Qwen-Image-Agent: Bridging the Context Gap in Real-World Image Generation

Zekai Zhang, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiaoyue Chen, Xiao Xu, Yan Shu, Yanran Zhang, Yixian Xu, Yuxiang Chen, Zhendong Wang, Zihao Liu, Zikai Zhou, Huishuai Zhang, Dongyan Zhao, Chenfei Wu*

Abstract

While text-to-image (T2I) models have achieved remarkable progress, they struggle with real-world requests that are often underspecified, implicit, or dependent on up-to-date knowledge. We identify this challenge as the Context Gap: the mismatch between the user context and the sufficient generation context for T2I models. To bridge this gap, we propose Qwen-Image-Agent, a unified agentic framework that integrates plan, reason, search, memory and feedback in a context-centric manner. Qwen-Image-Agent treats user input as partial context and progressively constructs the generation context through Context-Aware Planning and Context Grounding. Specifically, Context-Aware Planning identifies missing context and plans how it should be acquired and used, while Context Grounding gathers this context from reason, search, memory, and feedback. To evaluate agentic image generation, we further introduce Image Agent Bench (IA-Bench), a benchmark covering four core image agent capabilities: Plan, Reason, Search, and Memory. Experiments on IA-Bench, Mindbench and WISE-Verified show that Qwen-Image-Agent outperforms strong baselines and achieves state-of-the-art performance.

Figure 1: Qwen-Image-Agent examples, generated without providing visual references.

*Corresponding Author.

1 Introduction

Text-to-image (T2I) models have achieved impressive progress in generating high-quality images from natural language prompts (Labs, 2024; Stability AI, 2024b; Wu et al., 2025b). As these systems move into real-world applications such as marketing, product design, and slide creation, they are increasingly expected to solve practical visual tasks rather than merely render prompts.

Despite their generative ability, current T2I models remain limited on real-world tasks (He et al., 2026a). A key reason is a structural mismatch between training and deployment: models are optimized for fully specified prompts (Wu et al., 2025b), while real-world requests are often underspecified. In practice, successful generation may require inferring implicit user intent, retrieving up-to-date knowledge or visual references from web, and incorporating interaction history.

We refer to this mismatch as the **Context Gap**: the gap between the provided ***user context*** and the ***generation context*** required for T2I models. This gap motivates a paradigm shift from traditional ***direct image generation*** to ***agentic image generation***, where the system must identify missing context, acquire it, and use it effectively during generation. Recent work has explored components such as plan (Yao et al., 2026), reason (He et al., 2026a), search and tool use (Ye et al., 2026; Feng et al., 2026; He et al., 2026a), memory (He et al., 2026b), and self feedback (Jiang et al., 2026; Wang et al., 2025), but these efforts remain fragmented and do not provide a unified framework for context-centered generation.

To this end, we propose **Qwen-Image-Agent**, a unified agentic framework that integrates plan, reason, search, memory and feedback in a context-centric manner. Rather than treating user context as the final generation condition, our pipeline progressively constructs the full generation context through Context-Aware Planning and Context Grounding. Specifically, **Context-Aware Planning** operates at three levels: Information-level Planning identifies missing information and routes it to appropriate grounding strategies, Content-level Planning assembles grounded context into a detailed generation specification, and Generation-level Planning allocates context in multi-image and multi-turn scenarios. **Context Grounding** collects missing context from multiple sources, including reasoning for implicit intent inference, search for factual knowledge and visual references, memory for historical and personalized context, and feedback for iterative refinement. Overall, Qwen-Image-Agent is training-free, compatible with existing image generators, and supports both multi-image and multi-turn interaction.

Existing evaluations mainly emphasize rendering abilities (Ghosh et al., 2023; Hu et al., 2024) or isolated knowledge and reasoning abilities (Niu et al., 2025; Zhao et al., 2025), but fail to systematically assess the capabilities required for agentic image generation. To fill this gap, we introduce **Image-Agent-Bench (IA-Bench)**, a benchmark that evaluates four core agentic capabilities: Plan, Reason, Search, and Memory, over 17 real-world tasks, 730 test instances, and 1801 fine-grained binary checklist items. Each task is paired with a structured VLM-based evaluation protocol for reliable assessment.

Experiments on IA-Bench and prior benchmarks, including WISE-Verified (Niu et al., 2025) and Mind-Bench (He et al., 2026a), show that Qwen-Image-Agent substantially outperforms strong agentic baselines and achieves state-of-the-art results. Ablation studies further verify the complementary benefits of different grounded contexts. Our contributions are summarized as follows:

- We identify the **Context Gap**, i.e., the mismatch between user context and generation context as a fundamental challenge in real-world image generation. This provides a unified lens for understanding why current T2I systems fail in practical settings.
- We propose **Qwen-Image-Agent**, a unified and context-centric framework for agentic image generation that addresses the context gap through plan, reason, search, memory and feedback.
- We introduce **IA-Bench**, a benchmark for systematically evaluating agentic image generation along four capabilities: Plan, Reason, Search, and Memory.
- Experiments show that Qwen-Image-Agent substantially outperforms strong agentic baselines, and achieve state-of-the-art performance on IA-Bench, Mindbench and WISE-Verified.

2 Related Work

2.1 Agentic Image Generation

Recent work extends image generation and editing with agent capabilities such as planning, reasoning, memory, search, and self feedback. Planning-based methods decompose complex intents into intermediate steps (Yao et al., 2026); Reasoning-based methods handle implicit user intent for more intelligent generation and editing (He et al., 2026a); Search-based methods incorporate web search and image search to improve grounding in open-world scenarios (Feng et al., 2026; He et al., 2026a); Memory-based meth-

ods support long-horizon interactions through persistent memory (He et al., 2026b); and Feedback-based methods study test-time scaling for image generation (Jiang et al., 2026; Wang et al., 2025). However, from the perspective of generation context, existing methods remain fragmented in how they identify, acquire, and use the context required for real-world image generation. In contrast, Qwen-Image-Agent unifies plan, reason, memory, search, and feedback within a single framework, bridging the context gap in real-world image generation.

2.2 Benchmarks for Image Generation

Early image generation benchmarks mainly evaluate instruction following and text–image alignment, such as GenEval (Ghosh et al., 2023) for compositional attribute binding and DPGBench (Hu et al., 2024) for dense prompt following. More recent benchmarks target harder settings that are either knowledge-driven or reasoning-driven. Knowledge-driven benchmarks, such as WISE (Niu et al., 2025) and PhyBench (Meng et al., 2024), evaluate grounding in domain knowledge and physical commonsense. Reasoning-driven benchmarks, such as RISEBench (Zhao et al., 2025), test whether models can translate logical, causal, and spatio-temporal reasoning into visual outputs. Mind-Bench (He et al., 2026a) covers both aspects. However, existing benchmarks mainly evaluate partial agent abilities, especially reasoning or search, while largely overlooking planning and memory. To support holistic evaluation of agentic image generation, we introduce IA-Bench, which covers the full spectrum of agent capabilities with fine-grained, checklist-based evaluation.

3 Qwen-Image-Agent Framework

3.1 Formulation of Image Agents

We formalize image generation and edit as a conditional rendering problem. Given a user context $c_u = (P, I_{\text{ref}})$ with prompt P and optional reference images I_{ref}, **Direct image generation** renders output image y in a single forward pass, where p_{gen} is the image generator:

$$y \sim p_{\text{gen}}(\cdot \mid c_u). \tag{1}$$

In real-world scenarios, however, the provided user context is often incomplete for the desired visual task. We therefore distinguish **user context** c_u from the **generation context** c_g, which denotes the complete context needed for successful rendering. The earlier mentioned **context gap** is thus defined as the discrepancy between c_u and c_g.

Agentic image generation addresses this gap by treating p_{gen} as a renderer and introducing a context-construction process to resolve the context gap. At each step t, the agent maintains a state s_t, takes an action a_t, and receives an observation o_t, forming a trajectory

$$\tau = \{(s_t, a_t, o_t)\}_{t=1}^{T}. \tag{2}$$

The action space consists of basic operations to gather context, including `plan`, `reason`, `search`, `rewrite`, and `evaluate`. The state is defined as $s_t = (c_t, O_{t-1})$ where c_t is the current context under construction, and $O_{t-1} = \{o_1, \ldots, o_{t-1}\}$ is the set of accumulated intermediate results. Let $c(\tau)$ denote the final generation context induced by trajectory τ. The agentic generation process is then formulated as:

$$p_{\text{agent}}(y \mid c_u) = \sum_{\tau} p(\tau \mid c_u)\, p_{\text{gen}}(y \mid c_g = c(\tau)). \tag{3}$$

Under this formulation, the agent progressively builds the generation context along the trajectory before the final rendering step.

3.2 Overview of Qwen-Image-Agent

To bridge the context gap between user context and generation context required for image generators, we propose **Qwen-Image-Agent**, a unified agentic framework that integrates planning, reasoning, search, memory and feedback in a context-centric manner. As shown in Figure 2, it consists of two main modules: Context-Aware Planning and Context Grounding.

