

---

# RAG-IGBench: Innovative Evaluation for RAG-based Interleaved Generation in Open-domain Question Answering

---

**Rongyang Zhang<sup>1,2</sup>, Yuqing Huang<sup>1,2</sup>, Chengqiang Lu<sup>2</sup>, Qimeng Wang<sup>2</sup>, Yao Gao<sup>2</sup>,  
Yi Wu<sup>2</sup>, Yao Hu<sup>2</sup>, Yin Xu<sup>1</sup>, Wei Wang<sup>3</sup>, Hao Wang<sup>1,\*</sup>, Enhong Chen<sup>1\*</sup>**

<sup>1</sup>State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China  
<sup>2</sup>Xiaohongshu Inc.   <sup>3</sup>Xi'an Jiaotong University

## Abstract

In real-world scenarios, providing user queries with visually enhanced responses can considerably benefit understanding and memory, underscoring the great value of interleaved image-text generation. Despite recent progress, like the visual autoregressive model that unifies text and image processing in a single transformer architecture, generating high-quality interleaved content remains challenging. Moreover, evaluations of these interleaved sequences largely remain underexplored, with existing benchmarks often limited by unimodal metrics that inadequately assess the intricacies of combined image-text outputs. To address these issues, we present RAG-IGBench, a thorough benchmark designed specifically to evaluate the task of Interleaved Generation based on Retrieval-Augmented Generation (RAG-IG) in open-domain question answering. RAG-IG integrates multimodal large language models (MLLMs) with retrieval mechanisms, enabling the models to access external image-text information for generating coherent multimodal content. Distinct from previous datasets, RAG-IGBench draws on the latest publicly available content from social platforms and introduces innovative evaluation metrics that measure the quality of text and images, as well as their consistency. Through extensive experiments with state-of-the-art MLLMs (both open-source and proprietary) on RAG-IGBench, we provide an in-depth analysis examining the capabilities and limitations of these models. Additionally, we validate our evaluation metrics by demonstrating their high correlation with human assessments. Models fine-tuned on RAG-IGBench’s training set exhibit improved performance across multiple benchmarks, confirming both the quality and practical utility of our dataset. Our benchmark is available at <https://github.com/USTC-StarTeam/RAG-IGBench>.

## 1 Introduction

Multimodal content generation has gained prominence in modern information systems, particularly for interleaved image-text generation in open-domain question answering. This emerging field addresses the need for responses combining textual explanations with visual elements for more comprehensive user interactions [1, 10, 69]. The task aims to produce coherent sequences of interleaved text and illustrative images aligned with arbitrary queries [38, 22, 65, 62], supporting applications ranging from web content creation to visual storytelling [30] and chain-of-thought explanations. Unlike previous approaches [4, 75, 28, 20, 61, 25] that generated single-modality responses, this integrated approach better addresses complex real-world needs.

---

\*Corresponding author.

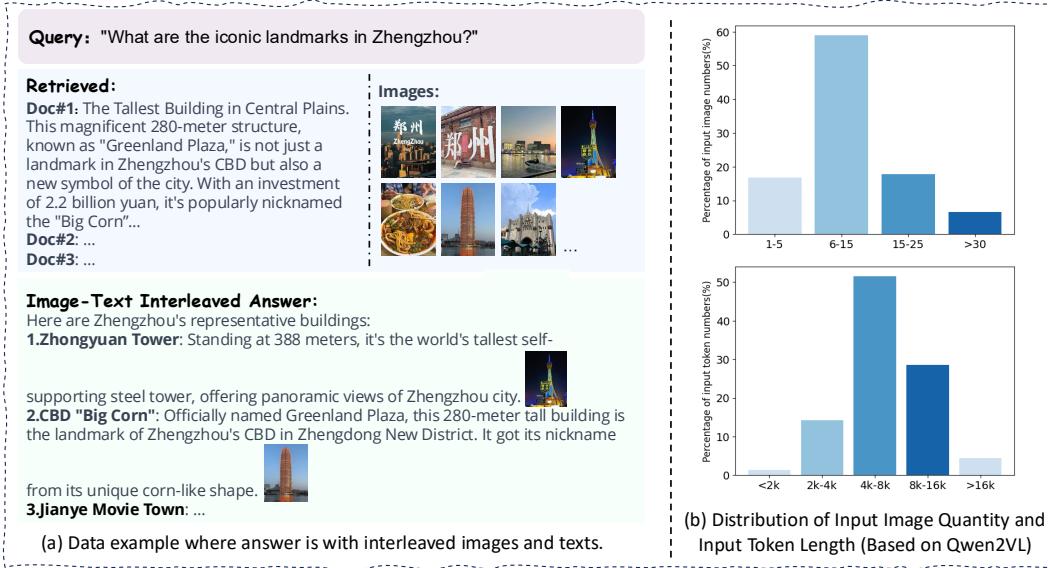


Figure 1: Data example and statistics of input.

Recent research [1] has explored combining multimodal large language models (MLLMs) with diffusion models for coherent image-text generation. However, this approach suffers from semantic inconsistencies between modalities and reduced image quality due to independent generation processes. While novel unified transformer architectures [66, 92, 70] attempt to handle image-text comprehension and generation simultaneously, they show limitations in following complex instructions, restricting their practical utility. Additionally, these transformer-based approaches require extensive datasets and computational resources, presenting further development challenges.

In addition, developing comprehensive benchmarks for interleaved image-text generation presents two significant challenges [23]. First, despite the availability of extensive open-source image-text datasets [9, 56, 91], their unlabeled nature and quality inconsistencies limit their applicability to benchmark evaluation, relegating them primarily to model pre-training. Second, current evaluation methodologies demonstrate considerable limitations [48, 83]. Conventional metrics such as Fréchet Inception Distance (FID) [27] assess only image quality while neglecting image-text integration. Although recent MLLM-based evaluation approaches [48, 83] attempt to address this gap through specialized instructions, they introduce inherent model biases, inconsistencies from stochastic variations, and require task-specific manual adaptation of evaluation criteria.

To enhance image-text content generation for open-domain applications, we propose an innovative approach based on Retrieval-Augmented Generation (RAG) [40]. As illustrated in Figure 1(a), our RAG-IG framework integrates MLLMs with retrieval mechanisms, enabling the processing of both user queries and retrieved multimodal content. The system generates markdown-formatted text with image placeholders, subsequently replaced with corresponding retrieved images to produce coherent interleaved image-text responses. This methodology significantly improves information clarity and presentation, yielding high-quality outputs that effectively integrate textual and visual elements while maintaining semantic consistency.

To address the evaluation challenges in image-text interleaved generation tasks, we propose a comprehensive benchmark framework that extends beyond current MLLM-based evaluation methods [48, 83]. Our evaluation framework systematically assesses generated content across three critical dimensions: textual quality, image quality, and image-text coherence. For textual assessment, we utilize ROUGE scores [46] to evaluate the accuracy and fluency of generated text. For image evaluation, we propose modified metrics based on edit distance [53] and Kendall correlation [50] to assess both image selection accuracy and sequential arrangement. The image-text coherence is measured through two complementary metrics: CLIP-score [26] and a novel semantic alignment score derived from embedding models. These metrics are then integrated to provide a holistic evaluation score for the generated content.

