ATP-Bench: Towards Agentic Tool Planning for MLLM Interleaved Generation

Yinuo Liu[1,2*], Zi Qian[1], Heng Zhou[1,3*], Jiahao Zhang[1], Yajie Zhang[1], Zhihang Li[1†], Mengyu Zhou[1], Erchao Zhao[1], Xiaoxi Jiang[1], and Guanjun Jiang[1]

[1] Qwen Business Unit, Alibaba Group, China
[2] Huazhong University of Science and Technology, China
[3] Zhejiang University, China

Abstract. Interleaved text-and-image generation represents a significant frontier for Multimodal Large Language Models (MLLMs), offering a more intuitive way to convey complex information. Current paradigms rely on either image generation or retrieval augmentation, yet they typically treat the two as mutually exclusive paths, failing to unify factuality with creativity. We argue that the next milestone in this field is *Agentic Tool Planning*, where the model serves as a central controller that autonomously determines when, where, and which tools to invoke to produce interleaved responses for visual-critical queries. To systematically evaluate this paradigm, we introduce *ATP-Bench*, a novel benchmark comprising 7,702 QA pairs (including 1,592 VQA pairs) across eight categories and 25 visual-critical intents, featuring human-verified queries and ground truths. Furthermore, to evaluate agentic planning independent of end-to-end execution and changing tool backends, we propose a Multi-Agent MLLM-as-a-Judge (MAM) system. MAM evaluates tool-call precision, identifies missed opportunities for tool use, and assesses overall response quality without requiring ground-truth references. Our extensive experiments on 10 state-of-the-art MLLMs reveal that models struggle with coherent interleaved planning and exhibit significant variations in tool-use behavior, highlighting substantial room for improvement and providing actionable guidance for advancing interleaved generation.

1 Introduction

Interleaved generation, which aims to jointly produce coherent sequences of text and images, represents a burgeoning frontier for multimodal large language models (MLLMs) [1, 16, 21, 37, 42]. Unlike text-only responses, interleaved content provides a more intuitive and efficient way to convey information [23, 31, 40, 41]. Figure 1 shows representative examples of this task. For instance, a researcher can interpret experimental results more effectively through figures than dense text (top panel); a user seeking styling advice benefits from previewing a modified hairstyle (bottom-right panel); and culinary instructions become more actionable

* Work done during an internship at Alibaba.
† Corresponding author.

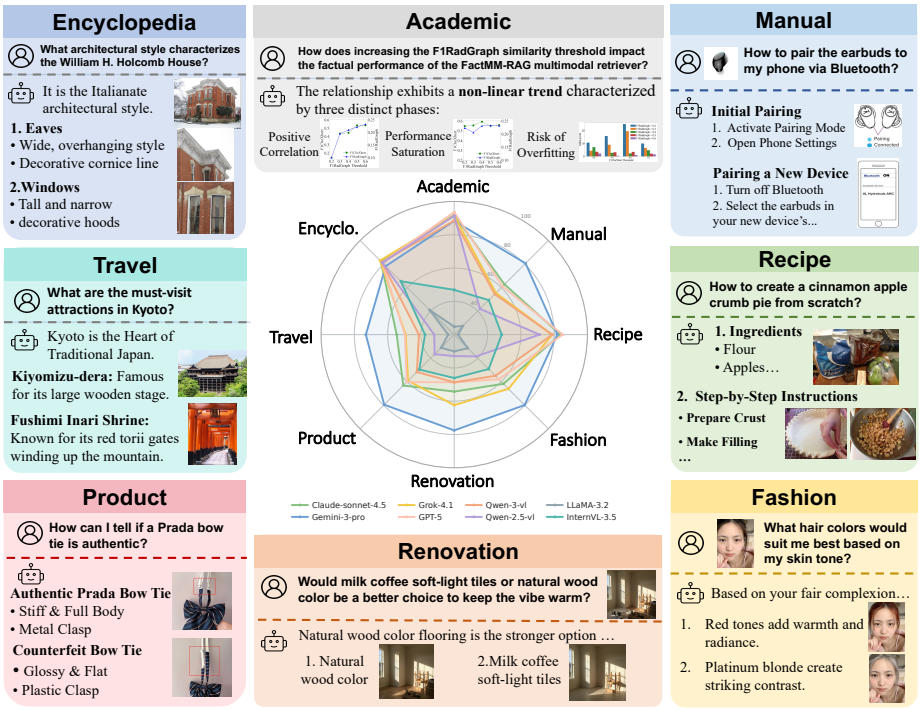


Fig. 1: Examples of eight task categories and corresponding model performance. Since our benchmark targets agentic tool-planning ability, evaluation outputs and ground truths contain tool-planning tags instead of rendered images (See Figure 2). For better interpretability, we present results obtained after end-to-end tool execution.

with step-by-step visual aids (right panel). By grounding abstract explanations in concrete visual context, interleaved generation can minimize users' cognitive load and bridge the gap between digital assistance and real-world application [6].

Research on interleaved generation has largely converged to two separate paradigms. The first paradigm focuses on *image generation* via either unified models [8, 29, 32] or diffusion-based pipelines [1, 26, 42], but it often lacks factual grounding and struggles with complex diagrams [41]. The second paradigm adopts *retrieval augmentation* by retrieving and citing images from external corpora [40, 41], but it is inherently limited in its ability to modify visuals and support query-specific creative generation. However, neither paradigm captures a common real-world setting where the model must both reference context images and generate query-specific visuals within the same interleaved response, leaving this practical requirement largely unmet.

We argue that the next milestone for interleaved generation is *Agentic Tool Planning*, which dissolves the boundary between reference and generation. In this paradigm, the model acts as a central controller, autonomously deciding *when*, *where*, and *which* tools to invoke, dynamically orchestrating capabilities such as

Table 1: Comparison of *ATP-Bench* with existing benchmarks. *ATP-Bench* is the first to integrate hybrid image sourcing (Reference & Generation) and dual query types (QA & VQA) with expert-level annotations for both queries and ground truths.

Dataset name	#Sample	Query type		Response IMG source		Annotated Query	Annotated GT
		QA	VQA	Reference	Generation		
OpenLEAF [1]	30	✓	✗	✗	✓	✗	✗
InterleavedBench [21]	815	✓	✓	✗	✓	✗	✗
OpenING [42]	5,400	✓	✓	✗	✓	✗	✗
RAG-IGBench [41]	6,057	✓	✗	✓	✗	✓	✓
MRAMG-Bench [40]	4,800	✓	✗	✓	✗	✓	✓
LLM-I [16]	30	✓	✓	✗	✓	✗	✗
Ours (*ATP-Bench*)	**7,702**	✓	✓	✓	✓	✓	✓

citing provided images, performing targeted edits, synthesizing new content, and acquiring real-world assets via web search to produce interleaved responses.

