OmniMapBench: Benchmarking Visual-Centric Reasoning on Diverse Map Documents

Yang Chen[1,2*], Yunwen Li[3*], Yufan Shen[1*], Minghao Liu[3,4*], Tianyu Zheng[3], Bin Fu[1], Qunshu Lin[4], Zhi Yu[2], and Botian Shi[1,5✉]

[1] Shanghai Artificial Intelligence Laboratory, Shanghai, China
shibotian@pjlab.org.cn,chenyang3@pjlab.org.cn
[2] Zhejiang University, Hangzhou, China
[3] M-A-P
[4] 2077AI
[5] Shanghai Innovation Institute, Shanghai, China

Abstract. Recent advancements in LVLMs necessitate robust benchmarks for complex, visually grounded reasoning. A critical limitation is identified in many document understanding benchmarks: visual content is often reducible to text, enabling high performance without genuine visual grounding. To address this limitation, OmniMapBench is introduced to foster visual-centric reasoning for map documents. The benchmark comprises 2,096 manually annotated question-answer pairs across 1,603 map documents from nine categories. It is designed to probe a hierarchy of skills, ranging from perception to multi-step visual reasoning. To quantify benchmark properties, a simple yet effective benchmark-level metric is proposed: the Visual Dependency Index (VDI), defined as the accuracy drop when images are replaced with question-agnostic descriptions. OmniMapBench exhibits higher VDI than established benchmarks, which quantitatively validates its focus on irreducible visual reasoning. Comprehensive evaluations of 25 leading LVLMs are conducted on OmniMapBench. A significant performance gap is observed, with the top-performing model achieving only 75.03% accuracy. This result underscores the challenges posed by OmniMapBench to current LVLMs. This work aims to catalyze progress in visual-centric reasoning for document understanding of LVLMs. The dataset and code are publicly available at https://github.com/SIGMME/OmniMapBench.

Keywords: Visual reasoning · Map understanding · MLLM

1 Introduction

Advances in the reasoning capabilities of Large Language Models (LLMs) [17,48, 69] have catalyzed the development of Large Vision-Language Models (LVLMs) [4,8,20,51,56]. Consequently, visual reasoning has emerged as a central research area [27, 66, 72, 88, 89]. The objective is shifting from perception tasks, such as

*Equal contribution. ✉Corresponding author.

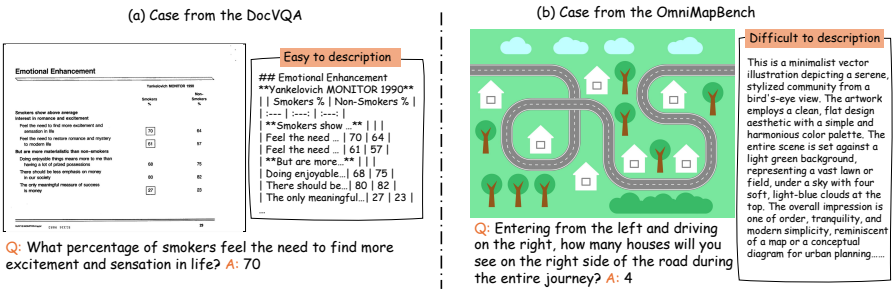

Fig. 1: Comparison of textualizability across benchmarks. (a) A DocVQA case with structured tabular content is easily described in text, and the question is solvable from the description. (b) In contrast, the *simple* map image from OmniMapBench cannot be sufficiently captured by a question-agnostic textual description to solve the question. The task requires visual-centric reasoning that is irreducible to a linear text format.

object recognition and captioning, towards complex inference grounded in visual content. To this end, various "think with images" [52,64,65,90] methods are explored, wherein models are trained to decompose intricate visual queries and execute programmatic actions via external tools [13, 77]. This paradigm facilitates a more adaptive and compositional approach to problem-solving, enabling models to address challenges that transcend simple visual perception.

The rapid evolution of model capabilities, however, places new demands on evaluation benchmarks. As shown in Figure 1, a significant limitation is identified in many existing document Visual Question Answering (VQA) benchmarks [46]: the visual content in these benchmarks is often reducible to structured text. For instance, document images can be converted into markdown representations through OCR and layout analysis [7, 10, 19, 32, 39, 63], which is a common practice in retrieval-augmented generation (RAG) pipelines. Charts and tables are frequently expressible as code or data structures [36,44,45]. In mathematical and geometric reasoning tasks, the visual diagrams are often text-complete [38,84,85], meaning that a verbal description of the geometric configuration suffices for problem solving [26,40,42]. This reducibility suggests that many existing benchmarks do not require genuine visual grounding and can be solved through text-based inference alone.