Context-Aware Planning identifies missing context, plans how to obtain it, determines how it should be used for generation and how to allocate it in multi-turn and multi-image scenarios.

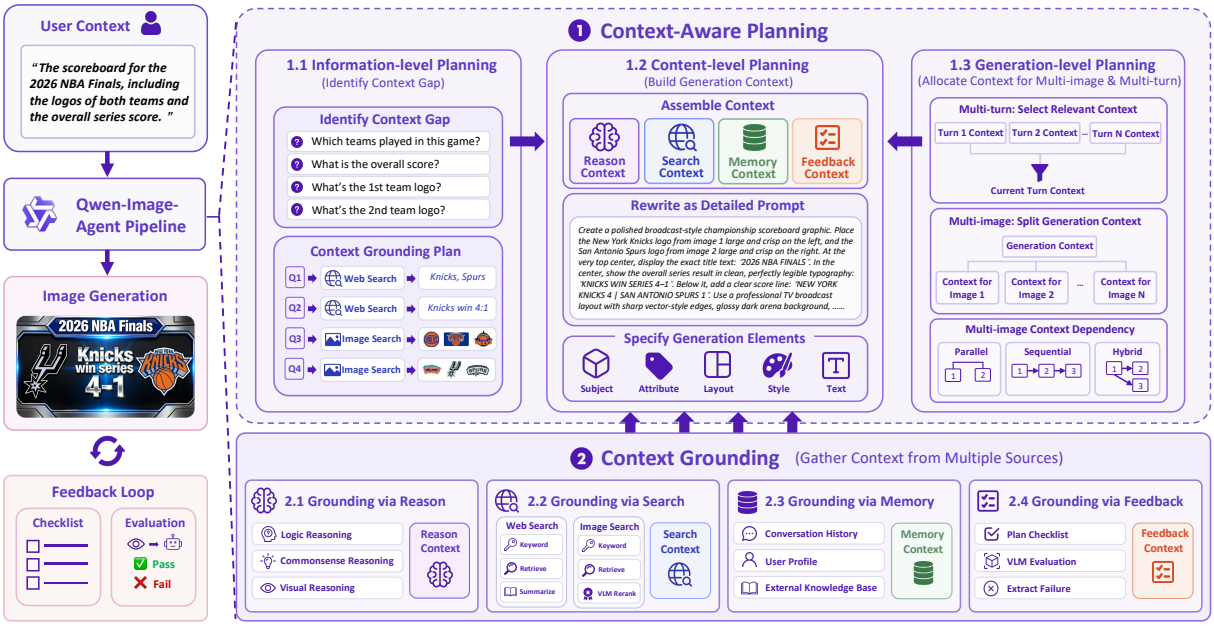

Figure 2: Overview of the Qwen-Image-Agent framework. Given a user context, the pipeline first identifies the context gap through information-level planning and gathers heterogeneous contexts. It then builds generation context through content-level planning. Qwen-Image-Agent further supports multi-turn and multi-image generation through generation-level planning.

Context Grounding gathers the missing information from multiple sources (including reason, search, memory and feedback) and organizes them in a context-centric manner.

Given a user context, the system first performs information-level planning to identify the context gap. It then grounds the missing information through reasoning, web search, and image search, producing reasoning context and search context. Together with memory context, these signals are fed into content-level planning, which builds a richer and more complete generation context for image synthesis. After an image is generated, the system evaluates the result through a feedback loop, and the newly obtained feedback context is incorporated back into content-level planning for iterative refinement. Finally, generation-level planning further extends the framework to support multi-turn and multi-image generation.

3.3 Context-Aware Planning

To systematically manage and utilize context throughout the generation process, we propose **Context-Aware Planning**. It operates at three levels: information-level, content-level, and generation-level.

Information-level planning identifies the context gap and plans how to resolve it. Given a user context, the system first raises explicit questions to characterize the missing information required for generation. Then, it routes each questions to a suitable context grounding strategy, including reasoning, web search and image search, as detailed in Section 3.4.

Content-level planning builds the generation context and plans the image content to be generated. Specifically, the system first assembles the context obtained during the context grounding stage, and then rewrites the user prompt into a detailed prompt that specifies key generation elements, including subject, attributes, layout, style, and textual elements.

Generation-level planning allocates generation context in multi-image and multi-turn scenarios. In multi-turn settings, excessively long contexts often lead to content drift or even generation collapse. To mitigate this issue, we select relevant information from previous turns while keeping the overall context length manageable. In multi-image settings, we distribute the generation context across individual images while accounting for multi-image context dependency, including parallel, sequential, and hybrid.

3.4 Context Grounding

To bridge the gap between user context and generation context, we propose **Context Grounding**, a unified module that collects context through reason, search, memory and feedback, and grounds generation with gathered context.

Grounding via Reason. User requests are often ambiguous, incomplete, or implicitly specified; therefore, generation needs to be grounded in additional context. Reasoning-based grounding addresses this issue by making implicit intents and requirements explicit. We consider three forms of reasoning: commonsense reasoning, logical reasoning, and visual reasoning. Specifically, for each question identified during Information-level Planning and assigned to reasoning, we employ a VLM to infer the corresponding answer. Together, these reasoning processes transform underspecified requests into concrete and explicit context items for downstream generation.

Grounding via Search. Some user requests depend on up-to-date factual information or IP-related visual references that cannot be inferred from the prompt alone. In such cases, we ground generation through search. For factual knowledge, we first extract search keywords from the user request, then perform web search and summarize the retrieved results into concise answers. For visual references, we retrieve candidate images from the web and employ a VLM to rank them, retaining the most relevant ones. Overall, search-based grounding enriches the request with external factual and visual context that cannot be obtained through reasoning alone.

Grounding via Memory. In multi-turn scenarios or long-horizon tasks, users may refer to knowledge or references mentioned in previous turns. In such cases, we ground generation with memory. Specifically, we incorporate the conversation history into the context and extract as well as update user profiles for long-horizon tasks. In addition, memory grounding extends to external memory sources, such as textual and visual knowledge bases. To support this, we implement a multimodal retriever that retrieves the most relevant textual and visual items from external memory and integrates them into the grounded context for generation.

Grounding via Feedback. Text-to-image models cannot directly inspect their own outputs, which often leads to discrepancies between the prompt and the generated image. In such cases, we ground generation through feedback. Specifically, after generation, we first plan a checklist of expected image attributes, and then employ a VLM to assess each generated result against this checklist. Items that fail the evaluation are converted into feedback context and combined with the previously grounded context to refine the prompt for the next round. Overall, feedback-based grounding closes the loop between generation and evaluation, enabling iterative correction toward better alignment with user intent.

4 IA-Bench

4.1 Motivation and Overview

Existing benchmarks for image generation mainly focus on rendering-oriented abilities, such as instruction following, visual fidelity, and aesthetic quality. However, real-world image generation often involves challenges beyond rendering alone: user requests may be underspecified, require external knowledge, demand multi-step decomposition, or depend on prior context. Addressing such requests requires models to infer implicit constraints, reason over intermediate decisions, retrieve relevant information, and maintain consistency across turns. These capabilities remain insufficiently studied in existing benchmarks, despite being particularly important for agentic image generation.

To address this gap, we introduce **Image Agent Bench (IA-Bench)**, a benchmark designed to evaluate the agentic capabilities involved in image generation. As illustrated in Figure 3, IA-Bench covers four core capabilities: **Plan**, **Reason**, **Search**, and **Memory**. The benchmark consists of *4 tasks, 17 subtasks, 730 instances and 1801 evaluation checklist items*. Together, they provide a structured evaluation framework for image generation systems across planning, reasoning, search, and memory dimensions.

Planning-Driven Tasks Planning-driven tasks evaluate whether a model can decompose a high-level goal into concrete visual arrangements and execute them in the final image. As illustrated in Figure 3, this category includes tasks such as **Composition**, **Enumeration**, and **Multi-Panel**. These tasks require the model to explicitly organize multiple objects, satisfy counting constraints, and place visual elements into structured layouts. For example, a composition task may ask the model to place a specified number of objects with different attributes into a coherent scene, while a multi-panel task may require generating a grid of images that jointly satisfy a higher-level instruction. Such tasks emphasize deliberate planning over purely local rendering quality.