Studying this paradigm requires a dedicated benchmark that jointly models both reference and generation. In contrast, existing benchmarks typically evaluate only one side in isolation. To bridge this gap, we present *ATP-Bench* (**A**gentic **T**ool **P**lanning Bench), a novel benchmark for evaluating MLLMs on user-centric interleaved tool-planning tasks. As summarized in Table 1, *ATP-Bench* is the first to jointly support hybrid image sourcing (reference & generation) and dual query types (QA & VQA), with expert-level annotations for both queries and ground truths. We ensure benchmark reliability through a systematic data collection and validation pipeline, along with a carefully designed evaluation strategy. We define a taxonomy of visual-critical categories and intents to guide query synthesis, with expert annotators verifying clarity and visual-criticality. Ground truth is constructed through a three-stage procedure with multi-perspective human evaluation. The resulting *ATP-Bench* comprises 7,702 QA pairs, including 1,592 VQA pairs, across eight categories and 25 visual-critical intents. To evaluate agentic tool-planning ability, we further propose a Multi-Agent MLLM-as-a-Judge (MAM) system that assesses tool-call precision, missing tool-use opportunities, and overall response quality without requiring ground truth or end-to-end execution. The reliability of MAM is further validated through a human agreement study.

Key Findings. Our comprehensive evaluation across 10 state-of-the-art reveals three key insights: (1) Existing MLLMs struggle to generate coherent interleaved tool plan, particularly for Travel and Renovation; (2) Gemini 3 Pro achieves the leading performance under our task setting; (3) Models exhibit distinct behavioral tendencies in terms of tool-call frequency and preference.

Our contributions are as follows: (1) We propose a novel paradigm, namely *Agentic Tool Planning*, for interleaved generation, and introduce *ATP-Bench*, a benchmark enabling the systematic study of tool planning capabilities under real-world interleaved settings for MLLMs. (2) We propose a MAM system to assess tool-call precision and identify missed tool-use opportunities, without requiring ground truth or end-to-end execution. (3) We conduct extensive ex-

periments on state-of-the-art MLLMs, revealing their tool-planning capabilities and potential biases, and providing actionable guidance for future research.

2 Related Work

Methods for Interleaved Generation. Interleaved generation produces coherent sequences that interleave textual reasoning with visual outputs. Early work explored unified autoregressive models that align text and image tokens end-to-end (e.g., Chameleon [32], Show-o [38], LLaVA-NeXT-Interleave [20], Orthus [19], Anole [8]). In contrast, pipeline-based systems [1,16,21,35,42] decouple language modeling from visual synthesis, allowing LLMs to invoke specialized vision modules, while retrieval-augmented methods improve grounding by retrieving and citing images from external corpora [40,41]. Nonetheless, most prior work favors either generation or retrieval, rather than supporting both in one interleaved response.

Evaluation for Interleaved Generation. Recent benchmarks have increasingly systematized interleaved-generation evaluation. General-purpose datasets (e.g., OpenLEAF [1], MMIE [37], InterleavedBench [21], OpenING [42], InterSyn [11]) typically use MLLM-as-a-judge [7] to assess image quality and cross-modal coherence. Retrieval-oriented benchmarks (RAG-IGBench [41], MRAMG-Bench [40]) emphasize factual grounding via retrieval, measuring recall, precision, and semantic alignment. Specialized frameworks further extend MLLM-based evaluation to structured/interactive settings, including LLM-I [16], ISG-Bench [6], and WEAVE [9]. Overall, existing benchmarks largely evaluate generation or retrieval in isolation with end-to-end metrics, prioritizing output fidelity and alignment over open-ended tool planning.

Tool-Augmented MLLMs. Recent advancements in tool-augmented MLLMs integrate external tools to enhance visual reasoning and action capabilities. Frameworks like ViperGPT [30] and MM-ReAct [39] employ prompting pipelines for code and API execution in multimodal tasks, while AssistGPT [12], LLaVA-Plus [22], and CLOVA [13] utilize cyclic pipelines involving planning, execution, and refinement feedback. MLLM-Tool [33] adopts a learning-based approach with multimodal encoders for tool selection and agentic execution. Although tool-augmented MLLMs have inspired new paradigms for interleaved generation, such as LLM-I [16], the lack of valid datasets and benchmarks still hinders effective evaluation of their tool planning abilities.

3 Task Formulation

Our goal is to plan an interleaved multimodal response R that tightly couples text with relevant images, given a visual-critical query q and a document set $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$. The set $\mathcal{D}$ is *source-agnostic*: it may be retrieved from an external knowledge base, or provided directly by the user as context. Each document $d_i \in \mathcal{D}$ is represented as a tuple $(T_i, \mathcal{I}_i)$, where T_i denotes the textual content and $\mathcal{I}_i$ is the set of images associated with the document. To generate R,

an MLLM consumes a concatenated input comprising an interleaved generation prompt p, the query q, and the document set $\mathcal{D}$. The model then produces an ordered sequence:

$$R = \{s_1, s_2, \ldots, s_m\}, \quad s_j \in \mathcal{S}_{\text{text}} \cup \mathcal{S}_{\text{tool}}, \tag{1}$$

where $\mathcal{S}_{\text{text}}$ denotes natural-language tokens and $\mathcal{S}_{\text{tool}}$ denotes tool-calling instructions for image integration.

Visual-Critical Queries. We define *visual-critical queries* as prompts for which an interleaved image–text response offers substantially higher utility and information density than a text-only response. Such queries typically involve:

- *Visual information augmentation*: images convey essential details that are difficult to capture precisely in language.
- *Cognitive acceleration*: visuals exploit parallel perception to speed up comprehension, especially in spatially intensive tasks and hands-on procedures.
- *Structural illustration*: diagrams make structure and relationships explicit.