To address the need for more visual-centric evaluation, maps are put forth as a document category that inherently resists textual reduction [18]. Unlike documents amenable to OCR [33, 46] or table [61] and charts [44] deconstructible into codes, the core semantic content of a map is embedded within its spatial topology, symbolic conventions, and continuous graphical features [18,35,78]. Information is conveyed through the interplay of contour lines, color-coded regions, relational positions, and complex legends that cannot be adequately captured by linear text descriptions. For example, the seemingly simple map in Figure 1(b) illustrates this difficulty: without foreknowledge of the specific query, it is challenging to formulate a textual description that exhaustively encodes all spatial

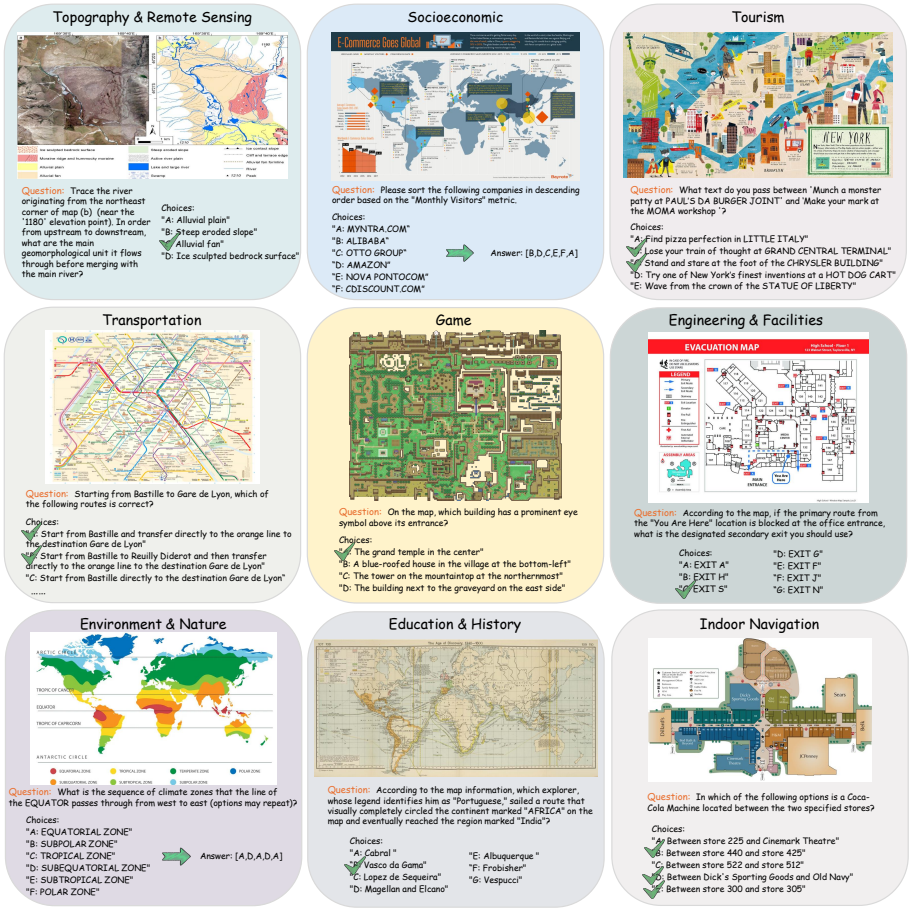

Fig. 2: Sample questions and answers from OmniMapBench across 9 map categories. The benchmark encompasses diverse visual styles and reasoning tasks, ranging from foundational topographic analysis to fictional virtual worlds and remote sensing imagery. Questions probe perception, single-step spatial reasoning, and multi-step relational reasoning grounded in map-specific visual cues.

relationships and symbolic details necessary for arbitrary reasoning tasks. Consequently, while human interpretation of a map may feel intuitive, the requisite synthesis of fine-grained perception and multi-step spatial reasoning presents a formidable challenge for current models, making maps an ideal medium for assessing genuine visual-centric reasoning.

This work introduces OmniMapBench, a benchmark designed specifically to target and evaluate visual-centric reasoning. As shown in Figure 2, the benchmark encompasses a diverse collection of maps, organized into 9 categories. This collection ranges from topographic and economic maps to historical and fantasy maps, ensuring high visual diversity. A set of approximately 2,096 question-answer pairs is constructed through a careful manual annotation process. To facilitate straightforward and objective evaluation, the tasks are structured into three formats: single-choice, multiple-choice, and ordering. These tasks are designed to probe a hierarchy of skills, including perception, single-step reasoning, and multi-step reasoning.

To quantitatively assess the reliance of a benchmark on visual information, this work proposes a simple yet effective metric. The Visual Dependency Index (VDI) measures the performance degradation observed when the full visual input is replaced by a text-only question-agnostic description generated by a vision-language model. A high VDI score signifies that a task is difficult to solve through textual shortcuts and thus requires genuine visual reasoning.

A comprehensive evaluation is conducted on 25 open-source and closed-source LVLMs. The results show that even the state-of-the-art model achieves 75.03% accuracy on OmniMapBench, indicating a substantial performance gap. Furthermore, experiments using the VDI protocol demonstrate that OmniMapBench possesses a significantly higher VDI compared to several established document VQA benchmarks. This finding quantitatively validates its reduced susceptibility to textualization shortcuts and confirms its focus on visual-centric reasoning.

The main contributions are summarized as follows:

- The introduction of OmniMapBench, a diverse, visual-centric benchmark for map understanding featuring 2,096 manually annotated and verified question-answer pairs of 1,603 map images across 9 map categories.
- The proposal of the Visual Dependency Index (VDI), a simple yet effective metric to quantify a LVLM benchmark's reliance on visual information.
- A comprehensive evaluation of 25 LVLMs is conducted, providing a comparison and identifying performance bottlenecks across map categories and visual-centric reasoning capabilities.