To establish a reliable benchmark, we have developed a systematic data collection and validation pipeline. The process begins with careful query selection and image-text content generation using

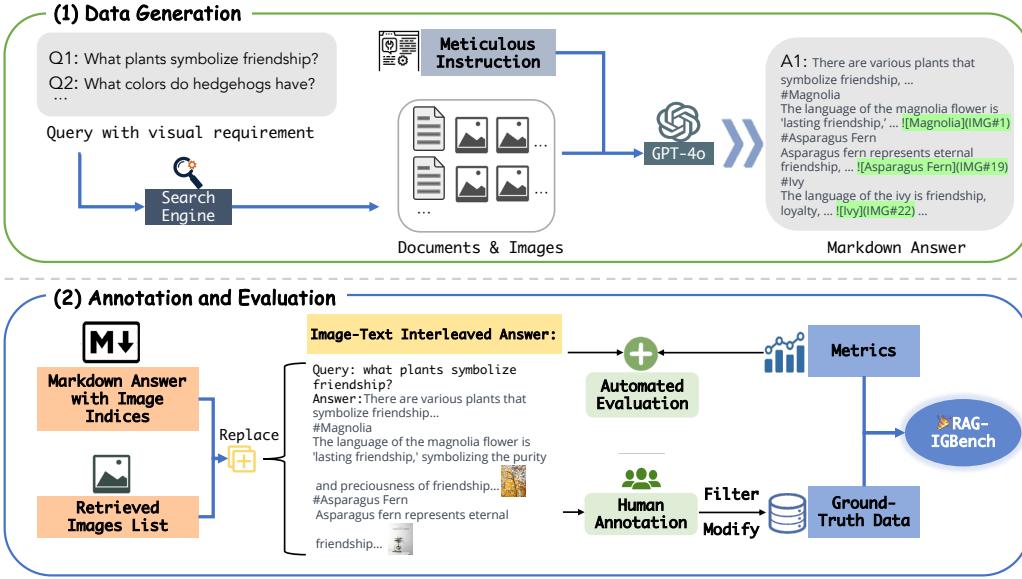


Figure 2: Overview of the pipeline of RAG-IG and the data construction of RAG-IGBench

our RAG-IG framework, followed by comprehensive human evaluation across multiple quality dimensions. We validate RAG-IG’s effectiveness from both data and metric perspectives. First, models fine-tuned on our benchmark dataset demonstrate improved performance across various multimodal evaluation tasks, confirming the high quality of our data. Second, we validate our evaluation metrics through a systematic study of 200 sampled queries, where we compare our automated metrics against human assessments. The strong correlation between automated and human evaluations, measured by both Pearson [57] and Spearman [58] coefficients, demonstrates the robustness of our evaluation framework. These results establish our benchmark as a reliable evaluation tool for interleaved image-text content.

Our main contributions can be summarized as follows: (1) We introduce an innovative method for generating interleaved image-text content based on RAG. This approach provides multi-modal answers to queries requiring images, effectively integrating text and visuals. (2) We establish a novel benchmark for interleaved image-text generation with a systematically collected dataset and validate the high quality of our collection. (3) We propose innovative automated metrics for evaluation and demonstrate their reliability through correlation experiments between automated scores and human assessment results. (4) We perform extensive experiments on our benchmark, evaluating a range of mainstream state-of-the-art MLLMs, including both open-source and proprietary models. The results provide valuable insights for future research in this field.

## 2 Related Work

**Multimodal Large Language Models for Multimodal Generation.** Recent advances in multimodal large language models (MLLMs) have substantially improved the integration of text and image modalities [47, 94, 18]. These models have evolved through pipeline-based and end-to-end approaches. Pipeline-based approaches combine diffusion models [54] for interleaved image-text generation [37] or utilize text-to-image retrieval to obtain visual information [44, 52, 85], while end-to-end approaches like Chameleon [66] and Show-o [84] employ unified transformer architectures for joint image-text understanding and generation. Though these advances have enhanced models’ capabilities in cross-modal understanding and generation, they face significant challenges in maintaining both high-quality image generation and image-text semantic alignment. RAG-based Interleaved Generation (RAG-IG) effectively addresses these limitations and requires fewer computational resources.

**Benchmarks for Multimodal Large Language Models.** Current evaluation benchmarks for multimodal large language models (MLLMs) face several critical limitations. While large-scale datasets like MINT-1T [37] and Obelics [39] serve as valuable pre-training resources, their lack of annotations and limited image-text coherence restricts their effectiveness for evaluation. Traditional benchmarks

[34], though effective for basic capability assessment, fall short in evaluating complex reasoning abilities. Recent benchmarks [43, 88, 49, 86, 82, 77, 71, 16, 33] have addressed sophisticated reasoning evaluation, yet they primarily focus on isolated tasks, overlooking the critical aspect of interleaved comprehension and generation. Although advancing efforts like [48], [83], and [93] attempt to evaluate interleaved content through model-based methods, these benchmarks are either limited by their dataset scale or restricted by the limitations of their evaluation metrics. Our benchmark addresses these limitations by providing a comprehensive evaluation framework for interleaved multimodal comprehension and generation.

**Evaluation Metrics For Multimodal Generation.** Evaluating multimodal generation for open-domain question answering presents significant challenges [1, 59, 81], which we have systematically categorized into three key aspects. First, single-modality metrics (e.g., BLEU[51], ROUGE[46] for text and FID [27], IS [55] for images) fail to capture the interdependence between different modalities. Second, cross-modal alignment metrics like CLIPScore [26] and X-IQE [13] evaluate image-text correspondence but overlook the intrinsic quality of individual modalities. Third, recent model-based approaches face significant limitations: GPT-based scoring [48, 89] and fine-tuned models scoring [83] introduce model-specific biases and suffer from evaluation instability [6, 78], and Interleaved Area approach [93] encounters considerable computational expense and limited scalability. Moreover, these frameworks require extensive task-specific adaptations, leading to inconsistent evaluation standards across different tasks and substantial manual effort in implementation.

### 3 RAG-IGBench

In this paper, we propose an approach for interleaved image-text generation based on Retrieval-Augmented Generation (RAG) and establish a comprehensive benchmark for systematic evaluation. The following sections first present the formal task definition for RAG-based interleaved generation, followed by a detailed description of our RAG-IGBench, including data format, statistics, construction methodology, and evaluation metrics.

#### 3.1 Task Definition

The task of interleaved image-text generation aims to produce responses that effectively combine textual and visual information for a given query. We adopt the RAG framework [40] to enhance the coherence and quality of the generated content. As illustrated in Figure 2, given a query  $q$ , we begin by retrieving a set of relevant documents  $D = \{d_1, d_2, \dots, d_n\}$  and their corresponding images  $I = \{I_1, I_2, \dots, I_n\}$  where  $n$  is the number of retrieved documents and  $I_i$  is the image set of the  $i$ -th document. Then, the retrieved content, in conjunction with the query and our elaborate instruction, is utilized as input for multi-modal large language models (MLLMs). Each image in the input is distinctly labeled with an index number. The MLLM generates a response in markdown format, where images are represented as ![image description](IMG#k), with k indicating the image index. We then replace these indices with corresponding image URLs to produce the final multimodal response.

However, implementing the RAG-IG paradigm with multiple images poses significant challenges: large image sets expand the input context window substantially, straining MLLM performance and computational resources [63, 29]. Therefore, we retrieve a maximum of three documents per query. Additionally, we exclude the queries that can be adequately addressed through text-only responses.

#### 3.2 Dataset Format and Statistics

Each sample in our dataset is structured as a tuple  $(q, gt, D, I)$ , where  $q$  represents the original user query,  $gt$  denotes the ground-truth answer in markdown format incorporating both textual and visual elements,  $D$  represents the retrieved contextual documents and  $I$  represents their corresponding images. As shown in Table 1, RAG-IGBench consists of 6,057 curated samples where each query is selected based on explicit visual requirements.

Table 1: Dataset statistics. The token statistics are calculated based on the tokenizer from Qwen2VL.

Statistic	Number	Percentage
Total queries	6,057	-
Questions with images	6,057	100%
Questions with answer label	6,057	100%
Avg./Max. Retrieved images	12.96/51	-
Avg./Max. Generated images	2.26/10	-
Avg./Max. Input tokens	7,914.22/39,959	-
Avg./Max. Answer tokens	291.40/2,967	-

Table 2: Comparisons between Benchmark and existing open-sourced multi-modal evaluation benchmarks.

Dataset Name	Multi-Images Input	Interleaved Input	Interleaved Generation	Average Images	Metric
MIRB [90]	✓	✗	✗	3.8	ACC
MEGA-Bench [11]	✓	✗	✗	2.0	45 metrics
Qbench2 [80]	✓	✗	✗	2.0	ACC
NLVR2 [64]	✓	✗	✗	2.0	ACC
BLINK [24]	✓	✗	✗	1.9	ACC
MuirBench [73]	✓	✓	✗	4.3	ACC
SEED-Bench-2 [42]	✓	✓	✗	1.5	ACC
VL-ICL Bench [95]	✓	✓	✗	1.3	Acc & MLLM
MMMU [87]	✓	✓	✗	1.2	ACC
INTERLEAVEDBENCH[48]	✓	✓	✓	1.7	GPT-4o based
MMIE [83]	✓	✓	✓	8.5	Fine-tuned VLM based
RAG-IGBench (Ours)	✓	✓	✓	<b>13.0</b>	Multi-dimensional metrics

### 3.3 Dataset Construction

The process of constructing benchmark data comprises three key stages, each of which plays a crucial role in ensuring the high quality of the final dataset.