Toolkit. To support principled visual integration within the interleaved response, we define a unified tool-calling space $\mathcal{S}_{\text{tool}}$. Each tool invocation follows a structured schema: `<tool>{"tool_name": ..., "description": ..., "params": ...}</tool>`. The toolkit comprises five specialized modules that cover complementary needs for visual acquisition and manipulation:

- *Reference*: anchors the response to specific visual evidence in $\mathcal{D}$ by specifying an `img_index`, enabling faithful citation of in-context images.
- *Diffusion*: synthesizes novel images for conceptual illustrations or artistic renderings absent from $\mathcal{D}$. It takes a semantically detailed `prompt` describing the intended content.
- *Search*: retrieves real-world visuals via external search engines such as Google image search. It accepts a targeted search `query` to ground the response in factual, up-to-date imagery.
- *Code*: generates programmatic, data-driven visualizations such as charts and mathematical plots. It requires a diagram `type` and a specification of the `data`, ensuring precision and interpretability.
- *Edit*: modifies a referenced image from $\mathcal{D}$ for localized refinement. Given an `img_index` and an edit `prompt`, it can add labels, highlight regions, or annotate salient visual features.

4 *ATP-Bench*

This section provides a comprehensive overview of *ATP-Bench*. *ATP-Bench* dataset consists of 7,702 QA pairs (including 1,592 VQA pairs) spanning eight categories and 25 visual-critical intents. The detailed statistics are provided in Table 2. We detail the construction methodologies for both the query and ground truth collections as illustrated in Figure 2a and 2b. Subsequently, we introduce MAM, the evaluation system specifically designed for this task, as shown in the right panel of Figure 2c.

Table 2: Statistics of *ATP-Bench*. Note that "#TC per GT" stands for the number of tool calls per ground truth answer.

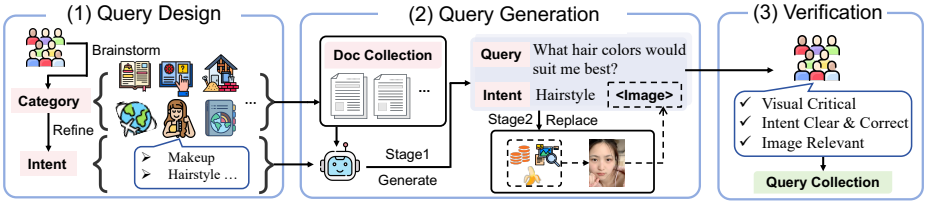
Category	Intent	#Query	#Doc per Q	#Img per Q	#TC per GT
Academic	Framework, Component, Comparison, Dataset, Result	196	1.00	3.34	1.96
Manual	Operation Guide, Function, Component	1,025	2.53	10.90	4.38
Recipe	Guideline, Ingredient	1,960	1.23	6.25	3.06
Fashion	Hairstyle, Makeup, Outfit, Photography	1,051	2.95	11.81	4.63
Renovation	Element, Style	440	2.92	11.98	4.79
Product	ProductIntro, Comparison, Authenticity	1,330	2.91	10.40	4.58
Travel	Planning, Navigation	734	2.91	16.00	8.00
Encyclopedia	Animals, Geography, Biography, Architecture	966	1.01	1.05	3.65
Overall	–	**7,702**	**2.15**	**8.87**	**4.32**

4.1 Query Collection

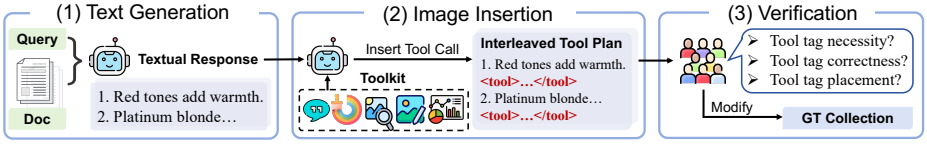
Design. We identify eight **high-visual-demand query categories** in which users are most likely to expect image-rich responses: *Academic*, *Manual*, *Recipe*, *Fashion*, *Renovation*, *Product*, *Travel*, and *Encyclopedia*. For each category, we collect visually grounded documents from existing multimodal benchmarks, including MRAMG-Bench, RAG-IGBench, and OVEN [17], which aggregate content from diverse sources such as Wikipedia, Xiaohongshu, and arXiv. We further define fine-grained **visual-critical intents** that intrinsically require visual support, resulting in **25 distinct intents** spanning the eight categories.
Generation. We employ Gemini 2.5 Pro [10] to generate queries via a two-stage pipeline. In the first stage, given a source document and its predefined intent taxonomy, the model identifies the most relevant intent(s) and synthesizes *text-only* queries conditioned on them. The prompts encourage a *natural, conversational tone* that resembles everyday user phrasing rather than explicit, tool-oriented requests. In the second stage, building on these text queries, we select intents that can be visually grounded and convert them into VQA-style queries by replacing key textual evidence with images. For such visually grounded queries, e.g., fashion advice based on user portraits, interior renovation suggestions based on photos, or encyclopedic questions about animals and landmarks, we curate a high-quality visual corpus from three sources, including web search, generative synthesis using *nano-banana* [14], and public benchmarks such as OVEN.
Verification. We employed a team of ten professional annotators to verify the quality of the generated queries. During this process, annotators removed queries with (i) ambiguous or unnatural queries, (ii) VQA pairs where the question is not grounded in the associated image, or (iii) queries that do not meet our definition of visual-critical queries. After filtering, approximately 95% of queries were retained. The queries are also relabeled with their corresponding categories and intent labels to ensure correctness, enabling reliable fine-grained analysis.

4.2 Ground Truth Collection

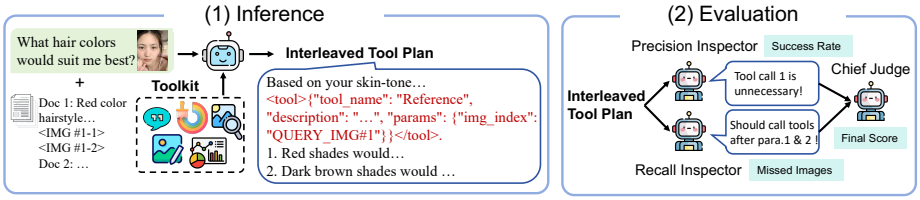
We used a three-stage process for ground truth generation.



(a) Query collection process of *ATP-Bench*.



(b) Ground truth collection process of *ATP-Bench*.



(c) Inference and evaluation process of *ATP-Bench*.

Fig. 2: Overview of *ATP-Bench* dataset construction and evaluation pipelines.

Textual Response Generation. We prompt an MLLM to generate a high-quality text-only answer grounded in the provided documents and images. The prompt enforces factual consistency with document text, accurate visual interpretation of the query and images, and coherent reasoning with complete query coverage. It also imposes task-specific formatting requirements, such as step-by-step instructions or a structured introduction with multi-aspect comparisons, and mandates an answer-first, concise style without filler.