2 Related Work

2.1 Large Visual Language Models

Large Visual Language Models (LVLMs) have developed rapidly and have shown significant progress in various visual language tasks. Both open-source models [56, 73] and powerful proprietary models [20, 50] have pushed the frontiers

of artificial intelligence [3, 8, 67, 70]. Early progress was marked by a substantial improvement in foundational skills, such as optical character recognition (OCR) [29, 46, 62], enabling models to read and interpret documents and charts with increasing accuracy [7, 28, 31, 41, 55, 79, 80]. Subsequently, LVLMs mastered compositional abilities like complex spatial instruction following and multi-turn visual dialogue [1, 15, 37], shifting the research frontier towards higher-order visual reasoning [58, 86]. To address visual-centric reasoning, the research focus is centered on enabling models to "think with images" [53, 65]. Three strategies are studied. In visual synthesis, auxiliary images are generated to support the reasoning process [14, 76, 87]. In focused perception, salient regions are selected and zoomed for fine-grained inspection [27, 52, 59, 60, 88, 90]. In code-based reasoning, executable programs are produced and executed to derive answers [13, 66, 72, 89].

2.2 LVLM Benchmarks

Evaluations for Large Vision-Language Models [34, 82] range from foundational understanding [11, 83] to complex visual reasoning [54]. Initial benchmarks established baselines for foundational understanding, such as visual question answering [5, 24, 43] and reading text within images [44, 46, 62]. Subsequently, the evaluative focus broadened to compositional reasoning, where models are required to interpret relationships within structured data or maintain context across conversational turns [15, 16, 81]. This progression naturally culminated in benchmarks targeting higher-order cognitive reasoning, which probe a model's inferential logic, often by requiring an explicit rationale for an answer [59, 74], or challenge it with complex, multi-step problems from knowledge-intensive domains like science and mathematics [40, 84, 91]. Current evaluation efforts focus on benchmarks of greater complexity, designed to probe deeper visual-centric reasoning [10, 12, 25, 30] and test the capability boundaries of SOTA models.

3 OmniMapBench Dataset

The construction of the OmniMapBench dataset is a meticulous, multi-stage process designed to ensure high diversity, quality, and reasoning complexity. As illustrated in Figure 3, the pipeline is structured into three main stages: (1) Data Acquisition, where a broad collection of map images is gathered and filtered; (2) Data Annotation, where model-aided human experts create question-answer pairs; and (3) Cross-Validation, where the annotated pairs undergo rigorous verification to produce the final benchmark.

3.1 Data Acquisition

The primary goal of the acquisition stage is to assemble a large and visually diverse corpus of map documents. The process commences with the collection of over 5,000 map images. A dual-modality approach is employed to balance scale and quality. First, an automated web scraping procedure is executed. To ensure

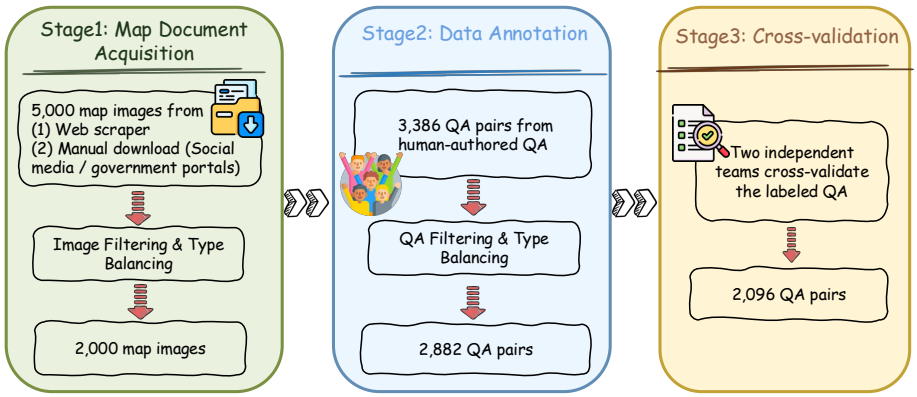

Fig. 3: Overview of the OmniMapBench dataset construction.

comprehensive coverage across the nine defined map categories, a LLM (GPT-5 [51]) is utilized to generate a wide array of search query variations from a set of seed keywords. Second, a targeted manual collection process is conducted to source high-quality, specialized maps. This is particularly important for categories such as transit maps and tourist diagrams, which demand high resolution and specific visual styles. These images are manually sourced from specialized platforms, including government portals and social media channels.

The initial corpus of 5,000 images is then subjected to a rigorous filtering protocol. Images are systematically removed based on three criteria: (1) high visual similarity to other images, indicating redundancy; (2) insufficient resolution or clarity, which would impede detailed analysis; and (3) content related to politically sensitive regions. After this screening, a refined set of approximately 2,000 high-quality map images is retained for the annotation stage.

3.2 Annotation and Validation

The annotation task is performed by professional human annotators. Human annotators are explicitly instructed to create **map-related challenging QA pairs**. To further standardize the evaluation and minimize ambiguity, QA pairs are formulated in objective formats, like single-choice, multiple-choice, or ranking questions. Strict guidelines are established for question creation. Each question must be answerable using only the visual information present in the map, without reliance on external knowledge. Annotators are also required to craft questions that span three predefined capability levels: basic perception (Level 1), single-step spatial reasoning (Level 2), and complex, multi-step relational reasoning (Level 3). This stage initially yields approximately 3,386 QA pairs. This collection then undergoes a preliminary filtering and type-balancing step, which reduces the set to 2,882 pairs for validation.