In the first stage, we guide MLLMs (including GPT4o[35], Claude-3.5-Sonnet[3], et al.) with meticulously crafted instructions to generate the raw question-answer data. This stage takes a query, retrieved documents, and relevant images as input into an MLLM, resulting in a Markdown-formatted answer that includes image indices. To improve answer correctness and ensure consistency between text and images, we took inspiration from the Chain of Thought (CoT) [79] and the ICL approach [21], which begins by instructing the model to analyze the input query, categorizing it into one of four types: "what-is", "how-to", "yes-or-no", and "head-to-head". The main reason for categorizing queries is that different types of queries have varying requirements for visual information. Please refer to the appendix B and F for detailed information and the specific generation prompt.

In the second stage, we annotate first-stage results through a rigorous process that is both time-consuming and labor-intensive. We first filter out cases lacking image indices in the answers. Annotators then evaluate the remaining data based on text quality, image quality, image-text consistency, and overall quality. To establish high-quality ground-truth answers, we conduct thorough image selection refinement, manually adding, removing, or replacing images in answers according to the candidate images. All annotators are multimodal assessment experts with high inter-rater reliability scores [ $\kappa > 0.85$ ], ensuring precise ground-truth answers.

In the third stage, we filter the QA pairs based on the annotation results of the second stage. QA pairs with low scores in any dimension are excluded from the dataset. Besides, we retain cases that have been successfully improved through manual refinement. This systematic filtering process results in our final benchmark dataset.

Through this three-stage process, we constructed our final RAG-IGBench . Answers are stored in markdown format with image indices rather than direct URLs, allowing for systematic evaluation and easy conversion to complete multimodal responses by mapping indices to their corresponding image URLs. It should be noted that the data was systematically reviewed for privacy concerns and any content containing potentially identifiable personal information was excluded.

### 3.4 Evaluation Metric

This section presents our evaluation metrics for multimodal responses. Following the annotation criteria established in 3.3, we evaluate responses across three dimensions: text quality, image quality, and image-text alignment. Unlike previous works that rely on MLLM-based scoring methods such as INTERLEAVEDBENCH [48] and MMIE [83], we implement rule-based metrics leveraging ground-truth references, which ensures unbiased and reproducible evaluation results. The following subsections detail our specific metrics for each dimension.

**Text Quality:** In the field of natural language processing (NLP), there are a variety of effective metrics to evaluate the quality of the generated text, including BLEU [51], ROUGE [46], and perplexity. In RAG-IGBench , we focus on measuring the semantic alignment between the generated text and the ground truth answers. We adopt the ROUGE-1 score as our evaluation metric due to its

effectiveness in capturing lexical overlap and semantic similarity between text passages. Although we also evaluated the ROUGE-2 and ROUGE-L metrics, empirical testing showed that ROUGE-1 provides better discriminative power to distinguish the quality of generated responses.

**Image Quality:** Previous image quality assessment studies have relied on metrics like the Fréchet Inception Distance (FID) [27] and the Inception Score (IS) [7], which measure statistical similarity between the generated and real distributions of characteristic features of the image. However, our RAG-IG framework selects images rather than generates them, making these traditional metrics inapplicable. Since both generated and ground-truth answers contain images selected from the same candidate pool, our evaluation focuses on two aspects: the selection accuracy and the ordering correctness of images. We formulate this as an ordered list comparison problem, employing Edit Distance to measure selection accuracy and Kendall Score to evaluate ordering consistency.

Specifically, let  $I_{generated} = \{img_1, img_2, \dots, img_m\}$  and  $I_{gt} = \{img_1, img_2, \dots, img_n\}$  where  $I_{generated}$  and  $I_{gt}$  are lists composed of images from the generated answer and ground truth, respectively, and  $m$  and  $n$  represent the lengths of two lists.

For the Edit Distance, we measure the similarity between the generated image sequence  $I_{generated}$  and  $I_{gt}$  by calculating the minimum number of operations (insertion, deletion, and substitution) required to transform one into the other. This computation can be efficiently implemented using dynamic programming. To handle varying sequence lengths, we normalize the score as:

$$\text{Edit Distance} = 1 - \frac{dp(m, n)}{\max(m, n)} \quad (1)$$

where  $dp$  is the dynamic programming matrix. This normalization ensures comparable scores across sequences of different lengths, with higher values indicating greater similarity.

For the Kendall Score, inspired by the Kendall rank correlation coefficient [50], commonly known as Kendall’s  $\tau$  coefficient, we calculate it by counting the proportion of concordant pairs between the two lists relative to all correct pairs in  $I_{generated}$ . To elaborate, let  $I_{correct} = \{img_1, img_2, \dots, img_o\}$  be the intersection of two image lists,  $I_{generated}$  and  $I_{gt}$ , where  $o$  denotes the length of the intersection, and the images in  $I_{correct}$  follow the order in  $I_{generated}$ . Then define  $\text{index}(img_k, I_{gt})$  refer to the index of the image  $img_k$  within  $I_{gt}$ . For each possible pair  $(img_i, img_j)$  in  $I_{correct}$ , where  $i < j \leq o$ , determine if it is a concordant pair by checking if  $\text{index}(img_i, I_{gt}) < \text{index}(img_j, I_{gt})$ . At last, we count the ratio of the number of concordant pairs to the total number of possible pairs formed by elements in  $I_{correct}$  as the final Kendall Score. Therefore, we can derive the following formula for the Kendall Score:

$$\text{Kendall Score} = \begin{cases} \frac{\sum_{1 \leq i < j \leq o} \mathbb{I}(\text{index}(img_i, I_{gt}) < \text{index}(img_j, I_{gt}))}{\frac{1}{2}o(o-1)}, & o > 1 \\ \frac{o}{\max(m, n)}, & o \leq 1 \end{cases} \quad (2)$$

This dual-metric enables a comprehensive evaluation that considers both the selection of appropriate images and their proper arrangement in the response.

**Image-Text Consistency:** In our RAG-IGBench , we first employ CLIP Score [26] to assess image-text consistency by measuring cosine similarity in a shared vector space. However, it has notable limitations: it struggles with complex semantics, inherits biases from the pre-training data, and captures limited context for individual images. To overcome these limitations, we propose the Alignment Score metric, which evaluates textual similarity between the context surrounding the correct image in the generated answer and its corresponding context in the ground-truth answer. This approach better captures the contextual relationship in interleaved multimodal content. It can be formulated as follows:

$$\text{Alignment Score} = \frac{\sum_{i=1}^o \text{Sim}(C_{img_i}, G_{img_i})}{o} \quad (3)$$

Here  $img_i \in I_{correct}$ , where  $C_{img_i}$  denotes the context of  $img_i$  in generated answer and  $G_{img_i}$  represents the context of  $img_i$  in ground-truth answer. This approach is predicated on the notion that the same image should be embedded in similar contextual settings. We evaluate image-text

alignment by comparing both the selected images and their contextual usage with the ground-truth answer. Strong alignment between the generated and ground-truth answers typically indicates proper integration of images within their textual context.

The CLIP Score and Alignment Score provide complementary measures for evaluating image-text consistency, with the former assessing direct semantic alignment and the latter measuring contextual similarity between image usage patterns.

### 3.5 Comparison With Other Multi-modal Benchmarks

As illustrated in the table 2, RAG-IGBench distinguishes itself from existing multimodal benchmarks in several significant ways: **(1) Multi-modal Multi-Image Input:** RAG-IGBench requires MLLMs to process multiple images alongside text simultaneously, presenting a more complex challenge than existing benchmarks that typically handle single-modality input or limited image counts; **(2) Open-Domain Interleaved Generation:** Unlike INTERLEAVEDBENCH [48] and MMIE [83], which focus on specific tasks like step-by-step instructions, our benchmark is designed for open-domain queries requiring interleaved image-text responses, enabling comprehensive and visually enriched answers across various domains; **(3) Novel Comprehensive Evaluation Metrics:** We introduce multi-dimensional evaluation metrics specifically designed for interleaved generation tasks, addressing three key dimensions: text quality, image quality, and image-text consistency (detailed in 3.4). Our RAG-IGBench overcomes the limitations of existing metrics that either lack task specificity or rely on MLLM-based evaluation, which can introduce potential bias and instability.

