Entity Consistency* — *Frequent unintended changes in color, shape, or pose; Visual Style Uniformity* — *Disruptive variations in lighting, perspective, or rendering type.*

- 4: *Entity Consistency* — *Sometimes recognizable but often altered inconsistently; Visual Style Uniformity* — *Mixed aesthetics (e.g., photorealistic and cartoon) or unstable color grading.*

- 5: *Entity Consistency* — *Core features partially retained but with noticeable drift; Visual Style Uniformity* — *General style trend present but with clear inconsistencies.*

- 6: *Entity Consistency* — *Most recurring entities remain identifiable across steps; Visual Style Uniformity* — *Style largely uniform with only minor, non-disruptive variations.*

- 7: *Entity Consistency* — *Appearance and spatial relationships stay consistent; Visual Style Uniformity* — *Cohesive look in medium, lighting, tone, and perspective throughout.*

- 8: *Entity Consistency* — *Attributes reliably preserved for easy tracking over time; Visual Style Uniformity* — *Fully unified media type, palette, lighting, and composition.*

- 9: *Entity Consistency* — *Precise visual identity maintained in every appearance; Visual Style Uniformity* — *Feels like frames from a single professional visual narrative.*

- 10: *Entity Consistency* — *Flawless preservation of shape, color, pose, and context across all steps; Visual Style Uniformity* — *All images indistinguishable in style—perfect visual harmony.*

5. **Completeness:** Verifies that all expected steps, components, or stages implied or explicitly requested in the input are present in the output, with sufficient explanatory depth in text and appropriate visual support where needed. Penalize omissions, underspecification, or glossing over critical subtasks.

*Scoring Guidelines:*

- 0: *Completeness* — *Ignores all explicit and implicit expectations; provides no steps, explanation, or visual support.*

- 1: *Completeness* — *Omits nearly all required content; lacks structure and meaningful text or images.*

- 2: *Completeness* — *Misses critical stages; explanations are extremely shallow and key visuals are absent.*

- 3: *Completeness* — *Skips major subtasks; lacks necessary detail and omits visuals for pivotal steps.*

- 4: *Completeness* — *Misses important components; explanations are insufficient and core parts lack proper visual support.*

- 5: *Completeness* — *Covers main points but glosses over several areas; critical subtasks receive only superficial treatment.*

- 6: *Completeness* — *Includes all essential steps with minor omissions; text is sufficient and visuals mostly appropriate.*

Table 7. The system prompts for obtaining detailed scores from GPT-based evaluators (Part III).

- 7: *Completeness* — All major steps present with adequate depth; visuals provided where needed for clarity.
- 8: *Completeness* — Rich explanations and high-quality, relevant visuals for all significant components.
- 9: *Completeness* — Includes every expected element and anticipates nuanced needs with insightful augmentation.
- 10: *Completeness* — Perfectly exhaustive; every requirement met with ideal detail and precisely matched visuals.

—  
Output JSON Structure:

Only four main categories are shown; each "Score" is the average of its subcategory scores.

```
{
  "scores": {
    "Prompt Adherence": {
      "Score": 0-10,
      "Justification": "Brief explanation of any issues identified"
    },
    "Narrative Coordination": {
      "Score": 0-10,
      "Justification": "Brief explanation of any issues identified"
    },
    "Image Text Consistency (Intra-output Coherence)": {
      "Score": 0-10,
      "Justification": "Brief explanation of any issues identified"
    },
    "Image Text Consistency (Input-to-Image Fidelity)": {
      "Score": 0-10,
      "Justification": "Brief explanation of any issues identified"
    },
    "Image Text Consistency (Input-to-Text Fidelity)": {
      "Score": 0-10,
      "Justification": "Brief explanation of any issues identified"
    },
    "Multi step Consistency (Entity Consistency)": {
      "Score": 0-10,
      "Justification": "Brief explanation of any issues identified"
    },
    "Multi step Consistency (Visual Style Uniformity)": {
      "Score": 0-10,
      "Justification": "Brief explanation of any issues identified"
    },
    "Completeness": {
      "Score": 0-10,
      "Justification": "Brief explanation of any issues identified"
    }
  }
}
```

- Be objective and thorough in your evaluation, providing clear justifications for your scores.
- Remember that you can accept image inputs directly, so you should analyze the images for each criterion.

Table 8. The system prompts for obtaining detailed scores from GPT-based evaluators (Part IV).

## References

- [1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 4
- [2] Jiu hai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025. 4
- [3] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 4
- [4] Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024. 6
- [5] Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025. 6
- [6] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 4
- [7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 4
- [8] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 4
- [9] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025. 4
- [10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 4
- [11] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 4
- [12] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 4
- [13] Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025. 4
- [14] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiu hai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025. 4
- [15] 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. 4
- [16] 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. 4
- [17] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyang Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 4
- [18] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyang Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 4
- [19] 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 *CVPR*, 2025. 4
- [20] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Showo2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025. 4
- [21] 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. 4
- [22] Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023. 4
- [23] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025. 4
- [24] Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo, Hao Zhang, Yuqi Lin, Yefei He, et al. Opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. In *CVPR*, 2025. 4, 6
- [25] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 4