To ensure maximum quality and correctness, the 2,882 candidate QA pairs are subjected to a rigorous cross-validation process. The validation is carried out

Table 1: Statistics of the OmniMapBench dataset.

(a) Image category distribution.

Category	#	%
Indoor Navigation	294	18.34
Education & History	276	17.22
Engineering & Facilities	269	16.78
Transportation	251	15.66
Tourism	211	13.16
Topography & Remote Sensing	117	7.30
Environment & Nature	83	5.18
Socioeconomic	63	3.93
Game	39	2.43
Total	**1,603**	**100**

(b) QA pair attributes.

Attribute	**Category**	#	%
QA Type	Single Choice	1,461	69.70
	Multi Choice	269	12.83
	Sort	366	17.46
Capability	L1 (Perc.)	441	21.04
	L2 (Single.)	802	38.26
	L3 (Multi.)	853	40.70
Language	English	1,274	60.78
	Chinese	822	39.22
Total		**2,096**	**100**

by two independent teams of annotators who did not participate in the initial annotation of the data they review. Each team assesses every QA pair for clarity, correctness of the answer, and adherence to the specified capability level. Any pair that is deemed ambiguous, incorrect, or miscategorized by either team is discarded. This multi-layered verification protocol effectively eliminates potential errors and subjective biases. The entire construction pipeline concludes with a final set of approximately 2,096 meticulously verified QA pairs, which form the OmniMapBench dataset.

3.3 Dataset Statistics

Table 1 presents a detailed breakdown of the OmniMapBench statistics. The image collection is composed of 1,603 images distributed across nine categories. The largest concentrations of images are found in Indoor Navigation (18.34%), Education & History (17.22%), and Engineering & Facilities (16.78%). The remaining categories contribute to a wide-ranging visual diversity, from schematic diagrams to photorealistic scenes, ensuring comprehensive coverage. Furthermore, all images are retained at their native resolutions to preserve the authentic diversity and fine-grained visual density of real-world maps. As a result, the image sizes vary broadly from 233×464 to $11{,}811 \times 9{,}442$ pixels, with a mean resolution of $2{,}055 \times 1{,}705$. The question-answering component contains 2,096 pairs. These are distributed by type into single-choice (69.70%), multiple-choice (12.83%), and sorting (17.46%). In terms of capability levels, questions are classified as Level 3 (Multi-step reasoning) at 40.70%, Level 2 (Single-step reasoning) at

Table 2: Visual Dependency Index (VDI) and Descriptive Saturation Index (DSI) across benchmarks and token budgets (evaluated with gpt-4.1-2025-04-14 [49]). Values are rounded to three decimals. At $k_{\max} = 1024$, for VDI, the **maximum** value is bolded and the second maximum is underlined; for DSI, the **maximum** value is bolded and the minimum is underlined, highlighting benchmarks with distinct characteristics in visual dependency and descriptive complexity.

Benchmark	$k_{\max} = 64$		$k_{\max} = 128$		$k_{\max} = 256$		$k_{\max} = 512$		$k_{\max} = 1024$	
	VDI	DSI	VDI	DSI	VDI	DSI	VDI	DSI	VDI	DSI
AI2D [26]	0.892	0.780	0.613	0.790	0.159	0.745	0.035	0.597	0.019	0.336
ChartQA [44]	0.559	0.842	0.343	0.891	0.183	0.847	0.147	0.698	<u>0.143</u>	0.394
DocVQA [46]	0.552	0.855	0.357	0.944	0.224	0.920	0.122	0.795	0.076	0.516
InfoVQA [45]	0.517	0.897	0.352	0.990	0.241	0.985	0.155	0.961	0.049	**0.762**
MathVista [40]	0.882	0.722	0.682	0.722	0.347	0.617	0.126	0.478	0.051	0.269
MMMU [84]	0.917	0.676	0.841	0.727	0.526	0.669	0.221	0.503	0.093	0.293
MMMU-Pro [85]	0.904	0.666	0.848	0.710	0.578	0.615	0.257	0.459	0.117	0.273
MMBench [38]	0.822	0.591	0.510	0.629	0.114	0.518	0.035	0.364	0.028	0.195
OCRBench [39]	0.222	0.507	0.149	0.492	0.095	0.378	0.044	0.244	0.043	0.124
OCRBench-v2 [19]	0.411	0.781	0.282	0.861	0.160	0.813	0.108	0.680	0.085	0.437
OmniMapBench	0.556	0.911	0.527	0.983	0.520	0.963	0.474	0.915	**0.338**	<u>0.629</u>

38.26%, and Level 1 (Perception) at 21.04%. The language distribution consists of 60.78% English and 39.22% Chinese, which facilitates cross-lingual evaluation.

3.4 Comparison with other Benchmarks

A high-level contrast is provided in Table 3. Prior document-understanding benchmarks primarily focus on content such as industrial documents, charts, or OCR-rich images that are often reducible to text. More recent map-specific benchmarks have emerged, but they tend to specialize in a single map type, such as transit maps (*e.g.*, ReasonMap [18]) or choropleth maps (*e.g.*, MapQA [9]). In contrast, OmniMapBench encompasses 9 map categories with diverse visual styles and manually annotated QA pairs. Comparison of visual dependency is shown in Sec. 4, with additional results provided in the Supplementary.