Image Insertion. To address visual gaps, we instruct an MLLM to proactively invoke specialized visual tools and place the generated outputs immediately after the relevant paragraph. Tool usage follows a clear **capability boundary**. *Reference* supports multi-step procedures, complex diagrams, and context-dependent examples. *Search* retrieves real-world entity images such as landmarks, artworks, notable figures, events, and maps. *Diffusion* generates creative concepts, design drafts, abstract visuals, and fashion demonstrations when no suitable reference exists. *Edit* modifies provided images, including highlighting, annotation, renovation mock-ups, and style previews. *Code* produces data-driven visualizations and blueprint-style diagrams. We further enforce strict integration rules: no figure-introducing phrases, single use per reference image, no mid-sentence insertion, and a high-value visualization policy to ensure clarity and coherence.

Fine-grained Annotation and Refinement. To ensure dataset fidelity and trustworthiness, we asked 15 annotators to review every tool call. A tool call is retained only if its expected image materially improves understanding, uses correct parameters, and is placed immediately after the relevant text to preserve narrative coherence. Annotators remove redundant or semantically mismatched calls, fix issues such as malformed syntax and inappropriate tool choices, and relocate misplaced tags to paragraph boundaries to avoid disrupting readability. We discard the entire sample when errors are systemic, such as severe markdown structure breakage or fundamental factual inaccuracies in the text.

4.3 MAM: Multi-Agent MLLM-as-a-judge

Evaluating model performance in our paradigm introduces challenges that conventional metrics cannot adequately address. First, the open-ended nature of the tasks means that ground truths function as high-quality references rather than definitive standards for every tool invocation. Second, we emphasize evaluating tool planning rather than end-to-end execution, thereby isolating the model's intrinsic planning ability from external API constraints. Third, evaluation must be multi-dimensional, assessing both the precision of tool calls, including their necessity, correctness, and placement, and the recall of missed visual elements when such support is required. To address these challenges, we propose a Multi-Agent MLLM-as-a-Judge system that evaluates tool-call precision, missed tool-use opportunities, and overall response quality without relying on ground truth or full execution. The framework comprises three specialized agents: the Precision Inspector, the Recall Inspector, and the Chief Judge.

The Precision Inspector. The Precision Inspector evaluates the precision and execution quality of each tool invocation in the model response using a two-stage procedure. It first verifies visual-critical necessity, requiring that the image provides clear added value over text, and checks tool boundary compliance to ensure the selected tool matches its technical capabilities. For invocations that pass these prerequisites, it then assesses semantic placement, structural coherence with surrounding content, parameter accuracy, and output format correctness, and assigns a score on a 0–2 scale where 0 indicates failure, 1 indicates partial satisfaction, and 2 indicates full satisfaction.

The Recall Inspector. The Recall Inspector identifies missed tool-use opportunities where the model responds with text only but should have invoked a tool. It flags cases where the text explicitly refers to an absent visual, where long textual descriptions of complex procedures, reference images, or real-world entities lack appropriate *reference* or *search* grounding, and where the model omits *diffusion*, *edit*, or *code* despite an implicit need for creative generation, image modification, or plotting.

The Chief Judge. The Chief Judge synthesize the overall interleaved pacing by integrating the Precision and Recall Reports into a holistic, quantitative score ranging from 0 to 100. This scoring mechanism is strictly anchored across five performance tiers: (i) *Excellent (80–100)* represents near-perfect execution, where all tools pass necessity and boundary checks with precise syntax and zero

missed visual opportunities; (ii) *Good but Flawed (60–80)* denotes a genuinely helpful response characterized by minor precision issues or at most one minor visual omission; (iii) *Mediocre (40–60)* indicates noticeable discrepancies, including multiple parameter errors, forced structural splits, or 1–2 missed opportunities identified by the Recall Inspector; (iv) *Poor (20–40)* signifies significant failures such as severe visual redundancy, tool boundary violations, or 2–3 instances where explicit visual contexts were improperly relegated to plain text; and (v) *Fatal (0–20)* reflects a complete breakdown of execution, marked by severe formatting disruptions, or the total omission of the user's core visual intent.

5 Experiment

5.1 Setup

Inference Settings. We evaluate 10 MLLMs, namely Claude Sonnet 4.5 [3], Claude Sonnet 4 [2], Gemini 3 Pro [15], Grok-4.1 Fast Reasoning [36], GPT-5 [25], GPT-4o [18], Qwen3-VL-Plus [4], Qwen2.5-VL-72B [5], LLaMA-3.2-11B [24], InternVL3.5-14B [34]. The prompt for interleaved generation is provided in the supplementary materials, which also use the same tool boundary as in the prompt for ground-truth generation. By default, we use a zero-shot strategy to evaluate the model performance. In addition, we conduct a few-shot (3-shot) experiment on Claude Sonnet 4.5, GPT-4o, Qwen2.5-VL-72B, LLaMA-3.2-11B.
MAM Settings. By default, we use Gemini 2.5 Pro as the agents in the MAM framework. We extend the agents to Claude Sonnet 4.5 and GPT-5 in our ablation study. The prompt for agents is provided in the supplementary materials.

5.2 Evaluation Metrics

Final Score (FS). The Final Score refers to the score assigned by the Chief Judge in our MAM system. It reflects the overall quality of the response.
Success Rate (SR). The Success Rate measures tool-call execution accuracy based on the Precision Inspector. A tool call is considered successful only if it is necessary, semantically appropriate, well-placed, and uses the correct tool with the correct parameters in the proper format.
Missed Images (MI). The Missed Images count is reported by the Recall Inspector, who reviews the response and identifies missed opportunities for tool calls based on omission criteria. Lower missed image counts mean the models are more capable of filling visual-critical gaps.
Tool Adoption Rate. Thw Tool Adoption Rate measures the percentage of queries in which the model invokes a specific tool at least once. It reflects the model's overall preference for incorporating a specific tool into responses.
Precision, Recall, and F1-score. We also evaluate tool invocation by comparing model-selected tool sets against our curated ground truth. Precision and Recall measure the accuracy and completeness of tool selection, respectively, while the F1-score provides a balanced assessment of overall performance.

Table 3: Final Score of models across eight categories (Renovation and Encyclopedia is abbreviated as Renova. and Encyclo. respectively).