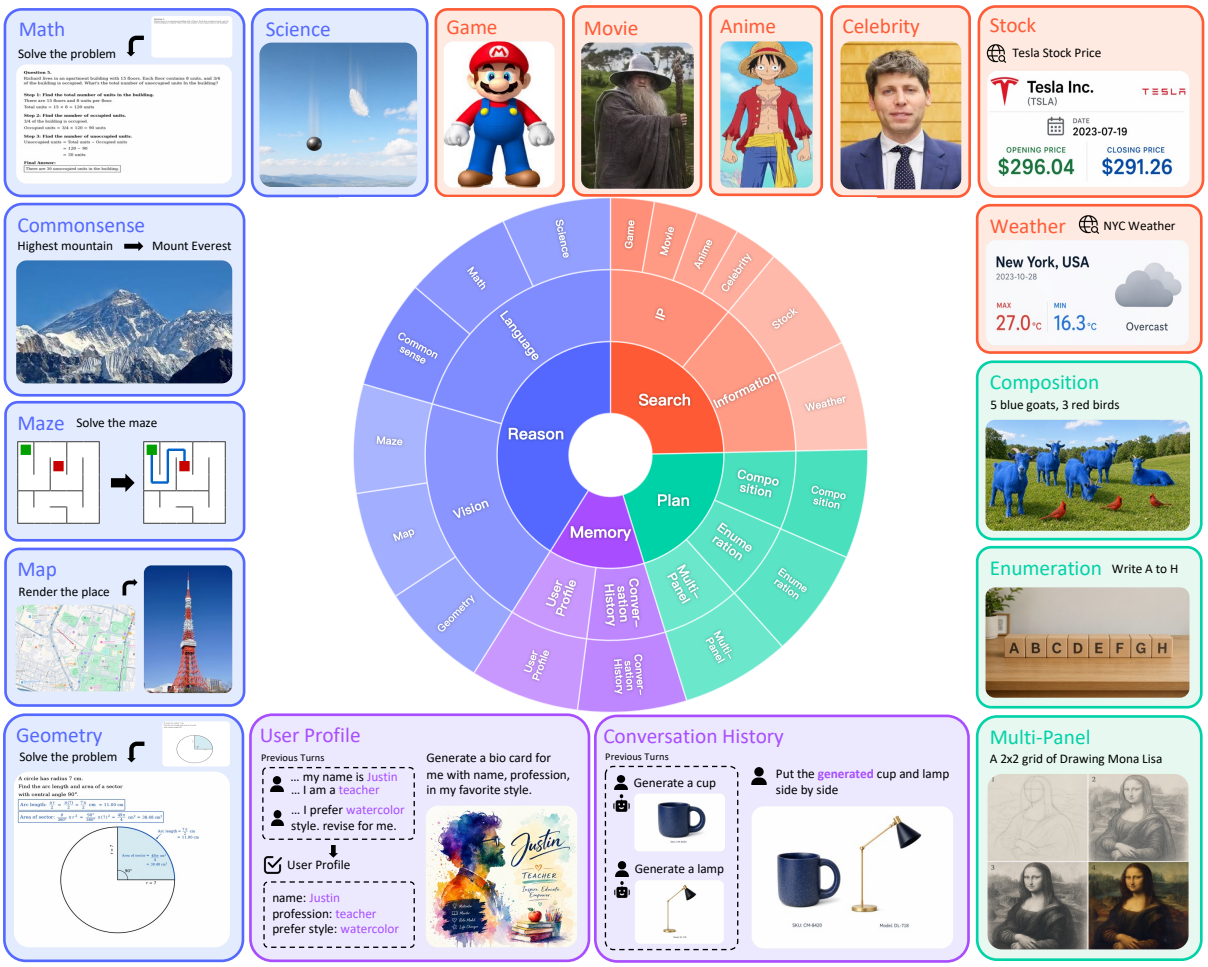

Figure 3: Overview of IA-Bench. IA-Bench covers 4 tasks, 17 subtasks, 730 instances and 1801 evaluation checklist items, providing a comprehensive evaluation of agentic image generation capabilities.

Reasoning-Driven Tasks Reasoning-driven tasks assess whether a model can infer latent constraints before generation and correctly ground the inferred result into the image. This category includes **Math**, **Science**, **Commonsense**, **Maze**, **Map**, and **Geometry**. These tasks involve three major types of reasoning: logical reasoning, commonsense reasoning, and visual reasoning. For example, a model may need to solve a math problem, infer the correct target from commonsense knowledge, or identify a valid path in a maze before rendering the final image. Unlike standard rendering tasks, success in this category depends on whether the model can first derive the correct intermediate conclusion and then faithfully express it in visual form.

Search-Driven Tasks Search-driven tasks assess whether a model can retrieve or ground external world knowledge that is not fully specified in the prompt. In IA-Bench, this category covers two major sources of knowledge: **IP**-related entities and **Information**. The **IP** branch includes tasks such as **Game**, **Movie**, **Anime**, and **Celebrity**, where the model must identify or accurately render well-known characters or people from cultural knowledge. The **Information** branch includes **Stock** and **Weather**, which require grounding up-to-date or structured real-world information into images. These tasks test whether image agents can go beyond prompt-local semantics and leverage retrieval or world knowledge to produce contextually correct outputs.

Memory-Driven Tasks Memory-driven tasks evaluate whether a model can preserve and reuse context across turns. This capability is essential for interactive image agents that must remain consistent with user preferences and prior dialogue history. IA-Bench includes **User Profile** and **Conversation History** task families. In user-profile tasks, the model must remember persistent user attributes, such as identity, profession, or preferred visual style, and incorporate them into later generations. In conversation-history tasks, the model must integrate previously generated content or earlier instructions into subsequent outputs, ensuring cross-turn consistency and correct composition. These tasks explicitly test whether the model

can maintain coherent long-range context rather than treating each generation request independently.

4.2 Benchmark Construction

IA-Bench is constructed through careful human annotation with explicit attention to both quality and difficulty. During prompt collection, we filter out instances that can be solved by memorization or pretrained visual priors rather than the intended capability. For example, in IP-related tasks, we exclude highly iconic characters that text-to-image models can often generate correctly without external search. For each task, we further verify feasibility and minimize ambiguity in evaluation.

For checklist construction, annotators first use LLMs to generate candidates, which are then manually reviewed and refined to ensure that each item is correct and necessary. For memory-oriented tasks, we further design dynamic evaluation checklists, as the reference may be determined by images generated in earlier interaction turns rather than a static target.

4.3 Evaluation Criterion

To enable objective and fine-grained evaluation, we adopt a checklist-based evaluation protocol. For each test instance i, let I^i_{gen} denote the generated image and $\mathcal{C}^i = \{c^i_j\}_{j=1}^{K_i}$ denote its associated checklist, where each item corresponds to a required visual condition. We use a VLM to determine whether the generated image satisfies each checklist item. We report two complementary metrics:

Pass Rate (PR) Pass rate measures strict task success. An instance is considered successful only when all checklist items are satisfied:

$$\text{PR} = \frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{K_i}\text{VLM}(I^i_{\text{gen}}, c^i_j).$$

Checklist Accuracy (CA) Checklist accuracy measures the average proportion of checklist items satisfied by the generated image:

$$\text{CA} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{K_i}\sum_{j=1}^{K_i}\text{VLM}(I^i_{\text{gen}}, c^i_j)\right).$$

Pass rate reflects strict end-to-end completion under all required constraints, whereas Checklist accuracy captures partial compliance in multi-constraint generation settings. Together, they characterize both holistic completion and partial fulfillment.

Image Agent score (IA-score) To summarize overall agent performance across different capability dimensions, we further report **IA-score**, a weighted aggregate score over the four core dimensions in IA-Bench: *Plan*, *Reason*, *Search*, and *Memory*. Specifically, IA-score is defined as

$$\text{IA-score} = 0.3 \times \text{Plan} + 0.3 \times \text{Reason} + 0.3 \times \text{Search} + 0.1 \times \text{Memory}.$$

Here, *Plan*, *Reason*, *Search*, and *Memory* denote the micro average evaluation scores for their respective dimensions. We assign higher weights to *Plan*, *Reason*, and *Search*, as these dimensions capture the core capabilities required for real-world image-agent tasks, while *Memory* is included as a complementary factor for measuring cross-step consistency and context retention.

5 Experiments

5.1 Experimental Settings

Benchmarks To comprehensively evaluate the ability of existing methods, we consider three benchmarks. First, our proposed **IA-Bench**, which measures four core agentic capabilities including Plan, Reason, Search and Memory. Second, **WISE-Verified** (Niu et al., 2025), a human-verified version of WISE, assesses semantic understanding and world knowledge in image generation models. Finally, **MindBench** (He et al., 2026a) evaluates the use of dynamic external knowledge and multi-step reasoning.