4 Visual Dependency Evaluation

A large portion of existing VQA benchmarks is solvable through text-mediated shortcuts. In this setting, the direct pathway `(Image, Question) -> Answer` is replaced by first producing a textual question-agnostic description of the image and then answering from text. High accuracy is often achieved with short descriptions, which indicates predominantly linguistic reasoning and limited visual dependency. To quantify this reliance on vision, the Visual Dependency Index (VDI) is proposed as the normalized accuracy drop when image input is

Table 3: High-level comparison with other benchmarks.

Benchmark	Document Type	Cases	Images
DocVQA [46]	Industrial document	~50k	~12k
ChartQA [44]	Chart	~2.5k	~1.6k
AI2D [26]	Science diagram	~15k	~5k
OCRBenchV2 [19]	OCR-rich image	~10k	~1.5k
InfoVQA [45]	Infographic	~3k	~0.5k
ReasonMap [18]	Transit map	~1k	~30
MapQA [9]	Choropleth map	~800k	~60k
MapQA [35]	OpenStreetMap	~3k	/
OmniMapBench	Diverse maps	2,096	1,603

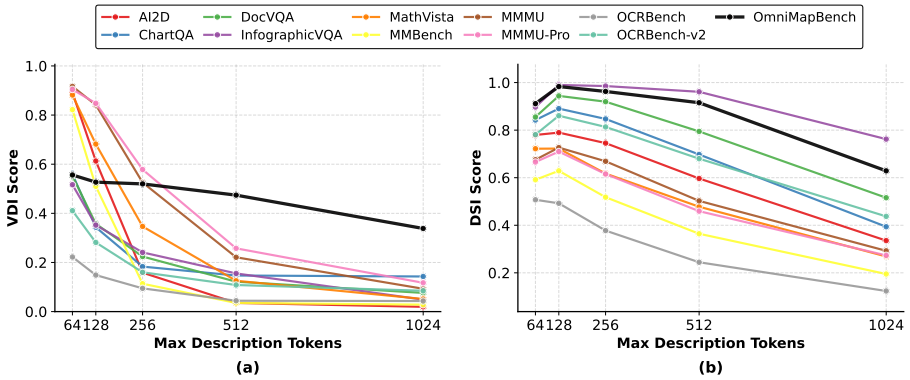

Fig. 4: Visualization of VDI (a) and DSI (b) across increasing description token budgets.

replaced by a textual surrogate under a token budget. A higher VDI indicates stronger dependence on visual cues that are not easily captured by text. Formal definitions are provided next.

4.1 Formal Definitions

To ensure a fair and standardized comparison across different benchmarks, the evaluation model is held constant. A single, powerful LVLM, denoted as $\mathcal{M}$, is selected to serve as the reference for measuring visual dependency. Let a benchmark be $\mathcal{B} = \{(I_i, Q_i, A_i)\}_{i=1}^{N}$, where I_i is an image, Q_i is a question, and A_i is the ground-truth answer. The standard performance of the reference model $\mathcal{M}$ on benchmark $\mathcal{B}$ is its accuracy on the direct inference task. This baseline, denoted as $\mathrm{Acc}_{\mathrm{full}}(\mathcal{B})$, is defined as:

$$\mathrm{Acc}_{\mathrm{full}}(\mathcal{B}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\mathcal{M}(I_i, Q_i) = A_i) \tag{1}$$

For the text-mediated pathway, the same model $\mathcal{M}$ is used to generate a description $D_{i,k_{\max}}$ for each image I_i, constrained by a maximum token budget of $k_{\max}$. Subsequently, $\mathcal{M}$ is queried again, this time operating in a language-only mode with an input composed of the generated description $D_{i,k_{\max}}$ and the original question Q_i. The text-only accuracy on benchmark $\mathcal{B}$, denoted as $\mathrm{Acc}_{\mathrm{text}}(\mathcal{B}, k_{\max})$, is then:

$$\mathrm{Acc}_{\mathrm{text}}(\mathcal{B}, k_{\max}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\mathcal{M}(D_{i,k_{\max}}, Q_i) = A_i) \tag{2}$$

The **Visual Dependency Index (VDI)** for a given benchmark $\mathcal{B}$, measured with respect to the reference model $\mathcal{M}$ and a token budget $k_{\max}$, is defined as the normalized performance drop:

$$\mathrm{VDI}(\mathcal{B}, k_{\max}) = \frac{\mathrm{Acc}_{\mathrm{full}}(\mathcal{B}) - \mathrm{Acc}_{\mathrm{text}}(\mathcal{B}, k_{\max})}{\mathrm{Acc}_{\mathrm{full}}(\mathcal{B})} \tag{3}$$

By fixing $\mathcal{M}$, $\mathrm{VDI}(\mathcal{B}, k_{\max})$ becomes a direct property of the benchmark $\mathcal{B}$ itself, quantifying its inherent dependency on visual information that resists textual summarization under a specific resource constraint $k_{\max}$. The VDI score ranges from approximately 0 to 1. A score near 0 indicates that the task can be almost entirely solved using textual information, while a score near 1 indicates a critical dependency on visual input.