Table 3: Evaluation results of mainstream MLLMs on RAG-IGBench. Bold and underlined entries represent the best performance in proprietary and open-source models, respectively.

Models	Text Quality		Image Quality		Image-Text Consistency		Mean↑
	Rouge-1↑	Edit Distance↑	Kendall Score↑	Alignment Score↑	CLIP Score↑		
Proprietary MLLMs							
GPT4o	<b>57.42</b>	<b>51.28</b>	<b>46.50</b>	<b>38.81</b>	<b>36.04</b>	<b>46.01</b>	
Claude3.5-sonnet	35.98	29.98	21.91	30.68	35.56	30.81	
Gemini-1.5-pro	46.35	42.22	34.57	35.07	34.85	38.61	
QwenVL-Max	49.24	44.66	38.02	34.55	38.28	40.95	
Open-source MLLMs							
Qwen2VL-7B	43.21	22.23	18.93	18.77	27.84	26.20	
Qwen2VL-72B	49.49	<u>36.66</u>	<u>31.40</u>	26.49	32.69	35.34	
Llava Onevision 72B	42.89	24.77	<u>19.66</u>	18.20	27.19	26.54	
InternVL2.5 8B	41.98	24.53	19.20	21.94	28.53	27.24	
InternVL2.5 78B	<u>50.71</u>	36.86	27.10	<u>35.68</u>	33.23	36.71	
NVLM-D-72B	38.43	13.97	11.87	<u>8.57</u>	13.82	17.33	
InternVL2-Llama3-76B	43.91	25.18	18.15	25.38	27.66	28.08	
Qwen2.5VL-7B	44.79	24.56	19.75	19.87	25.45	26.88	
Qwen2.5VL-72B	43.18	<u>40.38</u>	30.12	35.26	<u>40.03</u>	37.79	

## 4 Experiment

This section describes our experimental setup and presents the evaluation results of mainstream MLLMs on our RAG-IGBench , followed by a detailed analysis of model behaviors.

### 4.1 Experiment Setup

**Evaluation Models.** In this study, we evaluated the mainstream state-of-the-art MLLMs on the RAG-IGBench including four closed-source models (GPT-4o [35], Claude 3.5-Sonet [3], Gemini-1.5 [67], QwenVL-Max [5]) and six open-source models (Qwen2vl-7B/72B [76], Llava OneVision 72B [41], InternVL2.5 8B/78B [14], NVLM-D-72B [17], InternVL2-Llama3-76B [15], Qwen2.5VL 7B/72B [68]). The RAG-IG framework requires models to process multimodal inputs and multiple images simultaneously, which excludes certain models like LlaVA-1.6 [47] despite their strong image-text understanding capabilities. Furthermore, the evaluated models must strictly follow instructions to generate valid markdown results that can be transformed into image-text interleaved answers. Due to

Table 4: Correlation results with Human Evaluation of Image Quality, Image-Text Consistency and Overall Assessment.

Metric	Image Quality		Image-Text Consistency		Overall Assessment	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
FID	-0.223	-0.217	-	-	-	-
IS	-0.174	-0.139	-	-	-	-
CLIP-Score	-	-	-0.108	-0.094	-	-
GPT4o based	0.097	0.112	0.151	0.104	0.017	0.028
RAG-IGBench	<b>0.256</b>	<b>0.244</b>	<b>0.172</b>	<b>0.145</b>	<b>0.143</b>	<b>0.152</b>

these requirements, models with limited parameters and relatively weaker image-text understanding capabilities were not included in our evaluation. All models employed greedy decoding for answer generation, and all evaluations were conducted on a single machine equipped with 8 H800 GPUs.

**Implementation Details.** We utilize CLIP-ViT-Large-Patch14 [26] from OpenAI for CLIP Score calculation and Conan-embed-ding [45] for Alignment Score calculation. The choice of Conan-embedding is based on its superior performance in text embedding tasks. Besides, all images were constrained to a uniform width of 540 pixels for the main experiments. To facilitate a more systematic comparison across different models, we normalized both metrics to a standardized scale.

## 4.2 Main Results

We present our primary evaluation results on RAG-IG in Table 3, with proprietary models in the upper portion and open-source models in the lower portion. Our analysis reveals several key findings: (1) In terms of overall performance, GPT4o leads among all models, followed by Gemini-1.5-pro. Among open-source models, Qwen2.5VL-72B achieves the highest performance, closely followed by InternVL2.5 78B. Claude-3.5-sonnet demonstrates relatively lower performance, primarily due to incorrect images in its responses. (2) The performance gap between proprietary and open-source models varies significantly across different metrics. The small gap in CLIP Score between proprietary and open-source models suggests that current MLLMs have reached comparable proficiency in image-to-text tasks. However, larger performance disparities appear in the image quality dimension, which requires image-to-text matching capabilities. We attribute these significant differences to the limited exploration of such image-to-text matching tasks in previous multimodal research.

Analysis of Table 3 reveals a substantial performance gap between Qwen2VL-7B and Qwen2VL-72B. This disparity is not only in text quality but also in image quality and image-text alignment, demonstrating the significance of scaling up MLLM architectures.

Additionally, among open-source models, those beyond the Qwen-2VL series and InternVL2.5 demonstrate notably inferior performance. We attribute the superior performance of Qwen2VL and InternVL2.5 to their specialized image processing approaches: Qwen2-VL employs mROPE for adaptive resolution handling, while InternVL2.5 implements a dynamic resolution strategy, processing input images into tiles of  $448 \times 448$  pixels. This represents a promising direction for future MLLM advancement.

## 4.3 Analysis

In this subsection, we conduct a detailed analysis of the results, centering on four key research questions related to RAG-IGBench:

**RQ1.** How is the consistency of our metrics with human evaluation?

**RQ2.** How is the data quality of RAG-IGBench?

Benchmark	Qwen2VL-7B	Qwen2VL-7B-sft	Variation
RAG-IGBench(Dev)	26.12	36.10	<b>+9.98</b>
MMB [49]	81.62	82.04	<b>+0.42</b>
BLINK [24]	53.34	54.50	<b>+1.16</b>
Mantis-Eval [36]	49.06	57.55	<b>+8.49</b>
Q-Bench2 [80]	61.80	68.60	<b>+6.80</b>
NLVR2 [64]	84.01	80.56	<b>-3.45</b>
Avg.	59.33	63.23	<b>+3.90</b>

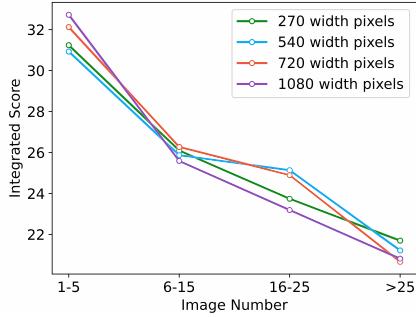


Figure 3: Ablation experiment result of different resolutions and amounts of input images.

**RQ3.** What challenges encountered by MLLMs in RAG-IG?

**RQ4.** How does the amount and resolution of input images influence the performance of MLLMs?

**The consistency of our metrics with human evaluation.** We evaluated our RAG-IGBench metrics through correlation studies with 200 randomly sampled cases. Our metrics assess image quality (Edit Distance and Kendall Score), image-text consistency (Alignment Score and CLIP Score), and overall assessment. We computed Pearson and Spearman correlations between these metrics and human evaluations, comparing them with conventional metrics (FID, IS, et al.) and GPT4o-based evaluation. Table 4 shows that our metrics outperformed conventional metrics and GPT4o-based evaluation in terms of correlation with human evaluation.

**The data quality of RAG-IGBench.** To illustrate the high quality of our dataset, we conducted a systematic evaluation by partitioning RAG-IGBench into training and test sets and subsequently fine-tuning the Qwen2VL series model on the training set. The experimental results, as shown in the table 5, demonstrate not only significant performance improvements on the test set but also notable gains across various existing multi-modal benchmarks. Particularly noteworthy is the enhanced performance on benchmarks that require multi-image comprehension capabilities, such as BLINK [24]. These comprehensive improvements across different evaluation metrics underscore the quality of our dataset.