Model Name	Checklist Accuracy (%)				Pass Rate (%)				IA-score
	Plan	Reason	Search	Memory	Plan	Reason	Search	Memory	
Closed-source Image Generation Models									
GPT-Image-1.5 (OpenAI, 2025)	55.1	55.6	55.2	87.6	23.3	36.7	35.0	72.0	35.7
Nano Banana (Google DeepMind, 2025b)	68.0	63.9	61.7	60.2	42.0	43.3	42.2	48.0	43.1
Nano Banana Pro (Google DeepMind, 2025a)	60.8	66.2	68.3	72.0	32.7	44.3	47.8	52.0	42.6
Seedream-5.0-Lite (ByteDance Seed, 2025)	71.3	58.3	50.1	66.4	46.0	37.0	21.1	48.0	36.0
Qwen-Image-2.0 (Zhao et al., 2026)	50.0	48.2	38.0	51.8	20.0	27.7	6.7	11.0	17.4
Open-source Image Generation Models									
SD-3.5-medium (Stability AI, 2024b)	15.6	6.9	20.4	5.9	0.0	4.0	3.3	0.0	2.2
SD-3.5-large (Stability AI, 2024a)	19.0	9.2	24.2	10.5	0.0	5.7	6.1	1.0	3.6
FLUX.2-dev (Labs, 2024)	29.4	23.2	33.1	52.9	5.3	15.0	9.4	11.0	10.0
Bagel (Deng et al., 2025)	20.0	12.6	15.1	4.7	0.7	4.0	0.6	0.0	1.6
Bagel w/ CoT (Deng et al., 2025)	22.6	26.7	12.8	5.9	2.0	19.0	0.6	0.0	6.5
Echo-4o (Ye et al., 2025)	22.1	11.6	17.1	7.9	0.7	4.0	0.6	0.0	1.6
Echo-4o w/ CoT (Ye et al., 2025)	17.4	10.1	9.3	7.6	0.0	4.0	0.6	0.0	1.4
Qwen-Image (Wu et al., 2025a)	30.0	28.2	35.1	41.1	4.7	17.7	6.1	9.0	9.4
Agentic Image Generation Models									
GenSearcher (Feng et al., 2026)	37.0	30.1	46.5	46.6	9.3	20.3	24.4	11.0	17.3
GEMS (He et al., 2026b)	70.6	28.4	49.4	52.6	41.3	18.3	18.9	13.0	24.9
MindBrush (He et al., 2026a)	56.1	51.8	53.6	53.1	28.0	32.7	35.6	13.0	30.2
SCOPE (Ren et al., 2026)	73.3	45.2	44.4	45.2	46.7	30.0	23.3	9.0	30.9
Qwen-Image-Agent	72.9	65.5	67.6	73.6	45.3	43.7	46.1	49.0	**45.4**

Table 1: Results on IA-Bench. We report checklist accuracy, pass rate, and the overall IA-score, all measured in percentage (%). For all metrics, higher values indicate better performance.

Baselines We compare Qwen-Image-Agent against **proprietary models**, including GPT-Image-1 (OpenAI, 2024), GPT-Image-1.5 (OpenAI, 2025), Nano Banana (Google DeepMind, 2025b), Nano Banana Pro (Google DeepMind, 2025a), FLUX.2-pro (Black Forest Labs, 2026b), FLUX.2-max (Black Forest Labs, 2026a), Seedream-5.0-Lite (ByteDance Seed, 2025) and Qwen-Image-2.0 (Zhao et al., 2026), as well as state-of-the-art **open-source models** including Stable Diffusion series (Podell et al., 2023; Rombach et al., 2022; Stability AI, 2024c;b;a), FLUX series (Labs, 2024; Labs et al., 2025; Black Forest Labs, 2026b), Janus series (Wu et al., 2024; Chen et al., 2025), Z-Image (Cai et al., 2025), Qwen-Image (Wu et al., 2025a), and **unified generation models** including UniWorld-V1 (Lin et al., 2025), Bagel (Deng et al., 2025), Echo-4o (Ye et al., 2025), and DraCo (Jiang et al., 2025). We also include a wide range of **agentic generation models** including GEMS (He et al., 2026b), MindBrush (He et al., 2026a), GenSearcher (Feng et al., 2026) and SCOPE (Ren et al., 2026). All baselines are evaluated in their default settings.

Implementation Details We employ Qwen-Image-2.0 as the image generation and edit backbone, and GPT-5.5-0424 as the MLLM backbone. Regarding search tools, we utilize Google Search API for web search and image search. We set the limit of text search to 5, and the limit of image search to 5. We further utilize Jina API to process visited web pages. To ensure a fair comparison, all agentic generation baselines are evaluated under the same experimental setting, using GPT-5.5-0424 as the MLLM backbone and Qwen-Image-2.0 as the image generation and edit backbone. For the feedback loop, we allow up to 3 feedback attempts on IA-Bench, while disabling the feedback loop on WISE-Verified and MindBench to enable direct comparison with non-agentic methods. In IA-Bench, for the baselines without multiturn abilities, we append the previous turn information as prompt for testing.

5.2 Quantitative Results

We present the quantitative results on IA-Bench in Table 1. As shown, Qwen-Image-Agent achieves the highest IA-score, outperforming strong closed-source baselines such as Nano Banana Pro and GPT-Image-1.5. Compared with the direct generation baseline Qwen-Image-2.0, our agentic framework improves the Q-score substantially, **from 17.4 to 45.4**. In comparison with other agentic image generation methods, Qwen-Image-Agent achieves strong performance across the Plan, Reason, and Search dimensions, which we attribute to its unified, context-centered framework. More importantly, it shows a particularly large improvement in the Memory dimension, which hightlights its practical value in real-world, multi-turn image generation scenarios.

From the overall comparison, we observe that agentic generation models consistently outperform direct generation models on core agentic capabilities such as Plan, Reason, and Search. At the same time, closed-source models still maintain a noticeable advantage in Memory compared with agentic methods.

Model Name	Culture	Time	Space	Biology	Physics	Chemistry	Overall
Nano Banana Pro (Google DeepMind, 2025a)	0.8975	0.8167	0.9333	0.8167	0.8667	0.8750	0.8760
GPT-Image-1.5 (OpenAI, 2025)	0.8900	0.6917	0.8833	0.8000	0.7583	0.7750	0.8250
Qwen-Image-2.0 (Zhao et al., 2026)	0.8219	0.6500	0.8992	0.7917	0.8000	0.7479	0.7954
Bagel (w/ CoT) (Deng et al., 2025)	0.7800	0.6333	0.5667	0.3750	0.5500	0.5083	0.6280
Bagel (Deng et al., 2025)	0.4125	0.3500	0.3083	0.2000	0.4417	0.2583	0.3520
Janus-Pro-7B (Chen et al., 2025)	0.3700	0.3500	0.2833	0.2833	0.4000	0.2333	0.3340
Janus-Pro-1B (Chen et al., 2025)	0.3050	0.2333	0.2333	0.2167	0.3083	0.2000	0.2650
Janus-1.3B (Wu et al., 2024)	0.3175	0.2833	0.1833	0.2250	0.3417	0.1833	0.2730
FLUX.2-dev (Black Forest Labs, 2026b)	0.6650	0.5667	0.6583	0.3667	0.5250	0.3750	0.5650
FLUX.2-klein-9B (Black Forest Labs, 2026b)	0.4900	0.3917	0.5500	0.3833	0.4833	0.2250	0.4400
FLUX.2-klein-4B (Black Forest Labs, 2026b)	0.4400	0.3667	0.4667	0.3167	0.3917	0.3333	0.4010
FLUX.1-dev (Labs, 2024)	0.5225	0.4000	0.5333	0.1750	0.3750	0.2417	0.4160
FLUX.1-schnell (Labs, 2024)	0.4650	0.3250	0.4667	0.2083	0.3833	0.1000	0.3640
SD-3.5-large (Stability AI, 2024a)	0.4900	0.4083	0.4417	0.3000	0.3750	0.2083	0.4040
SD-3.5-medium (Stability AI, 2024b)	0.4825	0.3750	0.3750	0.1833	0.3917	0.2000	0.3760
SD-3-medium (Stability AI, 2024c)	0.4700	0.4083	0.4000	0.2000	0.3750	0.2583	0.3850
SD-XL-0.9 (Podell et al., 2023)	0.4925	0.3667	0.2417	0.2667	0.3333	0.1833	0.3640
SD-1.5 (Rombach et al., 2022)	0.4450	0.3083	0.2083	0.2083	0.2167	0.1500	0.3090
Qwen-Image (Wu et al., 2025a)	0.6275	0.5250	0.5583	0.3417	0.4833	0.2500	0.5100
Qwen-Image-2512 (Wu et al., 2025a)	0.5950	0.4750	0.6000	0.3500	0.4917	0.2583	0.4990
UniWorld-V1 (Lin et al., 2025)	0.5150	0.4917	0.5500	0.2250	0.4000	0.1667	0.4260
Z-Image (Cai et al., 2025)	0.5475	0.4667	0.5083	0.3250	0.4750	0.1750	0.4530
Qwen-Image-Agent	**0.9200**	**0.9167**	**0.9333**	**0.8333**	**0.8667**	**0.9000**	**0.9020**

Table 2: Results on WISE-Verified. Best results are shown in bold.