MLLM	Academic	Manual	Recipe	Fashion	Renova.	Product	Travel	Encyclo.	Avg
Claude Sonnet 4.5	91.93	62.74	83.86	61.79	54.22	63.25	54.56	82.39	69.34
Claude Sonnet 4	88.12	63.37	82.85	62.21	54.33	63.87	54.77	**83.69**	69.15
Gemini 3 Pro	88.32	**80.58**	81.96	**79.83**	**77.53**	**79.51**	**73.07**	78.20	**79.88**
Grok-4.1	91.43	55.39	78.83	66.05	62.37	59.20	49.18	82.96	68.18
GPT-5	**93.59**	61.00	**85.13**	57.91	48.63	60.31	49.63	81.27	67.18
GPT-4o	86.27	56.38	79.86	48.76	43.19	53.34	43.94	73.59	60.67
Qwen3-VL-Plus	88.51	54.42	78.87	54.99	48.67	51.73	41.22	81.51	62.49
Qwen2.5-VL-72B	91.36	41.90	71.32	38.49	32.90	36.61	32.06	81.11	53.22
InternVL3.5-14B	46.90	49.34	48.52	49.24	46.06	49.68	37.58	65.37	49.09
LLaMA-3.2-11B	24.80	27.34	22.34	31.48	30.08	31.15	23.64	40.94	28.97

Table 4: Success Rate of models across eight categories.

MLLM	Academic	Manual	Recipe	Fashion	Renova.	Product	Travel	Encyclo.	Avg
Claude Sonnet 4.5	92.09	66.95	81.95	71.55	65.10	71.44	72.98	**84.06**	75.77
Claude Sonnet 4	87.31	66.60	80.49	71.80	62.58	71.46	70.77	79.85	73.86
Gemini 3 Pro	78.89	**82.97**	74.11	**85.92**	**86.43**	**86.44**	**87.86**	71.55	**81.77**
Grok-4.1	90.82	57.92	74.53	69.77	71.41	60.92	62.54	81.79	71.21
GPT-5	**94.54**	60.80	**84.41**	64.92	47.62	66.43	63.08	72.01	69.23
GPT-4o	85.83	55.24	79.13	54.07	38.34	63.56	55.96	60.89	61.63
Qwen3-VL-Plus	85.63	55.81	78.01	62.65	48.39	56.60	50.78	73.57	63.93
Qwen2.5-VL-72B	83.24	28.64	57.56	31.35	19.51	24.80	21.16	46.81	39.13
InternVL3.5-14B	11.23	10.73	6.10	11.73	9.94	11.54	10.89	8.77	10.12
LLaMA-3.2-11B	0.00	20.23	16.64	27.27	24.14	25.44	19.79	13.19	18.34

5.3 Results

Main Results We report results on tool-planning quality using three metrics: final score (Table 3), success rate (Table 4), and missed images (Table 5).

Gemini 3 Pro achieves leading performance across all metrics and most categories. Gemini 3 Pro attains the highest average final score of 79.88 and tool-use success rate of 81.77, while maintaining the lowest missed image count at 0.49. It shows clear advantages in tool-intensive domains, consistently reaching final scores around 77–81 with success rates above 82 in Manual, Fashion, Renovation, and Product. Even in Travel, it achieves the highest success rate of 87.86 while sustaining competitive overall performance.

Tier-2 models achieve comparable final scores but exhibit distinct tool-use trade-offs. Claude Sonnet 4.5, Claude Sonnet 4, Grok-4.1, and GPT-5 obtain similar average final scores, yet differ noticeably in tool-use behavior. Claude Sonnet 4.5 achieves a higher success rate than Claude Sonnet 4, but also incurs more missed images, suggesting a more conservative tool-use strategy. GPT-5 performs strongly in language-intensive domains such as Academic, but degrades substantially in tool-heavy categories, with lower final scores and success rates in Renovation and Travel, accompanied by increased missed images.

GPT-4o and Qwen3-VL-Plus form a mid-tier with weaker visual grounding coverage and execution robustness. GPT-4o and Qwen3-VL-Plus obtain similar average final scores, 60.67 and 62.49, but lag behind Tier-2 models in success rate and exhibit higher missed-image counts. The gap is most

Table 5: Missed Image of models across eight categories; lower is better.

MLLM	Academic	Manual	Recipe	Fashion	Renova.	Product	Travel	Encyclo.	Avg
Claude Sonnet 4.5	0.10	1.53	0.29	1.29	1.78	1.31	2.93	0.30	1.19
Claude Sonnet 4	0.06	1.18	**0.23**	1.09	1.55	1.15	2.38	0.25	0.99
Gemini 3 Pro	0.15	**0.44**	0.25	**0.46**	**0.52**	**0.52**	**1.31**	0.24	**0.49**
Grok-4.1	**0.05**	1.55	0.44	0.96	1.22	1.27	2.75	**0.22**	1.06
GPT-5	0.14	1.32	0.31	1.30	1.67	1.29	2.60	0.26	1.11
GPT-4o	0.28	1.50	0.51	1.65	1.80	1.68	3.03	0.36	1.35
Qwen3-VL-Plus	0.17	1.75	0.55	1.70	1.98	1.83	3.25	0.31	1.44
Qwen2.5-VL-72B	0.15	2.11	0.70	2.00	2.21	2.38	3.65	0.27	1.68
InternVL3.5-14B	0.98	1.84	1.65	1.49	1.71	1.69	3.28	0.34	1.62
LLaMA-3.2-11B	1.37	2.32	2.03	2.07	2.26	2.26	3.82	1.11	2.16

evident in tool-intensive categories such as Travel and Renovation, where both models show reduced success rates and substantially elevated missed images. Overall, their performance is constrained by insufficient visual gap detection and less robust tool execution when image use is required.

Open-source models underperform overall, with heterogeneous failure patterns. Qwen2.5-VL-72B performs well in knowledge-intensive categories such as Academic and Encyclopedia, but drops sharply in tool-heavy domains, accompanied by a low average success rate and elevated missed images, indicating limited tool planning ability. InternVL3.5-14B shows fewer missed images than LLaMA-3.2-11B but an extremely low success rate, suggesting frequent invalid tool calls due to formatting issues. LLaMA-3.2-11B performs worst overall, with low final scores, low success rates, and high missed images, reflecting compounded weaknesses in both visual gap detection and valid tool execution.

Across categories, Academic and Encyclopedia are the easiest, while Travel and Renovation are the hardest. Averaged across models, Academic and Encyclopedia achieve the highest final scores with the fewest missed images, reflecting limited need for complex tool coordination. In contrast, Travel is the weakest category overall, characterized by the lowest final scores across most models and substantially elevated missed-image counts, indicating failures in identifying tool-use opportunities. Renovation is also consistently challenging, but for a different reason: several models exhibit markedly reduced success rates despite only moderately increased missed images, suggesting that errors stem more from incorrect tool execution. Overall, Travel is primarily MI-dominated, whereas Renovation is more SR-dominated.