A key methodological consideration in this process is that generative models operate with a maximum token limit ($k_{\max}$), not a fixed output length. The actual number of generated tokens, k_{actual}, is a variable that depends on both the model and the complexity of the image content. To analyze this behavior and to provide a complementary view on benchmark complexity, the **Descriptive Saturation Index (DSI)** is introduced as a secondary metric. DSI quantifies the model's tendency to utilize its allocated token budget.

The DSI for a benchmark $\mathcal{B}$ at a given $k_{\max}$ is defined as the average ratio of the actual tokens generated to the maximum allowed:

$$\mathrm{DSI}(\mathcal{B}, k_{\max}) = \mathbb{E}_{i \sim \mathcal{B}} \left[\frac{k_{\mathrm{actual},i}}{k_{\max}} \right] \tag{4}$$

where $k_{\mathrm{actual},i}$ is the length of the description generated for the i-th sample. A DSI value close to 1 suggests the benchmark's content is descriptively dense, consistently requiring the full token budget. A lower DSI indicates the content is more easily summarizable. Together, VDI and DSI provide a more comprehensive characterization of a benchmark's demands on visual reasoning.

4.2 Experimental Setup

Benchmarks. The evaluation protocol is applied to OmniMapBench and ten established visual reasoning benchmarks, as detailed in Table 2. To ensure a

comprehensive comparison, this selection spans multiple domains. It includes benchmarks focused on document understanding (*e.g.*, DocVQA [46]), chart reasoning (*e.g.*, ChartQA [44]), mathematical problem-solving (*e.g.*, MathVista), and broad multi-disciplinary reasoning (*e.g.*, MMMU [85], AI2D [26]).

Procedure and Models. The procedure for deriving VDI and DSI curves is outlined in Algorithm 1. A benchmark $\mathcal{B}$ with N samples is considered. A single reference LVLM, `gpt-4.1-2025-04-14` [49], is fixed as $\mathcal{M}$ for all stages: (i) obtaining $\text{Perf}_{\text{full}}$ via direct visual inference on (I_i, Q_i); (ii) generating descriptions $D_{i,k_{\max}}$ under token budgets $K = \{64, 128, 256, 512, 1024\}$; and (iii) performing text-only inference with $(D_{i,k_{\max}}, Q_i)$. For each $k_{\max}$, the actual token length $k_{\text{actual},i}$ is recorded and exact-match accuracy $\text{Acc}_{\text{text}}(k_{\max})$ is computed. VDI and DSI are then computed following Equations 3 and 4. The prompts for generating descriptions without Q_i and answering with Q_i are shown in Figure 5 and Figure 6.

Prompt for Generating Image Descriptions

Please provide a text description for this image that is as detailed as possible. You can also use the corresponding code to describe information such as charts/tables. No more than {max_tokens} tokens. Only output the description, no other text.

Fig. 5: The prompt used to generate question-agnostic image descriptions for the VDI analysis

Prompt for Answering from Description

{Question}. Now, Please answer the question only based on the following image description: {image_description}

Fig. 6: The prompt used for the text-only answering stage of the VDI analysis. It forces the model to rely exclusively on the generated text description, thereby measuring performance without direct visual input.

4.3 Evaluation Results and Analysis

The empirical results for the VDI and DSI are presented in Figure 4 and Table 2. VDI decreases with larger token budgets across benchmarks. A slower decay is

Algorithm 1 Pseudocode for the Visual Dependency Evaluation Experiment Procedure

Input: Benchmark $\mathcal{B} = \{(I_i, Q_i, A_i)\}_{i=1}^{N}$, VLM $\mathcal{M}$,
Baseline performance $\text{Perf}_{\text{full}}$, Token budgets $K = \{k_{\max}^{(1)}, \ldots, k_{\max}^{(m)}\}$

for each $k_{\max} \in K$ **do**
  Initialize $S_{\text{text}} \leftarrow 0$ and $L_{\text{total}} \leftarrow 0$
  for $i = 1$ to N **do**
    $D_{i,k_{\max}} \leftarrow \mathcal{M}.\text{GenerateDescription}(I_i, k_{\max})$
    $k_{\text{actual},i} \leftarrow \text{Length}(D_{i,k_{\max}})$
    $L_{\text{total}} \leftarrow L_{\text{total}} + k_{\text{actual},i}$
    $\hat{A}_i \leftarrow \mathcal{M}(D_{i,k_{\max}}, Q_i)$
    $S_{\text{text}} \leftarrow S_{\text{text}} + \mathbb{I}(\hat{A}_i = A_i)$
  end for
  $\text{Acc}_{\text{text}}(k_{\max}) \leftarrow S_{\text{text}}/N$
  $\text{VDI}(k_{\max}) \leftarrow (\text{Perf}_{\text{full}} - \text{Acc}_{\text{text}}(k_{\max}))/\text{Perf}_{\text{full}}$
  $\text{DSI}(k_{\max}) \leftarrow (L_{\text{total}}/N)/k_{\max}$
end for

Output: VDI and DSI curves parameterized by K

observed for OmniMapBench; at $k_{\max} = 1024$, a VDI of 0.338 is registered, the highest among all benchmarks. In contrast, benchmarks such as AI2D and MMBench drop to near-zero VDI (0.019 and 0.028), indicating that long descriptions largely eliminate visual dependency for them.