Model Name	Knowledge-Driven					Reasoning-Driven					Overall
	SE	Wth	MC	IP	WK	SL	Poem	LifeR	GU	Math	
GPT-Image-1 (OpenAI, 2024)	0.32	0.06	0.22	0.02	0.16	0.32	0.10	0.24	0.10	0.12	0.17
GPT-Image-1.5 (OpenAI, 2025)	0.36	0.18	0.22	0.04	0.30	0.34	0.08	**0.34**	0.10	0.02	0.21
FLUX.2-pro (Black Forest Labs, 2026b)	0.38	0.12	0.08	0.00	0.20	0.44	0.64	0.18	0.04	0.02	0.21
FLUX.2-max (Black Forest Labs, 2026a)	0.44	0.12	0.10	0.04	0.38	0.40	0.50	0.20	0.02	0.06	0.23
Nano Banana (Google DeepMind, 2025b)	0.30	0.10	0.12	0.00	0.30	0.32	0.36	0.20	0.04	0.08	0.18
Nano Banana Pro (Google DeepMind, 2025a)	0.50	**0.36**	0.40	**0.16**	**0.56**	**0.62**	0.68	0.30	**0.16**	**0.46**	0.41
SDXL (Podell et al., 2023)	0.04	0.00	0.04	0.00	0.00	0.00	0.00	-	-	-	0.01
SD-3.5-medium (Stability AI, 2024b)	0.02	0.00	0.00	0.00	0.02	0.00	0.00	-	-	-	0.01
SD-3.5-large (Stability AI, 2024a)	0.04	0.00	0.02	0.00	0.02	0.00	0.06	-	-	-	0.01
FLUX.1-dev (Labs, 2024)	0.04	0.00	0.00	0.00	0.02	0.02	0.04	-	-	-	0.02
FLUX.1-kontext (Labs et al., 2025)	0.02	0.00	0.00	0.00	0.02	0.00	0.00	-	-	-	0.01
FLUX.1-krea (Labs, 2024)	0.04	0.00	0.04	0.00	0.02	0.00	0.02	-	-	-	0.02
Bagel (Deng et al., 2025)	0.02	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.00	0.08	0.02
Echo-4o (Ye et al., 2025)	0.04	0.00	0.00	0.00	0.00	0.02	0.06	0.02	0.02	0.02	0.02
DraCo (Jiang et al., 2025)	0.02	0.00	0.02	0.00	0.00	0.02	0.02	0.04	0.02	0.06	0.02
Z-Image (Cai et al., 2025)	0.02	0.00	0.08	0.02	0.00	0.00	0.00	-	-	-	0.02
Qwen-Image (Wu et al., 2025a)	0.08	0.00	0.04	0.00	0.00	0.04	0.00	0.04	0.00	0.00	0.02
Qwen-Image-2.0 (Zhao et al., 2026)	0.19	0.24	0.23	0.04	0.12	0.42	0.58	0.12	0.02	0.28	0.23
Qwen-Image-Agent	**0.60**	0.28	**0.70**	**0.16**	0.28	0.58	**0.82**	0.24	**0.20**	0.34	**0.42**

Table 3: Results on MindBench. Best results are in bold and the second best ones are underlined.

These findings suggest that IA-Bench is a valid and informative benchmark for evaluating image agents, while also shedding light on promising directions for future research in agentic image generation.

Moreover, Qwen-Image-Agent delivers outstanding performance on both WISE-Verified, which emphasizes world knowledge, and MindBench, which focuses on complex reasoning and the use of external knowledge. As shown in Table 2, on WISE-Verified, Qwen-Image-Agent achieves state-of-the-art performance, surpassing the previous SOTA model, Nano Banana Pro. The results on MindBench are reported in Table 3, where Qwen-Image-Agent also sets a new state of the art. In particular, compared with the direct generation baseline Qwen-Image-2.0, our agentic framework improves performance by 82.6%. These results further demonstrate the practical effectiveness and generalizability of our proposed agentic

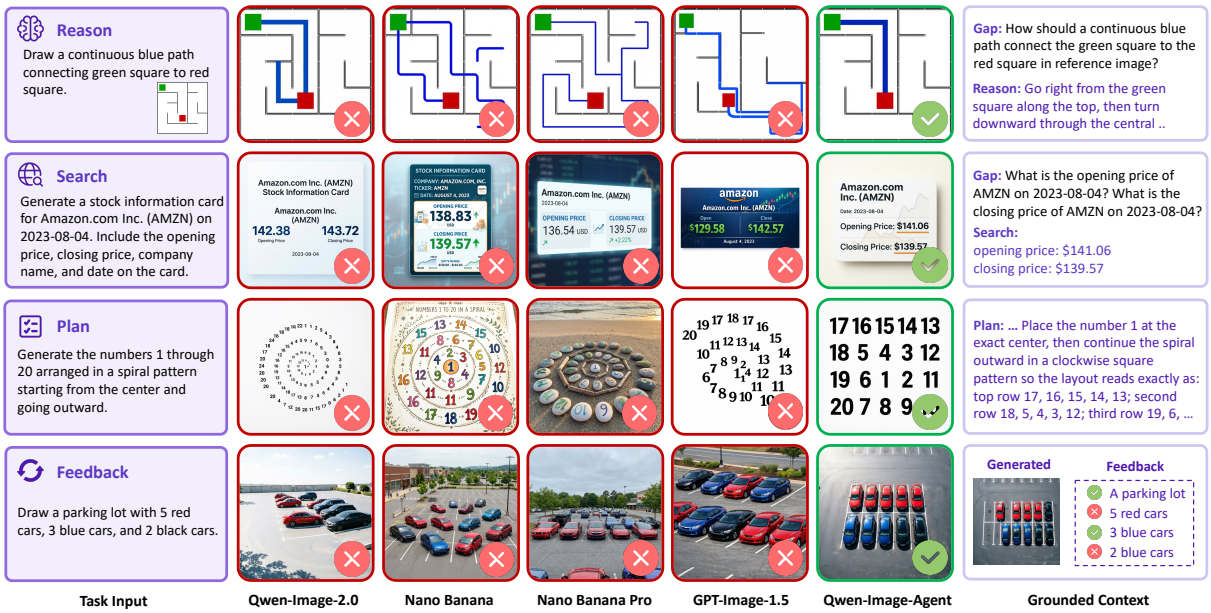

Figure 4: Qualitative Comparison of different models on IA-Bench, which demonstrates different capabilities of Qwen-Image-Agent, including Plan, Reason, Search, and Feedback.

Framework	MLLM backbone	Gen. backbone	Pass Rate (%)				IA-score
			Plan	Reason	Search	Memory	
Qwen-Image-Agent	GPT-55	Qwen-Image-2.0	45.3	43.7	46.1	49.0	45.4
w/o Reason	GPT-55	Qwen-Image-2.0	24.7	29.7	46.1	49.0	35.1
w/o Search	GPT-55	Qwen-Image-2.0	46.0	44.3	7.8	49.0	34.3
w/o Memory	GPT-55	Qwen-Image-2.0	45.3	43.7	46.1	0.0	40.5
w/o Feedback	GPT-55	Qwen-Image-2.0	40.0	41.3	42.8	49.0	42.1
Qwen-Image-Agent	GPT-55	Qwen-Image	19.3	30.7	31.1	40.0	28.3
Qwen-Image-Agent	Qwen	Qwen-Image-2.0	24.7	41.7	19.4	21.0	27.8

Table 4: Ablation study on Grounded Context, MLLM Backbone, and Generation Backbone, conducted on IA-Bench using Pass Rate as metric. Metrics with significant decreases are marked in green

framework across diverse image generation tasks.

5.3 Qualitative Results

Figure 4 presents a qualitative comparison between Qwen-Image-Agent and strong baselines, including Qwen-Image-2.0, NanoBanana, NanoBanana Pro, and GPT-Image-1.5. Although Qwen-Image-Agent is built upon Qwen-Image-2.0, it substantially improves generation quality on complex real-world tasks by bridging the context gap through our agentic pipeline. Instead of directly treating the user request as the final generation condition, Qwen-Image-Agent progressively transforms incomplete user context into sufficient generation context for image synthesis.

As shown in the figure, Qwen-Image-Agent can infer the correct maze trajectory in the reasoning case, retrieve accurate stock information in the search case, generate the specified spiral layout in the planning case, and verify object attributes and composition in the feedback case. In contrast, existing baselines often fail when the required context is implicit, missing, or needs to be grounded before generation. These examples demonstrate the effectiveness of our proposed pipeline and highlight the importance of addressing the context gap in real-world image generation.

5.4 Ablation Study

Ablations on Grounded Context To validate the effectiveness of grounded contexts in Qwen-Image-Agent, we conduct comprehensive ablation studies on different types of grounded contexts, including *Reason Context*, *Search Context*, *Memory Context*, and *Feedback Context*, using IA-Bench evaluation protocols.