**Challenges encountered by MLLMs in RAG-IG.** RAG-IG poses substantial challenges to MLLMs, particularly in their ability to comprehend and process interleaved image-text inputs containing multiple images. Through systematic error analysis, we find that not all queries produce valid outputs, and we have categorized these failure cases into three distinct categories: (1) Hallucination: Our experiments reveal that models frequently hallucinate image index numbers, generating indices that exceed the number of retrieved images (see Figure 6). (2) Instruction following ability: Despite explicit formatting instructions in prompt, models often generate syntactically incorrect Markdown structures. (3) Length of context windows: The combination of multiple images and retrieved texts results in extensive context lengths, thereby impacting model performance and memory efficiency. Consequently, we have to exclude samples with excessive images and restrict the maximum number of retrieved documents to three. Table 6 presents statistics on hallucination and format errors across model variants. We focus on open-source MLLMs where these issues are prevalent, as proprietary models show minimal errors.

**Influence of the amount and resolution of input images.** We conducted systematic ablation studies on Qwen2VL-7B, varying the number and resolution of input images (Figure 3). Our quantitative observations indicate the following: (1) Model performance decreases with more input images, even at low resolutions, indicating MLLMs’ inherent limitations in processing multiple images beyond token-length constraints. (2) Higher image resolutions improve performance with 1-5 images but degrade performance with more images, suggesting a critical trade-off between detail richness and context length management.

#### 4.4 Other Text Quality Metrics

We selected ROUGE as our primary text evaluation metric due to its widespread adoption and computational efficiency. However, considering that stronger text evaluation metrics are available,

Table 6: Statistics of invalid format and hallucinations generated by MLLMs

Models	Invalid format	Hallucination
Qwen2VL-7B	5	6
Qwen2VL-72B	0	5
Llava Onevision 72B	1579	12
InternVL2.5 78B	0	14
NVLM-D-72B	45	39
InternVL2-Llama3-76B	44	94

Table 7: Comparison of different text quality metrics.

Models	ROUGE	CIDEr	CIDEr-D	SPICE	CLAIR
GPT-4o	56.58	2.458	0.810	0.0998	88.461
Claude-3.5-sonnet	40.82	1.199	0.026	0.0584	87.940
Gemini-1.5-pro	45.59	1.827	0.364	0.0575	88.113
Qwen2VL-72B	50.92	2.339	0.549	0.0788	87.841
Qwen2VL-7B	44.70	1.990	0.307	0.0880	84.466

such as CLAIR [8] and SPICE [2], we also report the performance of several mainstream models under these metrics in Table 7 for comprehensive reference. Specifically, the CLAIR metric employs GPT-4o as the evaluation model.

The CIDEr [72] and SPICE metrics necessitate multiple reference sentences (i.e., multiple ground truths) to generate reliable evaluation outcomes. Given that RAG-IGBench provides only a single reference sentence, the reliability of these metrics is inherently compromised. While CLAIR demonstrates strong capability in differentiating text quality across various models, several considerations must be addressed: GPT-4o’s potential model bias, evaluation instability arising from API modifications, and the associated computational costs. In contrast, the ROUGE metric employs rule-based text quality assessment without requiring additional large language models, thereby facilitating more efficient and cost-effective evaluation.

In summary, while ROUGE metrics exhibit certain limitations in terms of evaluation reliability, CLAIR metrics present notable concerns regarding potential model bias and substantial additional computational overhead. Following a comprehensive assessment of these methodological trade-offs, we employ ROUGE metrics as our primary evaluation framework for the experimental analysis.

## 5 Limitation

Our RAG-IGBench, while demonstrating promising capabilities in evaluating the interleaved content generated by RAG-based methodologies, faces several important limitations that warrant discussion. First, although using model-generated answers as ground truth could potentially introduce biases, we have effectively mitigated this concern by employing diverse state-of-the-art models as generating models complemented by expert manual refinements. Second, while the data originates exclusively from the Xiaohongshu platform, this actually enhances content consistency while still providing substantial diversity—our dataset encompasses queries across nine categories including culture, health, education, etc., as demonstrated in the figure 4. The inherent high-quality standards of Xiaohongshu content, combined with our meticulous annotation process, ensure exceptional dataset quality despite the single-platform source. Third, regarding the predominantly Chinese original corpus, our utilization of SOTA LLMs like GPT-4o for translation ensures a high-fidelity English version with minimal semantic discrepancies. Finally, although the benchmark has been carefully curated to ensure high-quality question-answer pairs and feasible evaluation metrics, this comes at the expense of scale, thereby limiting its utility for pre-training purposes.

## 6 Conclusion

In this paper, we introduced RAG-IGBench, a novel benchmark designed specifically for evaluating interleaved image-text generation. Our benchmark advances the field through 4 key contributions: (1) A novel RAG-based approach for interleaved image-text generation. (2) A systematic benchmark with a meticulously curated dataset featuring diverse multimodal content. (3) Comprehensive evaluation metrics that assess text quality, image quality, and image-text coherence. (4) Extensive experimental analysis of both open-source and proprietary MLLMs. Our experimental results not only validate the benchmark’s effectiveness but also reveal important insights into the capabilities and limitations of current MLLMs in handling complex multimodal tasks. Future research directions include expanding the benchmark’s scale while maintaining its high-quality standards and developing enhanced evaluation methodologies for emerging multimodal generation tasks.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62441227, 62441239, U23A20319, 62502483) and the Anhui Province Science and Technology Innovation Project (Grant Nos. 202423k09020010, 202423k09020011, 2023z020006). We gratefully acknowledge the USTC supercomputing center and Xiaohongshu Inc. for providing computational resources that made this research possible.

## References

- [1] Jie An, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Lijuan Wang, and Jiebo Luo. Openleaf: Open-domain interleaved image-text generation and evaluation. *arXiv preprint arXiv:2310.07749*, 2023.
- [2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation, 2016.
- [3] Anthropic. The claude 3 model family: Opus, sonnet, haiku.
- [4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- [6] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. Benchmarking foundation models with language-model-as-an-examiner, 2023.
- [7] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.
- [8] David Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. Clair: Evaluating image captions with large language models. *arXiv preprint arXiv:2310.12971*, 2023.
- [9] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, 2021.
- [10] Dongping Chen, Ruoxi Chen, Shu Pu, Zhaoyi Liu, Yanru Wu, Caixi Chen, Benlin Liu, Yue Huang, Yao Wan, Pan Zhou, and Ranjay Krishna. Interleaved scene graph for interleaved text-and-image generation assessment, 2024.
- [11] Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuan-sheng Ni, Wang Zhu, Ziyang Jiang, Bohan Lyu, et al. Mega-bench: Scaling multimodal evaluation to over 500 real-world tasks. *arXiv preprint arXiv:2410.10563*, 2024.
- [12] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than \$3. <https://github.com/Deep-Agent/R1-V>, 2025. Accessed: 2025-02-02.
- [13] Yixiong Chen, Li Liu, and Chris Ding. X-iqe: explainable image quality evaluation for text-to-image generation with visual large language models. *arXiv preprint arXiv:2305.10843*, 2023.
- [14] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shen-glong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.
- [15] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.
- [16] Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023.
- [17] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint*, 2024.

- [18] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.
- [19] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fulí Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhua Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yachui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-rl: Incentivizing reasoning capability in llms via reinforcement learning, 2025.
- [20] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021.
- [21] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning, 2024.
- [22] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.
- [23] Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, et al. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*, 2024.
- [24] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive, 2024.
- [25] Hongchao Gu, Dexun Li, Kuicai Dong, Hao Zhang, Hang Lv, Hao Wang, Defu Lian, Yong Liu, and Enhong Chen. Rapid: Efficient retrieval-augmented long text generation with writing planning and information discovery. *arXiv preprint arXiv:2503.00751*, 2025.
- [26] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

- [27] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [28] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019.
- [29] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024.
- [30] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239, 2016.
- [31] Wenzuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.
- [32] Wenzuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2025.
- [33] Yuqing Huang, Rongyang Zhang, Xuesong He, Xuyang Zhi, Hao Wang, Xin Li, Feiyang Xu, Deguang Liu, Huadong Liang, Yi Li, et al. Chemeval: A comprehensive multi-level chemical evaluation for large language models. *arXiv preprint arXiv:2409.13989*, 2024.
- [34] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.
- [35] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [36] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. Mantis: Interleaved multi-image instruction tuning, 2024.
- [37] Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models, 2023.
- [38] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [39] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024.
- [40] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.
- [41] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024.
- [42] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024.