Tool Call Number Distribution. Figure 3 illustrates distinct tool-use strategies across models. Top-performing models cluster around three to five calls, indicating calibrated planning that avoids both under-invocation and redundancies. Mid-tier models skew toward fewer calls, often peaking around one to three, reflecting conservative triggering behavior that aligns with their higher missed-image counts in tool-heavy categories. In contrast, smaller open-source models exhibit more extreme behaviors: LLaMA-3.2-11B concentrates mass at zero calls, suggesting systematic under-use and MI-dominated errors, while InternVL3.5-14B shows a long tail toward high call counts, consistent with SR-dominated failures due to difficulties in producing valid tool calls.

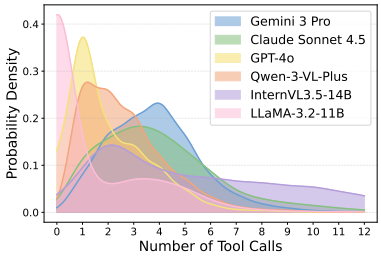


Fig. 3: Tool call number distribution for representative models.

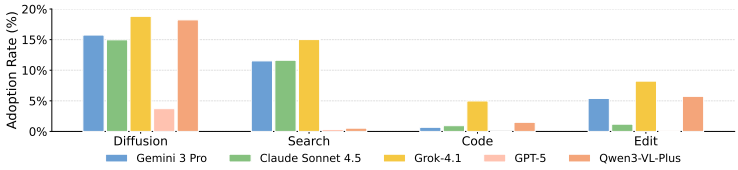


Fig. 4: Tool Adoption Rates per query for representative models. The reference tool is omitted due to its high dominance; all models utilize it in over 90% of queries.

Tool Adoption Rate. Figure 4 reports per-query adoption rates of non-reference tools. *Diffusion* is the most frequently invoked tool across models, followed by *search*, while *edit* is used moderately and *code* is rarely employed. Beyond this overall trend, models display distinct strategies. Gemini 3 Pro shows a balanced distribution across *diffusion*, *search*, and *edit*, aligning with its strong planning ability. Claude Sonnet 4.5 adopts a more conservative approach with lower auxiliary tool usage. Grok-4.1 is the most tool-active and diverse, including comparatively higher code usage, reflecting a more aggressive externalization strategy. GPT-5 relies minimally on non-reference tools, while Qwen3-VL-Plus primarily favors *diffusion* and *edit*, suggesting a generation-oriented interaction pattern.

Tool Success Rate. Figure 5 shows per-call success rates by tool and model. Overall, *search* and *diffusion* are most reliable, while *code* and *edit* show larger gaps. Gemini 3 Pro maintains consistently high SR across all tools, indicating strong tool-call generation. Claude Sonnet 4.5 performs similarly well and achieves the best results on *edit*, reflecting robust editing-oriented interactions. Grok-4.1 excels on *search* and remains competitive on *diffusion* and *edit*, but is weaker on *reference*. In contrast, Qwen3-VL-Plus performs well on *diffusion* and *search* but drops significantly on *code* and *edit*, suggesting difficulty in generating correct tool calls in code- or edit-intensive scenarios.

Ground Truth Based Evaluation. To assess tool-invocation accuracy and completeness, we report Precision, Recall, and F1-score by comparing inferred and ground truth tool sets. This metric accounts for the fact that while exact execution traces may diverge, the core set of tools necessary to resolve a query should remain consistent. As shown in Table 6, strong models indicated by our

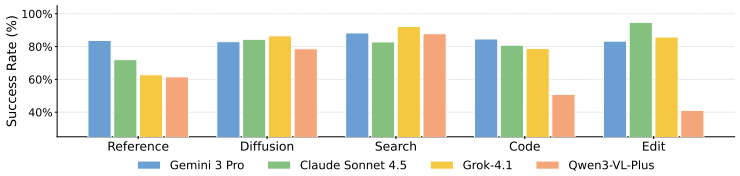


Fig. 5: Tool Success Rates per tool call for representative models on each tool.

Table 6: Tool set precision, recall and F1-score evaluated by ground truth.

Model	Precision	Recall	F1-score
Claude Sonnet 4.5	89.64	78.15	80.49
Claude Sonnet 4	88.05	79.95	80.71
Gemini 3 Pro	89.30	79.63	**81.21**
Grok-4.1	85.86	**81.69**	80.68
GPT-5	**91.42**	69.81	76.37
GPT-4o	80.16	61.31	67.12
Qwen3-VL-Plus	85.93	73.31	75.58
Qwen2.5-VL-72B	88.87	72.28	76.53
InternVL3.5-14B	5.00	3.82	4.13
LLaMA-3.2-11B	12.42	13.58	11.98

Table 7: Few-shot experiments on three models.

Method	FS↑	SR↑	MI↓
GPT-4o	60.35	63.05	1.39
GPT-4o + 3 Shots	**73.19**	**82.01**	**0.83**
Qwen2.5	53.88	39.06	1.64
Qwen2.5+ 3 Shots	**72.86**	**72.88**	**0.65**
LLaMA-3.2	29.60	25.35	2.10
LLaMA-3.2 + 3 Shots	**30.20**	**26.62**	**2.01**

MAM system (e.g., Gemini 3 Pro, Claude, Grok-4.1) achieve F1-scores above 80%. We further compare the F1-scores ranking with our Final Score and find a high Spearman correlation ($\rho = 0.879$, $p < 0.001$), indicating close agreement between tool set matching and MAM's multi-judge consensus.

6 Ablation Study

In this section, we present additional experimental results on a subset of our dataset. By default, we sample 800 queries, with 100 queries from each category.

Impact of In-Context Tool Demonstrations. Table 7 shows that adding in-context demonstrations substantially benefits models with sufficient instruction-following ability. GPT-4o sees notable gains in final score, success rate, and fewer missed images, reflecting better tool calibration. Qwen2.5-VL-72B improves even more in success rate, indicating more accurate tool calls and adherence to formats. LLaMA-3.2-11B, however, shows only marginal changes, implying few-shot prompting cannot offset its core tool-planning and execution limits.