As shown in Table 4, removing any grounded context leads to a clear drop in its corresponding evaluation dimension. This not only verifies the effectiveness of our context design, but also supports the validity of IA-Bench, as each dimension is sensitive to the capability it is intended to measure. We also observe that removing *Reason Context* degrades both Reason and Plan. This is because some implicit user requirements, such as enumeration, are resolved during reasoning and then reflected in planning. By contrast, removing *Feedback Context* causes a relatively smaller drop, which we attribute to the strong rendering accuracy of Qwen-Image-2.0. Overall, these results support our main claim that bridging the context gap greatly improves real-world image generation.

Ablations on MLLM Backbone To study the impact of the MLLM backbone, we conduct ablations on the backbone choice. By default, we use GPT-5.5-0424 as the MLLM backbone. In the ablation setting, we replace it with Qwen-Plus as the LLM backbone and Qwen-VL-Max as the VLM backbone. As shown in Table 4, replacing the default MLLM backbone causes substantial degradation across most metrics, showing that MLLM intelligence is critical to the overall system. In particular, it is important for layout-aware planning, keyword generation and information integration in search, and relevant context selection in memory.

Ablations on Generation Backbone To investigate the impact of image renderers under a fixed generation context, we conduct ablations on the image generation and editing backbones. By default, we use Qwen-Image-2.0 as the image generation and edit backbone. In the ablation setting, we use Qwen-Image as the generation backbone and Qwen-Image-Edit as the edit backbone. Table 4 shows that changing the generation backbone leads to consistent performance drops across all metrics. This suggests that strong generation and editing capability is also necessary for the full system. Even with a complete prompt and correct planning, some tasks remain difficult due to renderer limitations, such as counted composition, visually grounded reasoning, and accurate visual reference following.

5.5 Discussion

Through our experiments, we identify and summarize several important challenges and common failure modes in agentic image generation. These findings explain where current systems still struggle, shedding light on the main bottlenecks beyond direct image rendering.

Unidentified Context Gaps One of the central challenges in agentic image generation is identifying the gap between user context and generation context. Still, in some user cases, the context gap remains too implicit to be reliably identified, such as when the model must infer a historical event from a specific date and location stated in prompt. We find that such failures cannot be addressed by a stronger Generation backbone, since the bottleneck lies before rendering. Instead, they largely depend on the intelligence of the MLLM backbone for recognizing the missing context. Thus, we adopt a stronger MLLM backbone and substantially improves the overall system.

Ambiguous Boundary between Reason and Search The boundary between reasoning and search is often unclear. Some facts can be solved either by parametric knowledge or by external retrieval, depending on the capability boundary of the MLLM backbone. In our framework, we treat commonsense facts as solvable by internal reasoning, and define two categories that require explicit search: Precise Facts, which demand exact factual accuracy such as specific numbers, dates, and names, and Dynamic Facts, which change over time. We find that this definition helps decouple reasoning from search in a principled way, and our ablation results further support the effectiveness of this design.

Excessive Image Search Although image search provides useful visual grounding, excessive image search may hurt final generation quality: (1) Current editing models are generally less robust than direct generation models, and multi-reference editing is often more brittle than single-reference conditioning. (2) Irrelevant or weakly related reference images introduce harmful visual bias and degrade the final output. In particular, we observe that some agentic baselines, such as GenSearcher, tend to overuse image retrieval, which introduces distracting visual references and degrades the output. This issue is closely related to the IP capability of the Generation backbone. We therefore adapt the boundary of image search to the capability of the underlying generator. In our case, Qwen-Image-2.0 still lags behind the strongest models on IP-related tasks, we thus explicitly invoke image search for clear IP reference needs, while enforcing relatively strict constraints to avoid unnecessary visual retrieval.

Context Explosion in Multiturn Generation A major challenge in multiturn generation is context explosion, especially the rapid growth of image-token context. Across multiple turns, the system may need to process user-provided image references, previously generated images, and images retrieved

from search, all of which consume substantial visual context. We observe cases where such accumulated multimodal context already exceeds the token limits of strong baselines such as Nano Banana and Nano Banana Pro, leading to generation failure. To mitigate this issue, our system performs relevance-based context selection rather than naively retaining all historical inputs. This substantially alleviates context explosion and is critical for maintaining stable performance in long-horizon multiturn interactions.

Weak Feedback Supervision We also observe that the gains from feedback are relatively limited in our current setting. We attribute this to two main reasons. (1) First, our current feedback mechanism is implemented as a prompt-based feedback loop at the generation stage. In future work, we plan to extend feedback beyond post-hoc critique, so that it can also supervise context-gap identification and context grounding earlier in the pipeline. (2) Second, because we target general-purpose scenarios, we currently rely on VLM-generated feedback checklists as a generic feedback signal. In many real applications, however, one can introduce more explicit and task-specific supervision, such as predefined downstream metrics, generation quality criteria, or learned reward models. Such signals would provide clearer and more targeted feedback, and could potentially support stronger test-time scaling.

High Latency and Cost The full agentic pipeline inevitably introduces higher latency and cost than direct generation, since it may involve plan, reason, search, context integration, generation, and feedback loop. To mitigate this, we organize both information-level planning and generation-level planning with DAG-based execution, enabling as much parallelism as possible. Still, the overall pipeline remains substantially more expensive than one-shot generation. This highlights the need for more efficient agentic pipelines, potentially through training-based optimization or better tool-use policies.

6 Conclusion

In this work, we identify the context gap as a central challenge in real-world image generation. To address it, we propose **Qwen-Image-Agent**, a unified agentic framework that integrates plan, reason, search, memory and feedback in a context-centric manner. We further introduce **IA-Bench**, a benchmark for systematically evaluating four core capabilities of agentic image generation: Plan, Reason, Search, and Memory. Overall, our work highlights a shift from direct image generation to agentic image generation, and provides a unified context-centric perspective for understanding this transition. We hope our work offers practical guidance for building future image agents that can go beyond direct prompt rendering and better address real-world user needs.

References

Black Forest Labs. Flux 2 max: Next generation image synthesis. https://bfl.ai/models/flux-2-max, 2026a.

Black Forest Labs. Flux 2 pro: State-of-the-art quality at maximum speed. https://bfl.ai/models/flux-2, 2026b.

ByteDance Seed. Deeper thinking, more accurate generation: Introducing seedream 5.0 lite. https://seed.bytedance.com/zh/blog/deeper-thinking-more-accurate-generation-introducing-seedream-5-0-lite, 2025. Accessed: 2025-06-19.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

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.

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.

Kaituo Feng, Manyuan Zhang, Shuang Chen, Yunlong Lin, Kaixuan Fan, Yilei Jiang, Hongyu Li, Dian Zheng, Chenyang Wang, and Xiangyu Yue. Gen-searcher: Reinforcing agentic search for image generation. 2026. URL https://api.semanticscholar.org/CorpusID:286975158.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *ArXiv*, abs/2310.11513, 2023. URL https://api.semanticscholar.org/CorpusID:264288728.

Google DeepMind. Gemini image pro: High-quality image generation. https://deepmind.google/models/gemini-image/pro/, 2025a. Accessed: 2026-01-26.

Google DeepMind. Gemini image: High-quality image generation. https://deepmind.google/models/gemini-image/flash/, 2025b. Accessed: 2026-01-26.

Jun He, Junyan Ye, Zilong Huang, Dongzhi Jiang, Chenjue Zhang, Leqi Zhu, Renrui Zhang, Xiang Zhang, and Weijia Li. Mind-brush: Integrating agentic cognitive search and reasoning into image generation. *ArXiv*, abs/2602.01756, 2026a. URL https://api.semanticscholar.org/CorpusID:285269721.

Zefeng He, Siyuan Huang, Xiaoye Qu, Yafu Li, Tong Zhu, Yu Cheng, and Yang Yang. Gems: Agent-native multimodal generation with memory and skills. 2026b. URL https://api.semanticscholar.org/CorpusID:286974454.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *ArXiv*, abs/2403.05135, 2024. URL https://api.semanticscholar.org/CorpusID:268296755.

Dongzhi Jiang, Renrui Zhang, Haodong Li, Zhuofan Zong, Ziyu Guo, Jun He, Claire Guo, Junyan Ye, Rongyao Fang, Weijia Li, et al. Draco: Draft as cot for text-to-image preview and rare concept generation. *arXiv preprint arXiv:2512.05112*, 2025.

Kaixun Jiang, Yuzheng Wang, Junjie Zhou, Pandeng Li, Zhihang Liu, Chen-Wei Xie, Zhaoyu Chen, Yun Zheng, and Wenqiang Zhang. Genagent: Scaling text-to-image generation via agentic multimodal reasoning. *ArXiv*, abs/2601.18543, 2026. URL https://api.semanticscholar.org/CorpusID:285050929.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

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. URL https://arxiv.org/abs/2506.15742.

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.

Fanqing Meng, Wenqi Shao, Li Ray Luo, Yahong Wang, Yiran Chen, Quanfeng Lu, Yue Yang, Tianshuo Yang, Kaipeng Zhang, Yu Qiao, and Ping Luo. Phybench: A physical commonsense benchmark for evaluating text-to-image models. *ArXiv*, abs/2406.11802, 2024. URL https://api.semanticscholar.org/CorpusID:270560653.