- [43] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.
- [44] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
- [45] Shiyu Li, Yang Tang, Shizhe Chen, and Xi Chen. Conan-embedding: General text embedding with more and better negative samples, 2024.
- [46] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [47] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [48] Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. Holistic evaluation for interleaved text-and-image generation. *arXiv preprint arXiv:2406.14643*, 2024.
- [49] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.
- [50] A Ian McLeod. Kendall rank correlation and mann-kendall trend test. *R package Kendall*, 602:1–10, 2005.
- [51] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [53] Eric Sven Ristad and Peter N Yianilos. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532, 1998.
- [54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.
- [55] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.
- [56] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.
- [57] Philip Sedgwick. Pearson’s correlation coefficient. *Bmj*, 345, 2012.
- [58] Philip Sedgwick. Spearman’s rank correlation coefficient. *Bmj*, 349, 2014.
- [59] Huiyang Shao, Qianqian Xu, Peisong Wen, Peifeng Gao, Zhiyong Yang, and Qingming Huang. Building bridge across the time: Disruption and restoration of murals in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20259–20269, October 2023.
- [60] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model, 2025.

- [61] Tingjia Shen, Hao Wang, Chuan Qin, Ruijun Sun, Yang Song, Defu Lian, Hengshu Zhu, and Enhong Chen. Genki: Enhancing open-domain question answering with knowledge integration and controllable generation in large language models, 2025.
- [62] Tingjia Shen, Hao Wang, Jiaqing Zhang, Sirui Zhao, Liangyue Li, Zulong Chen, Defu Lian, and Enhong Chen. Exploring user retrieval integration towards large language models for cross-domain sequential recommendation. *arXiv preprint arXiv:2406.03085*, 2024.
- [63] Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*, 2024.
- [64] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs, 2019.
- [65] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality, 2024.
- [66] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024.
- [67] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [68] Qwen Team. Qwen2.5-vl, January 2025.
- [69] Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, et al. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024.
- [70] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024.
- [71] Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *arXiv preprint arXiv:2311.16101*, 2023.
- [72] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015.
- [73] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024.
- [74] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. Vi-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025.
- [75] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.
- [76] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution, 2024.
- [77] Xiya Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*, 2024.

- [78] Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization, 2024.
- [79] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.
- [80] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023.
- [81] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, Xiaohong Liu, Guangtao Zhai, Shiqi Wang, and Weisi Lin. Towards open-ended visual quality comparison, 2024.
- [82] Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *arXiv preprint arXiv:2406.06007*, 2024.
- [83] Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, et al. Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models. *arXiv preprint arXiv:2410.10139*, 2024.
- [84] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.
- [85] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.
- [86] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.
- [87] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.
- [88] Kai Zhang, Lingbo Mo, Wenhua Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024.
- [89] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v(ision) as a generalist evaluator for vision-language tasks, 2023.
- [90] Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*, 2024.
- [91] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yunling Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024.
- [92] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

- [93] Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo, Hao Zhang, Yuqi Lin, Yefei He, Lirui Zhao, Shuo Liu, Tianhua Li, Yuxuan Xie, Xiaojun Chang, Yu Qiao, Wenqi Shao, and Kaipeng Zhang. Opening: A comprehensive benchmark for judging open-ended interleaved image-text generation, 2025.
- [94] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [95] Yongshuo Zong, Ondrej Bohdal, and Timothy M Hospedales. Vl-icl bench: The devil in the details of benchmarking multimodal in-context learning. *CoRR*, 2024.

## A Data Resource and Statistics during the data construction process

**Resource.** All data in our benchmark are sourced from Xiaohongshu, a popular Chinese social media platform. We have conducted a rigorous content check to ensure that all data does not involve personal privacy, patents, or intellectual property issues. The personnel involved in manual annotation are professional annotators who have undergone specialized training. Importantly, all annotators are active users of the Xiaohongshu platform, making them highly familiar with the platform’s image-text content and user preferences.

**Statistics.** We provide detailed data statistics during the data construction process in Table 8. Since we pre-filtered queries without visual requirements, most queries in the human annotation exhibit visual needs, and most model-generated answers contain images. During the annotation process, we categorized text quality, image quality, image-text consistency, and overall quality into three levels: 0, 1, and 2. Note that image quality and image-text consistency are scored individually for each image. "High text quality" and "High overall quality" refer to cases with a score of 2, while "High image quality" and "High image-text consistency" indicate an average score above 1.5 per image. "Refined cases" refers to instances where we rewrote cases that had only one low-scoring dimension while all other dimensions scored high, thereby obtaining additional ground-truth data. Finally, regarding the source distribution of ground truth, we employed three state-of-the-art closed-source MLLMs (GPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro) for raw data generation, thus mitigating bias issues from relying on a single model and achieving more robust and fair evaluation.

Table 8: Data statistics during the data construction process.

Statistics Of Raw Cases	Number	Percentage
Total raw cases	18634	100%
Generated by GPT-4o	9054	48.58%
Generated by Claude-3.5-sonnet	5550	29.78%
Generated by Gemini-1.5-pro	4032	21.63%
Query w. visual demands	17768	95.35%
Answer w. images	18321	98.32%
High text quality	11536	61.91%
High image quality	10400	55.81%
High image-text consistency	7277	39.05%
High overall quality	5334	28.67%
Refined cases	723	3.88%
Final dataset	6057	32.51%

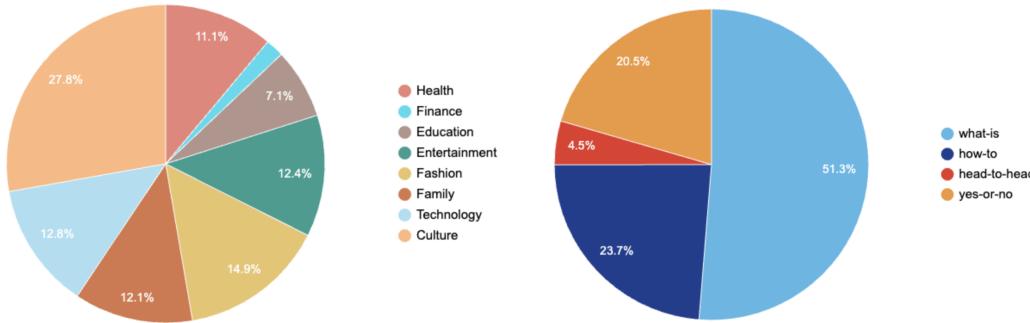


Figure 4: The distribution of distinct types of queries in RAG-IGBench.

## B Detailed Annotation Process

Section 3.3 has already provided an overview of the annotation process. The following section offers supplementary details regarding the first phase: raw answer generation.

In the first stage, we guide MLLMs (including GPT4o[35], Claude-3.5-Sonnet[3], et al.) with meticulously crafted instructions to generate the raw question-answer data. This stage takes a query, retrieved documents, and relevant images as input into an MLLM, resulting in a Markdown-formatted answer that includes image indices. To improve answer correctness and ensure consistency between text and images, we took inspiration from the Chain of Thought (CoT) methodology [79], which begins by instructing the model to analyze the input query, categorizing it into one of four types:

- The "what-is" query requires a direct response, ideally supplemented with an image to provide a clearer and more vivid understanding of the entity. For example, "What are the iconic landmarks in Zhengzhou?".
- The "how-to" query calls for a detailed, step-by-step action plan, complete with illustrations for each step to effectively guide the user through the process. For example, "How to tell if a carnation is a cold storage flower?".
- The "yes-or-no" query requires a straightforward affirmative or negative answer, complemented with images to enhance credibility while explaining the reason. For example, "Is *Gonjiam: Haunted Asylum* worth watching?".
- The "head-to-head" query involves comparisons, and it is most effective to present these with visual side-by-side images for immediate clarity. For example, "What is the difference between Calla Lily and Taro?".

The main reason for categorizing queries is that different types of queries have varying requirements for visual information. After query analysis, the model engages in thoughtful reasoning to plan the logic and structure of the answer. Figure 4 shows the distribution of four query categories. To further facilitate model comprehension, we incorporate an example in the instructions, in a manner analogous to the ICL approach [21]. For specific prompt details, please refer to F.