A complementary view is provided by DSI in Figure 4(b). OmniMapBench maintains high saturation across budgets, remaining at 0.629 at $k_{\max} = 1024$, the second highest at the maximum budget. This behavior indicates information-dense visual content that compels extensive token utilization, whereas several benchmarks (e.g., OCRBench at 0.124 and MMBench at 0.195) exhibit much lower saturation at large budgets. Taken together, the high DSI and VDI validate that OmniMapBench resists textual reduction and requires genuine visual-centric reasoning.

4.4 Ablation on the Reference Model

To assess sensitivity to the choice of the reference model, an ablation is conducted by replacing M with another LVLM. The results demonstrate consistent trends across different reference models. Detailed experiment results are available in the Supplementary.

5 Benchmarks

This section details the comprehensive evaluation of leading LVLMs on the OmniMapBench dataset.

Table 4: Main evaluation results on OmniMapBench. Models are grouped by category and sorted by Final Score. The best result is **bolded**, and the second-best is underlined.

Category	Model	Level 1	Level 2	Level 3	Final Score
Proprietary MLLMs	Gemini-3.1-Pro [22]	88.24	81.53	62.88	**75.03**
	Gemini-2.5-Pro [20]	73.70	70.70	46.31	61.40
	Doubao-seed-1.6-vision [8]	72.94	63.65	41.80	56.65
	GPT-5 [51]	63.10	61.47	48.53	56.54
	GPT-5-mini [51]	65.60	62.59	45.13	56.11
	GPT-4.1-mini [49]	60.59	47.13	35.64	45.27
	Claude-Sonnet-4-5 [4]	58.93	52.63	31.29	45.04
	GPT-4.1 [49]	48.52	45.14	37.87	42.88
	Claude-Opus-4-1 [2]	56.12	44.47	33.50	42.25
	GPT-5-nano [51]	45.56	40.52	32.83	38.44
	GPT-4o [47]	40.09	31.80	26.73	31.47
Open-Source MLLMs	Qwen3.5-397B-A17B [57]	83.89	79.01	60.37	<u>72.22</u>
	Qwen3.5-Plus [57]	85.32	77.64	57.47	70.80
	Qwen3.5-27B [57]	82.93	77.10	53.97	68.62
	Qwen3.5-122B-A10B [57]	83.13	76.52	53.54	68.37
	Kimi-K2.5 [68]	74.77	74.80	56.71	67.64
	Qwen3.5-35B-A3B [57]	81.10	72.27	49.93	64.75
	Qwen3.5-Flash [57]	79.27	73.87	48.62	64.56
	Gemma-4-31B-it [23]	60.77	58.35	40.68	51.67
	GLM-4.5V [71]	66.67	54.24	35.80	49.33
	InternVL3.5-241B-A28B [73]	68.48	52.99	30.36	47.04
	Intern-S1 [6]	61.22	49.13	32.71	44.99
	Gemma-4-26B-A4B-it [23]	51.25	49.75	32.83	43.18
	Qwen3-VL-8B [56]	60.09	44.89	29.07	41.65
	Gemma-3-27B-it [21]	34.92	31.80	25.44	29.87

5.1 Setup

Evaluated Models. A comprehensive suite of 25 representative LVLMs is evaluated to provide a broad view of the current landscape. These models are categorized into proprietary and open-source systems. The proprietary models include leading LVLMs such as GPT-5, Gemini-3.1-Pro. The open-source selection encompasses prominent models like Qwen3.5 and GLM series. A complete list of all evaluated models is provided in Table 4.

Evaluation Protocol. For all question types (single-choice, multiple-choice, and sorting), a strict accuracy metric is employed. A model's response is considered correct only if it exactly matches the ground-truth. To elicit the maximum reasoning capabilities of the models, a Chain-of-Thought [75] prompting strategy is applied. The prompts and decoding configurations are available in the Supplementary.

5.2 Results and Analysis

The primary evaluation results are presented in Table 4 and Figure 7. Table 4 decomposes performance across three capability levels, while Figure 7 provides a category-specific breakdown across the nine map types.

Overall Performance. A pronounced disparity across models is observed. The best final accuracy is achieved by Gemini-3.1-Pro at 75.03%. Among open-source models, Qwen3.5-397B-A17B attains the highest accuracy at 72.22%, demonstrating high competitiveness with the top proprietary system, with a narrow gap of less than 3 percentage points. Notably, the leading open-source models, such as the Qwen3.5 series, significantly outperform the majority of other proprietary models listed, challenging the notion that proprietary systems consistently hold a significant advantage. This indicates that while OmniMapBench remains a difficult benchmark, the gap in visual-centric reasoning between the best open-source and proprietary models is closing.