Yuwei Niu, Munan Ning, Mengren Zheng, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *ArXiv*, abs/2503.07265, 2025. URL https://api.semanticscholar.org/CorpusID:276929205.

OpenAI. Gpt-image-1: Models and capabilities for image generation. https://platform.openai.com/docs/models/gpt-image-1, 2024. Accessed: 2026-01-29.

OpenAI. Gpt-image-1.5: Enhanced visual reasoning and creative generation. https://platform.openai.com/docs/models/gpt-image-1.5, 2025. Accessed: 2026-01-29.

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.

Tianfei Ren, Zhipeng Yan, Yiming Zhao, Zhen Fang, Yu Zeng, Guohui Zhang, Hang Xu, Xiaoxiao Ma, Shiting Huang, Ke Xu, Wenxuan Huang, Lionel Z. Wang, Lin Chen, Zehui Chen, Jie Huang, and Feng Zhao. Scope: Structured decomposition and conditional skill orchestration for complex image generation. 2026. URL https://api.semanticscholar.org/CorpusID:288148127.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Stability AI. Stable diffusion 3.5 large. https://huggingface.co/stabilityai/stable-diffusion-3.5-large, 2024a.

Stability AI. Stable diffusion 3.5 medium. https://huggingface.co/stabilityai/stable-diffusion-3.5-medium/, 2024b.

Stability AI. Stable diffusion 3 medium. https://huggingface.co/stabilityai/stable-diffusion-3-medium, 2024c.

Kaishen Wang, Ruibo Chen, Tong Zheng, and Heng Huang. Imagent: A unified multimodal agent framework for test-time scalable image generation. *ArXiv*, abs/2511.11483, 2025. URL https://api.semanticscholar.org/CorpusID:283055363.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, De mei Li, Hang Zhang, Hao Meng, Hu Wei, Ji-Li Ni, Kai Chen, Kuang Cao, Liang Peng, Lin Qu, Min Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiao-Xue Xu, Yi Wang, Yichang Zhang, Yong-An Zhu, Yujian Wu, Yu-Jiao Cai, and Ze-Yang Liu. Qwen-image technical report. *ArXiv*, abs/2508.02324, 2025b. URL https://api.semanticscholar.org/CorpusID:280422608.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.

Mingde Yao, Zhiyuan You, King-Man Tam, Menglu Wang, and Tianfan Xue. Photoagent: Agentic photo editing with exploratory visual aesthetic planning. *ArXiv*, abs/2602.22809, 2026. URL https://api.semanticscholar.org/CorpusID:286082495.

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.

Ruijie Ye, Jiayi Zhang, Zhuoxin Liu, Zihao Zhu, Siyuan Yang, Li Li, Tianfu Fu, Franck Dernoncourt, Yue Zhao, Jiacheng Zhu, Ryan A. Rossi, Wenhao Chai, and Zhengzhong Tu. Agent banana: High-fidelity image editing with agentic thinking and tooling. *ArXiv*, abs/2602.09084, 2026. URL https://api.semanticscholar.org/CorpusID:285462350.

Bin Zhao, Chenfei Wu, De mei Li, Haoliang Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyan Lin, Kaiyuan Gao, Kuang Cao, Kun Yan, Liang Peng, Lihan Jiang, Niantong Li, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiao Xu, Xiaoyu Chen, Xihua Wang, Yan Shu, Yanran Zhang, Yi Wang, Yilei Chen, Ying Ba, Yixian Xu, Yujia Wu, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhendong Wang, Zihao Liu, Zikai Zhou, Anke Yang, Chen Cheng, Chenxu Lv, Dayiheng Liu, Fan Zhou, Han Xiong, Hongzhu Shi, Hu Wei, Hui Zhao, Ivy Liu, Jianwei Zhang, Jiawei Zhang, Kai Chen, Kang He, Le Xue, Lin Qu, Li Tang, Lu-Lu Feng, Min Wu, Minmin Sun, Na Ni, Rui Men, Shuai Bai, Si Zheng, Tao Lan, Tianqi Zhang, Tingkun Wen, Wei Wang, Wei Qiao, Weiyi Lu, Wenmeng Zhou, Xiaodong Deng, Xiaoxiao Xu, Xin Yu Fang, Xiong hui Chen, Yanan Wang, Yang Fan, Yichang Zhang, Yi-Xuan Xu, Yu Wu, Zhiyuan Ma, and Zhi Cai. Qwen-image-2.0 technical report. 2026. URL https://api.semanticscholar.org/CorpusID:288256176.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Hao Li, Zicheng Zhang, Guangtao Zhai, Junchi Yan, Hua Yang, Xue Yang, and Haodong Duan. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *ArXiv*, abs/2504.02826, 2025. URL https://api.semanticscholar.org/CorpusID:277510499.

A Appendix

A.1 Case Study

In this section, we present several case studies to demonstrate the agentic image generation capabilities of Qwen-Image-Agent, including Plan, Reason, Search, Memory, and Feedback, supporting both multi-image and multi-turn generation.

User Context

Generate the numbers 1 through 20 arranged in a spiral pattern starting from the center and going outward.

Gathered Context

Context Gap How should the numbers 1 through 20 be arranged in a spiral pattern starting from the center and going outward?
Reason Arrange them in an outward spiral like this:\n\n17 16 15 14 13\n18 5 4 3 12\n19 6 1 2 11\n20 7 8 9 10.

Generation Context

Create a clean, flat, top-down typographic graphic on a pure white background, showing exactly twenty bold black sans-serif numerals arranged in a square spiral grid. Each integer from 1 through 20 must appear exactly once, with no duplicates and no omissions; the numeral 1 appears only one time, at the center of the spiral. The spiral starts at the center with 1, moves right to 2, then continues clockwise and outward in strict sequential order. The final arrangement must read exactly as a four-row grid: top row: 17 16 15 14 13; second row: 18 5 4 3 12; third row: 19 6 1 2 11; bottom row: 20 7 8 9 10. Use evenly spaced numerals centered in an invisible grid, with crisp edges, high contrast, generous white margins, and no arrows, labels, decorative elements, extra text, repeated numbers, or missing numbers.

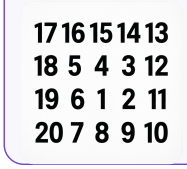

Figure 5: Case Study of planning ability. Qwen-Image-Agent solves the enumeration problem by planning the arrangement.

User Context

Draw a continuous blue path connecting green square to red square.

Gathered Context

Context Gap How should a continuous blue path connect the green square to the red square in the reference image?
Reason Go right from the green square along the top, then turn downward through the central vertical corridor to the red square.

Generation Context

Add a continuous blue path to image 1, starting at the green square. Draw the path to go right from the green square along the top route, then turn downward through the central vertical corridor, and continue until it reaches the red square. Make the line unbroken, clearly visible, and precisely aligned within the corridor path.

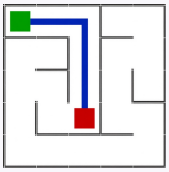

Figure 6: Case Study of reasoning ability. Qwen-Image-Agent solves the maze problem by reasoning the concrete path.

User Context

Present a clear 45° top-down view, vertical isometric 3D miniature cartoon scene, highlighting the iconic landmark in the center with fine modeling. The scene features soft and delicate textures, realistic PBR materials, and soft, realistic lighting effects. Weather elements are cleverly integrated into the urban architecture. Use a concise and unified composition for the urban landscape and atmospheric conditions, adopting minimalist aesthetics with a soft solid color background to highlight the main content. The overall visual style is fresh and soothing. Display a weather icon prominently in the top center, with the date (ultra-small font) and temperature range (medium font) below it. The city name (large font) is located below the weather graphic. The text should be consistent with the local language of the input city. Weather query time: August 14, 2025. City Name: [Toronto]

Gathered Context

Context Gap What was the weather for Toronto on August 14, 2025?
Search Mainly sunny, high 28°C; a few clouds at night, low 18°C.
Context Gap What is the iconic landmark of Toronto?
Reason CN Tower.
Context Gap Which local language should be used for Toronto?
Reason Canadian English (en-CA).

Generation Context

A vertical 45° top-down isometric 3D miniature cartoon scene of Toronto, Canada, composed like a clean weather app illustration on a soft solid pale sky-blue background. The CN Tower is the clear central landmark, standing tall and finely modeled with its slim concrete shaft, circular observation deck, antenna spire, and subtle realistic PBR surfaces; it rises from a compact stylized downtown base of simplified glass office towers, small streets, trees, and waterfront hints arranged in a concise, unified urban miniature. The scene represents mainly sunny weather on August 14, 2025: warm golden sunlight falls softly from the upper left, creating gentle realistic shadows, delicate highlights on glass and metal, and a fresh soothing atmosphere. Clever weather details are integrated into the architecture, with tiny sunlit reflections on windows, a few small soft white clouds floating between the buildings, and bright summer light wrapping around the CN Tower without obscuring it. Use soft and delicate textures, rounded miniature edges, realistic PBR materials, and a minimalist pastel color palette of pale blue, warm cream, light gray, soft green, and muted lake-blue. At the top center, prominently display a simple sunny weather icon: a warm yellow sun with a few small white cloud accents. Directly below the icon, show the date in ultra-small Canadian English text: "Aug 14, 2025". Beneath the date, show the temperature range in medium text: "18°C–28°C". Below the weather graphic, display the city name in large clean English lettering: "Toronto". The typography is crisp, centered, and modern, with the text clearly separated from the miniature city so it remains legible.