Impact of Tool Capability Boundaries. We identify an ambiguous region where images can be produced by either *search* or *diffusion*, such as "La Tour Eiffel in rain." To quantify how models resolve this ambiguity, we evaluate three prompt variants: Original, Search Enhanced, and Diffusion Enhanced. As shown in Figure 6, Claude Sonnet 4.5, Grok-4.1, and Gemini 3 Pro shift toward *search* under Search Enhanced prompts and strongly adopt *diffusion* under Diffusion Enhanced prompts, while GPT-5 remains conservative. This indicates that tool selection is highly sensitive to capability definitions, supporting agents' adaptation to evolving tool boundaries.

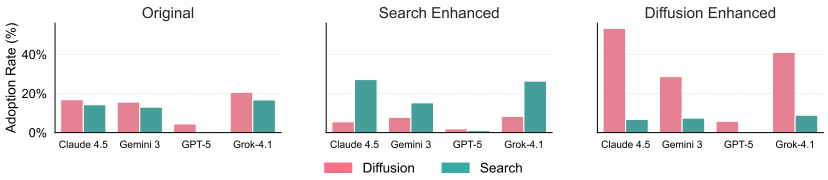


Fig. 6: Impact of tool boundary on Tool Adoption Rate. Claude 4.5 and Gemini 3 denote Claude Sonnet 4.5 and Gemini 3 Pro, respectively.

Table 8: Impact of the judge model and inter-judge rank correlation. Claude, Gemini, and GPT denote Claude Sonnet 4.5, Gemini 2.5 Pro, and GPT-5, respectively.

	Claude			Gemini			GPT		
	FS↑	SR↑	MI↓	FS↑	SR↑	MI↓	FS↑	SR↑	MI↓
Claude Sonnet 4	67.49	84.48	2.20	70.24	77.31	1.10	72.19	89.75	1.49
Gemini 3 Pro	**71.33**	**84.65**	**1.85**	**78.73**	**80.95**	**0.55**	72.42	**89.84**	1.48
Grok-4.1	67.69	81.28	**1.85**	68.23	71.55	1.07	**74.99**	89.50	**1.32**
GPT-4o	53.06	69.87	2.65	60.35	63.05	1.39	68.37	83.41	1.50
Qwen3-VL-Plus	57.35	75.74	3.07	63.5	65.69	1.35	68.37	83.37	1.70
Qwen2.5-VL-72B	46.41	51.44	3.36	53.88	39.06	1.64	62.68	61.56	1.79
InternVL-3.5-14B	47.22	6.83	2.84	50.14	11.46	1.63	63.13	68.30	1.79
LLaMA-3.2-11B	23.69	25.47	4.12	29.6	25.35	2.10	57.03	66.72	1.86

(a) Impact of the judge model.

		Claude	Gemini	GPT
FS	**Claude**	–	0.952	0.970
	Gemini	0.952	–	0.898
	GPT	0.970	0.898	–
SR	**Claude**	–	1.000	0.881
	Gemini	1.000	–	0.881
	GPT	0.881	0.881	–
MI	**Claude**	–	0.922	0.952
	Gemini	0.922	–	0.946
	GPT	0.952	0.946	–

(b) Spearman ρ ($p < 0.01$).

Impact of Judge Model. Table 8 compares results under Claude Sonnet 4.5, Gemini 2.5 Pro, and GPT-5 as judges. The overall ranking remains largely consistent: top models stay strong across judges, while mid-tier and weaker systems remain lower, indicating stable cross-judge trends. This is corroborated by high inter-judge Spearman rank correlations as shown in Table 8b. Judgment differences primarily stem from score calibration rather than rank reversal. Specifically, Claude Sonnet 4.5 assigns lower scores and more missed images to weaker models; Gemini 2.5 Pro better separates systems and yields fewer missed images; and GPT-5 gives higher scores and fewer missed images to weaker models.

Human Agreement Study. We sampled 400 queries from all models and asked the annotators to examine the evaluation reports of three agents, rating agreement on a 0–2 scale. The Precision Inspector, Recall Inspector, and Chief Judge achieved agreement rates of 84.00%, 85.88%, and 88.00%, respectively, indicating high consistency between human and MLLM judges.

Human Evaluation on End-to-End Execution. Table 9 reports human evaluation results on end-to-end interleaved generation. For each model, we sample 100 tool plans and execute all tool calls using implemented backends: in-context images for *Reference*, nano-banana [14] for *Diffusion*, Google Image Search via Serp API [28] for *Search*, GPT-5 [25] generated Python scripts for *Code*, and Doubao Seedream 4.0 [27] for *Edit*. We evaluate model performance using inappropriate images (# Inappr.), missed images (MI), and a five-point final score (FS). Results show that Gemini 3 Pro achieves the best overall perfor-

Table 9: Human evaluation on end-to-end execution. Note that fewer # *Inappr.* images do not imply a higher Success Rate, as malformed tool calls fail to generate any output.

Model	# Inappr.	MI ↓	FS ↑
Claude Sonnet 4.5	1.02	0.26	3.60
Claude Sonnet 4	1.59	0.34	3.35
Gemini 3 Pro	0.82	**0.06**	**3.89**
Grok-4.1	1.17	0.10	3.53
GPT-5	0.78	0.39	3.63
GPT-4o	0.54	0.88	3.24
Qwen3-VL-Plus	0.70	0.77	3.01
Qwen2.5-VL-72B	0.40	0.48	2.14
InternVL3.5-14B	0.08	1.37	1.26
LLaMA-3.2-11B	0.27	1.16	1.29

mance, followed by GPT-5 and Claude Sonnet 4.5. In contrast, InternVL3.5-14B and LLaMA-3.2-11B show low inappropriate image counts but much lower FS, mainly due to tool-call failures. MAM also strongly correlates with human evaluation, with Spearman rank correlations $\rho = 0.8909$ for FS ($p < 0.001$) and $\rho = 0.8303$ for MI ($p < 0.01$), supporting its reliability as an automatic judge.

7 Discussion

This work has several limitations, which we leave for future work. (1) Our interleaved generation setting focuses on text–image outputs and does not cover richer modalities such as audio or video. (2) Our toolkit is restricted to five tools and therefore does not capture broader agentic capabilities. (3) We focus on direct evaluation using MLLM-as-a-judge, and do not study alternative pipelines such as captioning images for LLM-as-a-judge.

8 Conclusion

In this work, we identify *Agentic Tool Planning* as a key next step for interleaved generation, where MLLMs autonomously coordinate tools to produce interleaved tool plans. We introduce *ATP-Bench*, the first benchmark that unifies hybrid image sourcing and dual query types under expert-annotated visual-critical intents. We further propose MAM, a Multi-Agent MLLM-as-a-Judge framework that disentangles tool-call precision, missed images, and response quality without requiring ground truth or end-to-end execution. Experiments on 10 state-of-the-art MLLMs show that current systems still struggle with coherent tool planning, and display notable differences in tool-use behavior.