## C Dataset and Code Release

Following NeurIPS Dataset and Benchmark Track guidelines, we publicly release the RAG-IGBench dataset on HuggingFace at this link: <https://huggingface.co/datasets/Muyi13/RAG-IGBench>. RAG-IGBench is available under the Creative Commons Attribution License (CC BY 4.0).

We include example code and instructions for (1) inference using closed- and open-source models, (2) automatic evaluation of generated answers, and (3) examining and analyzing the evaluated results. Our evaluation and metric computation scripts, along with responses from all closed- and open-source models, are accessible in the GitHub repository <https://github.com/zry13/RAG-IGBench> under the Apache 2.0 license. We include the dataset card and README for the resources

## D Additional Experiment of Reasoning MLLMs

Generating content that seamlessly integrates textual and visual modalities poses significant challenges. Recently, research into the reasoning capabilities of MLLMs has intensified rapidly, making it particularly meaningful to examine their potential to enhance performance in such tasks. Below, we present additional experiments conducted on four different multimodal reasoning models. Specifically, VL-Rethinker-7B [74] and Vision-R1-7B [31] are open-source models obtained by fine-tuning QwenVL2.5-7B with GRPO and other techniques, whereas Claude3.7 (claude-3-7-sonnet-20250219) and OpenAI o1 (o1-2024-12-17) are closed-source models. For better comparison, we also include the results of QwenVL2.5-7b, GPT4o, and Claude3.5-sonnet from the main experiment, and the reasoning MLLMs are highlighted in bold in the table.

Table 9: Experiments of Reasoning MLLMs on RAG-IGBench. ↑ indicates that higher values are better.

Models	Rouge-1↑	Edit Distance↑	Kendall Score↑	Alignment Score↑	CLIP Score↑	Mean↑
Qwen2.5VL-7B	44.79	24.56	19.75	19.87	25.45	26.88
VL-Rethinker-7B [74]	46.70	27.42	23.24	21.54	28.99	29.58
Vision-R1-7B [31]	41.11	11.54	9.26	9.00	12.28	16.64
Claude3.5-sonnet	35.98	29.98	21.91	30.68	35.56	30.81
Claude3.7-sonnet-thinking	38.19	36.02	25.34	45.43	39.19	36.83
GPT4o	57.42	51.28	46.50	38.81	36.04	46.01
OpenAI o1	38.07	39.16	35.25	35.22	40.48	37.64

Both VL-Rethinker-7B and Vision-R1-7B are built upon the Qwen2.5VL-7B foundation model and trained using GRPO. However, VL-Rethinker-7B achieves a 2.7% improvement in overall performance, while Vision-R1-7B shows a 10.24% decline in comprehensive performance. We

attribute this disparity primarily to their distinct training paradigms: Vision-R1-7B is predominantly optimized for mathematical reasoning tasks, while VL-Rethinker-7B incorporates a broader spectrum of real-world reasoning scenarios. This limited scope is the primary reason for Vision-R1-7B’s inferior performance.

Regarding proprietary reasoning models, Claude-3.7-Sonnet-thinking with reasoning capabilities significantly outperforms the standard Claude-3.5-Sonnet on RAG-IGBench, demonstrating that enhanced reasoning abilities can improve performance on the RAG-IG task. Although the o1 model performs considerably worse than GPT-4o, we believe this is because o1’s multimodal understanding capabilities are inherently weaker than those of GPT-4o.

## E GRPO Experiment Results Based on Our Metrics

Recent developments in the DeepSeek-R1 [19] series of work introduced Group Relative Policy Optimization (GRPO), a novel training paradigm that employs rule-based reward functions to enhance the reasoning capabilities of LLMs significantly. Subsequently, numerous studies [60, 12, 32] have attempted to extend GRPO to MLLMs. However, most of these efforts have been limited to tasks with easily verifiable answers, such as Math, OVD, and REC. The reason is that open-domain question answering presents challenges for GRPO implementation due to the difficulty in efficiently evaluating responses through rule-based approaches.

Our RAG-IGBench addresses this limitation through the innovative metrics we proposed, which effectively solve the evaluation challenge. Leveraging these newly developed metrics, we have designed novel reward functions capable of scoring generated multimodal responses, thereby enabling GRPO experiments in open-domain multimodal contexts.

Table 10 presents our preliminary experimental results. We conducted Supervised Fine-Tuning (SFT) and Group Relative Policy Optimization (GRPO) experiments on both the 3B and 7B versions of the Qwen2.5VL-Instruct model. We compared the different improvements brought by SFT and GRPO, and further performed GRPO training on the SFT checkpoint to demonstrate the compatibility between these two approaches. For SFT training, we set the initial learning rate to 1e-5 with a warmup ratio of 0.05, and trained for a total of 3 epochs. For GRPO training, we set the generation number to 8 and trained for 5,000 steps. We designed the reward model and training scripts based on the VLM-R1 [60] codebase, and we acknowledge the open-source work of [60]. The training scripts for this experiment have also been made publicly available.

Models	Text Quality		Image Quality		Image-Text Consistency		Mean↑
	Rouge-1↑	Edit Distance↑	Kendall Score↑	Alignment Score↑	CLIP Score↑		
Qwen2.5VL-3B-Instruct	41.49	17.71	10.44	16.78	19.80	16.30	
	+ SFT	50.81	32.99	22.81	35.31	39.15	33.11
	+ GRPO	42.01	31.66	24.23	30.52	41.20	31.63
	+ SFT and GRPO	50.95	31.65	23.36	35.74	40.10	33.31
Qwen2.5VL-7B-Instruct	45.82	24.41	17.07	22.41	28.18	22.89	
	+ SFT	52.91	34.35	24.30	35.47	39.25	33.77
	+ GRPO	45.43	31.19	25.22	33.41	41.77	33.00
	+ SFT and GRPO	52.35	34.63	24.69	35.38	38.95	33.81

Table 10: The experiment result of GRPO training based on our innovative metrics.

The experimental results demonstrate that GRPO training yields improvements in the dimensions of image quality and image-text consistency, while showing minimal changes in text quality. Compared to SFT, GRPO achieves greater enhancements in Kendall score and CLIP score, though the improvement in edit distance is relatively modest. Furthermore, the results of the combined SFT+GRPO training confirm the compatibility of these two training approaches. We believe this is because GRPO enhances the model’s ability to understand multiple images, whereas SFT primarily improves the model’s textual capabilities.

## F Prompt Templates

Here, we present the prompt template used for generating interleaved image-text content. We employ Jinja2 as a templating system, where during inference, the prompt is populated with specific queries and relevant image-text content to form complete instructions.