Figure 7: Case Study of web search ability. Qwen-Image-Agent solves the problem by retrieving external knowledge from web.

User Context

Generate an image of Gary De'Snake from Zootopia 2 wearing a red and green scarf lying under a Christmas tree.

Gathered Context

Context Gap What does Gary De'Snake from Zootopia 2 look like?
Search keyword: Gary De'Snake


(selected)

Generation Context

Place Gary De'Snake from image 1 lying curled comfortably beneath a decorated Christmas tree. Add a cozy red-and-green striped scarf wrapped around his neck, with the scarf ends resting on the floor. Surround him with warm white tree lights, shiny red and gold ornaments, wrapped presents, pine needles, and a soft golden holiday glow.

Figure 8: Case Study of image search ability. Qwen-Image-Agent solves the problem by retrieving visual reference from web.

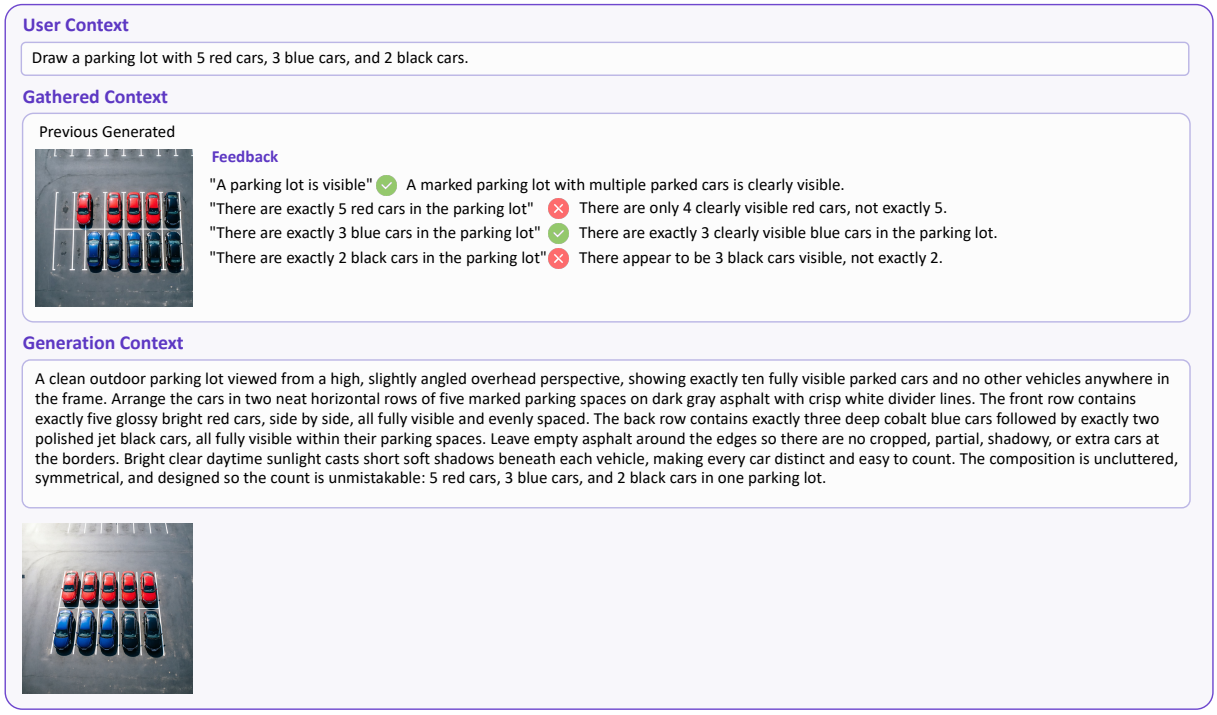

Figure 9: Case Study of feedback ability. Qwen-Image-Agent solves counted composition through self correction.

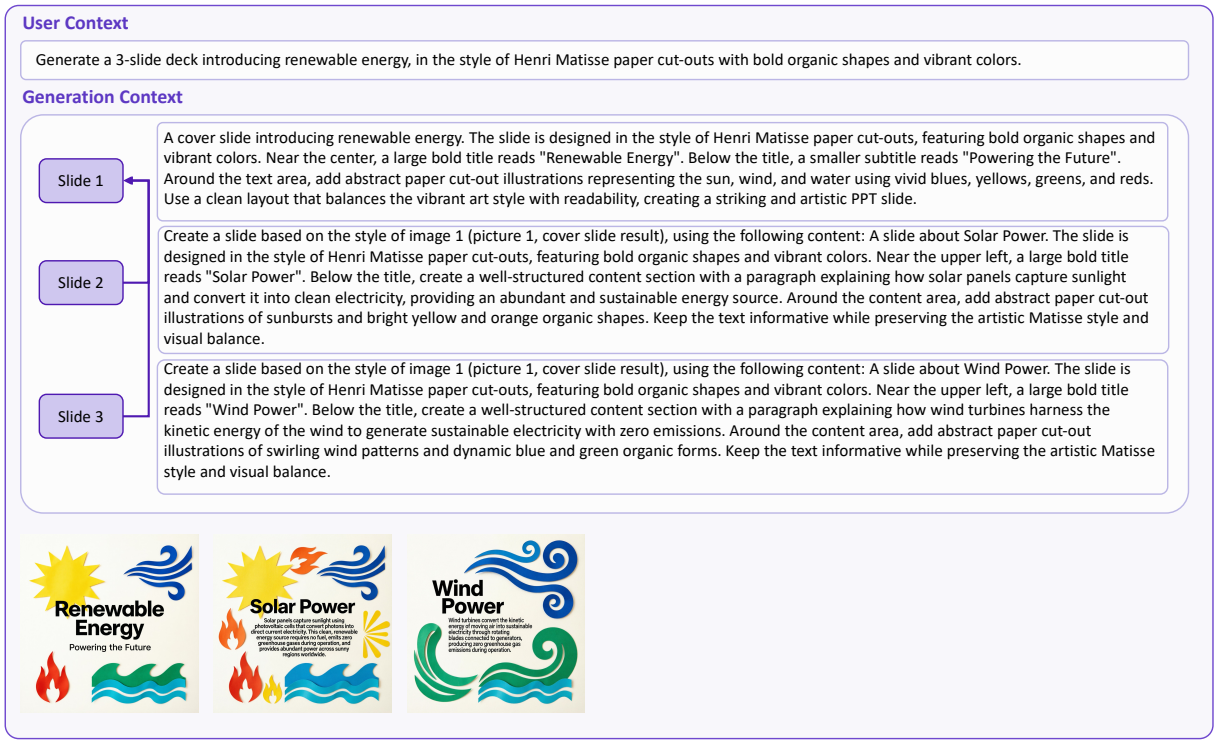

Figure 10: Case Study of multi-image ability. Qwen-Image-Agent enables multi-image generation through splitting and allocating generation context.

Conversation History

Turn 1: Generate a product image: a mini portable projector in white, with a keystone correction dial, HDMI port, and a small built-in speaker grille. Include the text "Brightness: 200 ANSI Lumens" below the product.

Turn 2: Generate a product image: a pull-down projection screen with a matte white surface, a black border, and a spring-loaded auto-retract mechanism. Include the text "Diagonal: 84 Inches" below the product.

Turn 3: Generate a product image: a USB-C to HDMI adapter in space gray aluminum, with a braided cable and 4K@60Hz support. Include the text \"Resolution: 4K@60Hz\" below the product.

User Context

Generate an image with the previous generated portable projector on the left and the previous generated projection screen on the right, side by side. Label the Brightness below the portable projector and the Diagonal below the projection screen.

Gathered Context

Memory

(selected)

Turn 1 Context

Generate a product image ... "Brightness: 200 ANSI Lumens" below the product

Turn 2 Context

Generate a product image ... "Diagonal: 84 Inches" below the product

(selected)

Generation Context

Place the portable projector from image 1 on the left and the projection screen from image 2 on the right, side by side on a clean white product-comparison layout. Add the label "Brightness: 200 ANSI Lumens" centered directly below the portable projector, and add the label "Diagonal: 84 Inches" centered directly below the projection screen. Keep both products fully visible, evenly scaled, and aligned horizontally.

Figure 11: Case Study of memory ability. Qwen-Image-Agent solves the multiturn problem by selecting relevant memory context.