Acknowledgements

This work was supported by Qwen Business Unit through Alibaba Research Intern Program. We thank Alibaba Group for providing research resources and annotation support, and the annotators for their valuable assistance.

References

1. An, J., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, L., Luo, J.: Openleaf: Open-domain interleaved image-text generation and evaluation. arXiv preprint arXiv:2310.07749 (2023)
2. Anthropic: Introducing claude 4 (2025), https://www.anthropic.com/news/claude-4, accessed: 2026
3. Anthropic: Introducing claude sonnet 4.5 (2025), https://www.anthropic.com/news/claude-sonnet-4-5, accessed: 2026
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923, accessed: 2026
6. Chen, D., Chen, R., Pu, S., Liu, Z., Wu, Y., Chen, C., Liu, B., Huang, Y., Wan, Y., Zhou, P., et al.: Interleaved scene graphs for interleaved text-and-image generation assessment. arXiv preprint arXiv:2411.17188 (2024)
7. Chen, D., Chen, R., Zhang, S., Wang, Y., Liu, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P., Sun, L.: Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In: Forty-first International Conference on Machine Learning (2024)
8. Chern, E., Su, J., Ma, Y., Liu, P.: Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. arXiv preprint arXiv:2407.06135 (2024)
9. Chow, W., Pan, J., Liang, Y., Zhou, M., Song, X., Jia, L., Zhang, S., Tang, S., Li, J., Zhang, F., et al.: Weave: Unleashing and benchmarking the in-context interleaved comprehension and generation. arXiv preprint arXiv:2511.11434 (2025)
10. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
11. Feng, Y., Sun, J., Li, C., Li, Z., Ai, J., Zhang, F., Chang, Y., Zhou, S., Zhang, S., Dai, Y., et al.: A high-quality dataset and reliable evaluation for interleaved image-text generation. arXiv preprint arXiv:2506.09427 (2025)
12. Gao, D., Ji, L., Zhou, L., Lin, K.Q., Chen, J., Fan, Z., Shou, M.Z.: Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. arXiv preprint arXiv:2306.08640 (2023)
13. Gao, Z., Du, Y., Zhang, X., Ma, X., Han, W., Zhu, S.C., Li, Q.: Clova: A closed-loop visual assistant with tool usage and update. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13258–13268 (2024)
14. Google: Introducing gemini 2.5 flash image (2025), https://developers.googleblog.com/introducing-gemini-2-5-flash-image, accessed: 2026
15. Google: A new era of intelligence with gemini 3 (2025), https://blog.google/products/gemini/gemini-3, accessed: 2026
16. Guo, Z., Zhang, F., Jia, K., Jin, T.: Llm-i: Llms are naturally interleaved multimodal creators. arXiv preprint arXiv:2509.13642 (2025)

17. Hu, H., Luan, Y., Chen, Y., Khandelwal, U., Joshi, M., Lee, K., Toutanova, K., Chang, M.W.: Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12065–12075 (2023)
18. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
19. Kou, S., Jin, J., Liu, Z., Liu, C., Ma, Y., Jia, J., Chen, Q., Jiang, P., Deng, Z.: Orthus: Autoregressive interleaved image-text generation with modality-specific heads. arXiv preprint arXiv:2412.00127 (2024)
20. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024)
21. Liu, M., Xu, Z., Lin, Z., Ashby, T., Rimchala, J., Zhang, J., Huang, L.: Holistic evaluation for interleaved text-and-image generation. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 22002–22016 (2024)
22. Liu, S., Cheng, H., Liu, H., Zhang, H., Li, F., Ren, T., Zou, X., Yang, J., Su, H., Zhu, J., et al.: Llava-plus: Learning to use tools for creating multimodal agents. In: European conference on computer vision. pp. 126–142. Springer (2024)
23. Mayer, R.E.: Multimedia learning. In: Psychology of learning and motivation, vol. 41, pp. 85–139. Elsevier (2002)
24. Meta: Llama 3.2: Revolutionizing edge ai and vision with open, customizable models (2025), `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`, accessed: 2026
25. OpenAI: Introducing gpt-5 (2025), `https://openai.com/index/introducing-gpt-5`, accessed: 2026
26. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
27. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025)
28. SerpAPI: Serpapi: Google search api (2026), `https://serpapi.com/`, accessed: 2026
29. Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Emu: Generative pretraining in multimodality. arXiv preprint arXiv:2307.05222 (2023)
30. Surís, D., Menon, S., Vondrick, C.: Vipergpt: Visual inference via python execution for reasoning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11888–11898 (2023)
31. Taneja, K., Singh, A., Goel, A.K.: Towards a multimodal document-grounded conversational ai system for education. In: International Conference on Artificial Intelligence in Education. pp. 92–99. Springer (2025)
32. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
33. Wang, C., Luo, W., Dong, S., Xuan, X., Li, Z., Ma, L., Gao, S.: Mllm-tool: A multimodal large language model for tool agent learning. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6678–6687. IEEE (2025)

34. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
35. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023)
36. xAI: Grok 4.1 (2025), `https://x.ai/news/grok-4-1`, accessed: 2026
37. Xia, P., Han, S., Qiu, S., Zhou, Y., Wang, Z., Zheng, W., Chen, Z., Cui, C., Ding, M., Li, L., et al.: Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models. arXiv preprint arXiv:2410.10139 (2024)
38. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024)
39. Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., Wang, L.: Mm-react: Prompting chatgpt for multimodal reasoning and action. arXiv preprint arXiv:2303.11381 (2023)
40. Yu, Q., Xiao, Z., Li, B., Wang, Z., Chen, C., Zhang, W.: Mramg-bench: a comprehensive benchmark for advancing multimodal retrieval-augmented multimodal generation. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3616–3626 (2025)
41. Zhang, R., Huang, Y., Lu, C., Wang, Q., Gao, Y., Wu, Y., Hu, Y., Xu, Y., Wang, W., Wang, H., et al.: Rag-igbench: Innovative evaluation for rag-based interleaved generation in open-domain question answering. arXiv preprint arXiv:2512.05119 (2025)
42. Zhou, P., Peng, X., Song, J., Li, C., Xu, Z., Yang, Y., Guo, Z., Zhang, H., Lin, Y., He, Y., et al.: Opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 56–66 (2025)