1 You are an answer aggregation system that generates informative  
2 answers based on the user's query and relevant documents.  
3  
4 User queries can be categorized into the following types:  
5 1) what-is: definition type queries, for which the answer should be  
6 direct and include entity images when possible to make the  
7 response more illustrative;  
8 2) how-to: procedure type queries, for which the answer should outline  
9 a process with steps preferably accompanied by corresponding  
10 images;  
11 3) yes-or-no: verification type queries, for which the answer should  
12 be direct;  
13 4) head-to-head: comparison type queries, for which the answer should  
14 be presented in a table format.  
15  
16 Below you will receive a specific user query, relevant numbered  
17 documents (e.g., DOC#1), and corresponding numbered images (e.g.,  
18 IMG#1). Please determine the query category, provide an ideal  
19 answer, include appropriate reasoning, and insert suitable images  
20 to enhance your answer when necessary.  
21  
22 Input and output should follow the JSON format below:  
23  
24 Input:  
25 {  
26 "query": "xxx", # User query, str  
27 "documents and corresponding images": [ # Retrieved relevant  
28 documents and corresponding images, List[dict]  
29 {  
30 "document": "DOC#1\xxxx", # Relevant numbered document  
31 "images": [IMG#1, IMG#2, xxx] # Corresponding numbered  
32 images list  
33 },  
34 {  
35 "document": "DOC#2\xxxx",  
36 "images": [IMG#3, IMG#4, xxx]  
37 }  
38 ]  
39 }  
40  
41 Output:  
42 {  
43 "reason": "xxx", # Rationale for determining query category,  
44 approach for organizing the answer, and layout planning, str  
45 "category": "xxx", # Query category, must be one of ["what-is", "  
46 how-to", "yes-or-no", "head-to-head"], str  
47 "answer": "xxx" # Ideal answer, str  
48 }  
49  
50 The generated answer should meet the following requirements:  
51 1. Use \*\*markdown\*\* syntax with appropriate formatting;  
52 2. Properly cite documents as superscript references, e.g., xxx<sup>[3]</sup>xxx, where DOC#3 is the document number. Avoid  
53 explicitly mentioning phrases like 'data source' or 'according to  
54 the document'. Each document should be cited only once;  
55 3. Insert key images as core content, e.g., xxx![dummy](IMG#1)xxx,  
56 where IMG#1 represents the first image. Image placement should be  
57 appropriate, selection should be relevant, placement should be

```

precise, multiple images should maintain consistent style, and
each image should be inserted only once;
4. Formatting should prioritize rationality, aesthetics, and
simplicity.

Here is a question-answering example:
{
    "query": "difference between polo and Ralph Lauren"
    Output: "The user query is asking about the difference between
Polo and Ralph Lauren, which falls into the image-text
category. By providing comparative logo images of the brands,
we can visually demonstrate their differences while
complementing with brief textual explanations to better
address the user's question.",
    "category": "image-text",
    "answer": "The differences between Polo and Ralph Lauren are
mainly reflected in their brands and logo designs. Ralph
Lauren is an American luxury brand, and its men's clothing
line is called Polo Ralph Lauren, with a logo featuring a
single rider on a single horse, holding a mallet in the right
hand and leaning toward the left of the image[1](DOC#1)</
sup>. Polo Sport is a domestic copycat brand, also with a
single rider on a single horse logo, but holding a mallet in
the left hand and leaning toward the right of the image[2](DOC#2)</sup>.
Additionally, there's U.S. POLO ASSN.,
whose logo features two riders on two horses, with one rider
positioned in front of the other[3](DOC#3)</sup>. The
image below shows the differences between these brand logos:
![logo comparison](IMG#1)"
}
48
49 Here is the given input:
{
50
51     "query": {{query}},
52     "documents and corresponding images": [
53         {% for note in notes %}
54             {
55                 "document": "{{note['doc']]}}",
56                 "images": {{note['img']]}}
57             }
58             {% endfor %}
59     ]
60 }
61
62 Following the question-answering example, and based on the provided
documents and image information, please provide your output.
Remember to appropriately cite documents in your answer, insert
relevant images in suitable positions when necessary, and ensure
your output follows the JSON format.

```

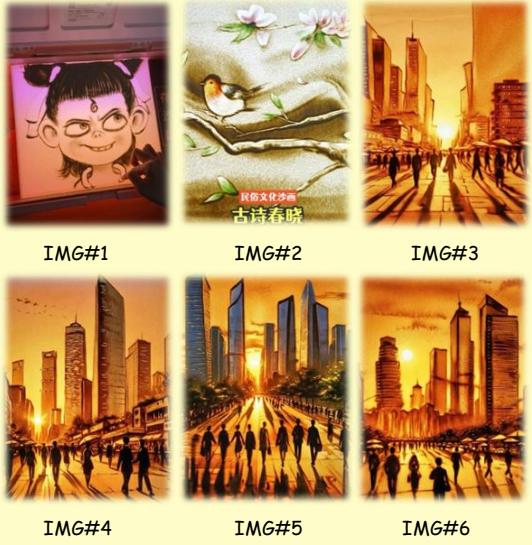
## G Qualitative Example

### G.1 Comparative generation results across different MLLMs

As illustrated in the figure 5, we show the outputs generated by distinct multimodal large language models responding to identical queries under our instructions. While all models successfully produced syntactically valid markdown-formatted responses, they exhibit notable variations in their presentation of image-text integration and visual-textual alignment.

## Query: Sand painting compositions

### Retrieved Images



### InternVL2.5-78B:

Sand painting is an ancient art form that is created by using fine sand on a sand tray. Sand paintings can depict a variety of vivid scenes, from natural scenery to urban bustle, which can be presented by the skillful hands of sand painters. Here are some examples of sand paintings:

- ![Sand painting 1](IMG#1)
- ![Sand painting 2](IMG#2)
- ![Sand painting 3](IMG#3)
- ![Sand painting 4](IMG#4)
- ![Sand painting 5](IMG#5)
- ![Sand painting 6](IMG#6)

These works show the diversity and unique charm of sand painting art.

### Qwen2.5VL-72B:

Sand painting is a unique art form that uses fine sand to create various patterns and pictures on a sand tray. Sand paintings can express a variety of themes and styles, from cartoon characters to natural landscapes to urban scenery, showing the diversity and creativity of sand painting art.

First, sand painting can be used to express cartoon characters, such as the image of Nezha in sand painting works, which is vivid and interesting, full of childlike fun and imagination. ![Nezha Sand Painting](IMG#1)

Second, sand painting can also be used to express natural landscapes and the artistic conception of ancient poems. For example, the picture of the ancient poem "Spring Dawn" in the sand painting works depicts the birdsong and flowers of spring through fine sand grains, showing the beauty and poetry of nature. ![Ancient Poetry Spring Dawn Sand Painting](IMG#2)

In addition, sand painting can also be used to express urban scenery and modern life. In a busy commercial district, sand painters use sand to depict high-rise buildings, pedestrians and sunsets, showing the prosperity and tranquility of the city. ![Urban Scenery Sand Painting](IMG#3)

Sand painting art has attracted the attention and love of many art lovers with its unique expression techniques and rich creative themes.

Figure 5: Comparison of results generated by different models

## G.2 Failure examples

The figure 6 illustrates an example of model error, demonstrating instances where the generated responses reference image indices that exceed the number of retrieved images. As analyzed in section 4.3, this represents a form of hallucination phenomenon. These findings highlight that addressing hallucinations in multimodal large language models constitutes a critical research direction. The observed inconsistencies between textual outputs and visual inputs underscore the challenges in achieving reliable cross-modal reasoning, suggesting that developing robust mechanisms to mitigate such hallucinations remains an important frontier in advancing multimodal AI systems.

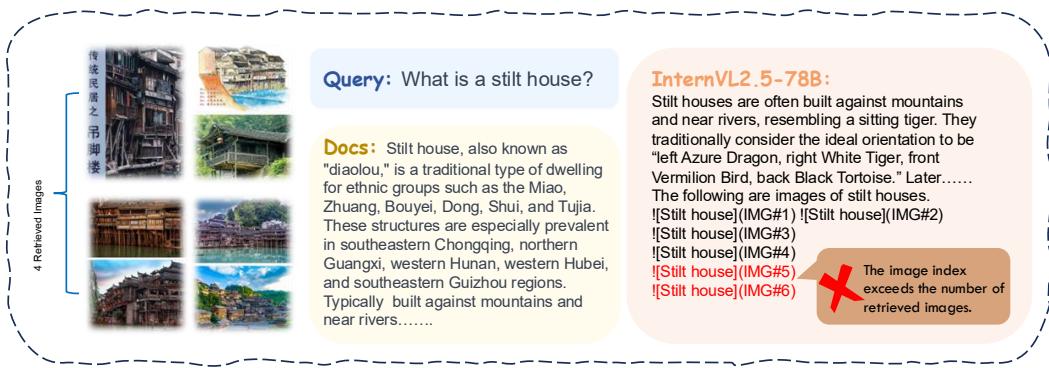


Figure 6: Examples of model failures.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The abstract and introduction clearly reflect the main contributions and scope of this paper, elucidating a novel approach and an innovative evaluation framework proposed in the field of multimodal generation.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: See appendix A.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: This paper provides open access to the source code and datasets, along with detailed documentation of the experimental procedures

### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: See appendix C.

### 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper outlines the key configurations and methodological settings implemented in 4. More granular information is provided in the publicly released codebase.

### 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Though we were unable to conduct extensive repeated evaluations due to computational resource limitations, we provide the data statistics as shown in the table 1.

### 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: See section 4.

#### 9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Our research adheres to the NeurIPS Code of Ethics.

#### 10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: See appendix A.

#### 11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

#### 12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All codes, datasets, and models used in the paper have been properly cited with their original sources.

#### 13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new assets introduced in this paper are accompanied by documentation, which is provided alongside the assets and includes detailed instructions for dataset evaluation and usage.

#### 14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing or research with human subjects.

#### 15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing or research with human subjects.

#### 16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

**Answer:** [NA]

**Justification:** The development of the core methods in this research does not involve LLMs as any important, original, or non-standard components.