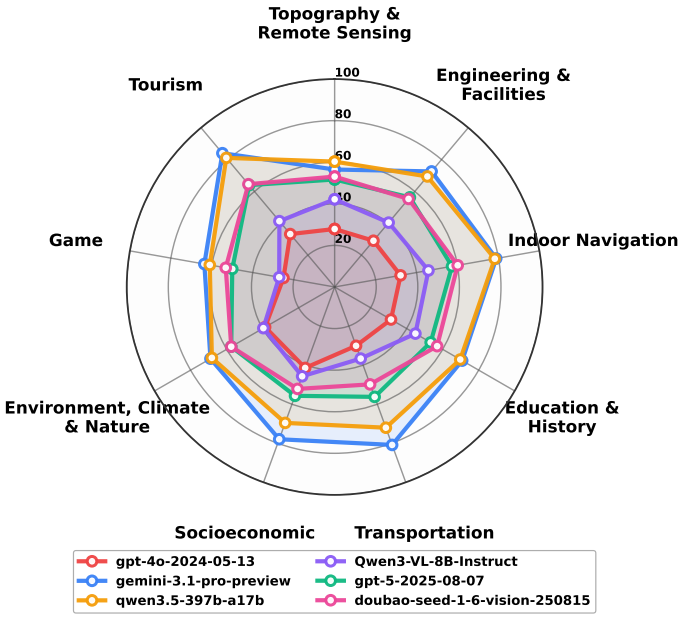

Fig. 7: Radar chart comparing the performance of six representative models across nine map categories in OmniMapBench.

No-Image Baseline. To assess whether OmniMapBench can be solved by relying on language priors, world knowledge, or answer-choice bias, a no-image ablation

is conducted. In this setting, only the textual question and the corresponding answer choices are provided to the models, without any visual input or textual image description. Three representative models are evaluated under this protocol. A substantial accuracy drop is observed, with the average performance decreasing from 58.87% in the standard setting to 23.35% in the blind setting. This degradation indicates that OmniMapBench cannot be reliably solved through textual shortcuts or inherent biases alone. Detailed per-model results for this baseline are provided in the Supplementary.

Performance by Capability Level. A clear trend of declining performance with increasing reasoning complexity is observed, from Level 1 (Perception) to Level 3 (Multi-step reasoning). For instance, Gemini-3.1-Pro records accuracies of 88.24%, 81.53%, and 62.88% at Levels 1, 2, and 3, respectively. These results indicate that while models handle basic perception and direct information extraction relatively well, multi-step spatial and visual-centric reasoning remains the principal bottleneck.

Performance by Map Category. Figure 7 reveals considerable performance variance across different map types. This suggests that model capabilities are not uniform across the diverse visual domains represented in OmniMapBench. Performance is generally higher on map types that contain more structured or conventional textual information, such as *Tourism*. In contrast, lower scores are observed for categories heavily reliant on the interpretation of dense, non-standard symbols and complex topological relationships, like *Topography & Remote Sensing*. This finding indicates a direction for future work in improving the LVLMs.

6 Conclusion

This work introduces OmniMapBench, a visual-centric benchmark for map document understanding, encompassing 2,096 manually verified QA pairs across 9 categories and spanning perception to multi-step reasoning tasks. The Visual Dependency Index (VDI) is proposed to quantify resistance to textual reduction under token budgets. Comprehensive evaluations reveal that state-of-the-art LVLMs achieve 75.03% accuracy, with high residual VDI confirming the benchmark's reliance on genuine visual grounding. The findings expose significant performance gaps in visual-centric reasoning. The dataset, metrics, and protocol provide a principled framework for assessing visual-centric reasoning of map documents. It is hoped that this benchmark catalyzes progress toward more robust multimodal reasoning and scalable perception-reasoning integration.

Acknowledgements

This work is supported by Shanghai Artificial Intelligence Laboratory.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. Anthropic: Claude-opus-4-1. https://www.anthropic.com/news/claude-opus-4-1 (2025)
3. Anthropic: Claude-Sonnet-4. https://www.anthropic.com/claude/sonnet (2025)
4. Anthropic: Claude-sonnet-4-5. https://www.anthropic.com/news/claude-sonnet-4-5 (2025)
5. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
6. Bai, L., Cai, Z., Cao, Y., Cao, M., Cao, W., Chen, C., Chen, H., Chen, K., Chen, P., Chen, Y., et al.: Intern-s1: A scientific multimodal foundation model. arXiv preprint arXiv:2508.15763 (2025)
7. Blecher, L., Cucurull, G., Scialom, T., Stojnic, R.: Nougat: Neural optical understanding for academic documents. arXiv preprint arXiv:2308.13418 (2023)
8. ByteDance: Seed1-6. https://seed.bytedance.com/en/seed1_6 (2025)
9. Chang, S., Palzer, D., Li, J., Fosler-Lussier, E., Xiao, N.: Mapqa: A dataset for question answering on choropleth maps (2022), https://arxiv.org/abs/2211.08545
10. Chen, L., Xie, W., Liang, Y., He, H., Zhao, H., Yang, Z., Huang, Z., Wu, H., Lu, H., charles, Y., Bao, Y., Fan, Y., Li, G., Shen, H., Chen, X., Xu, W., Si, S., Cai, Z., Chai, W., Huang, Z., Liu, F., Liu, T., Chang, B., Hu, X., Chen, K., Ren, Y., Liu, Y., Gong, Y., Li, K.: Babyvision: Visual reasoning beyond language (2026), https://arxiv.org/abs/2601.06521
11. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
12. Chen, Y., Liu, M., Shen, Y., Li, Y., Huang, T., Fang, X., Zheng, T., Huang, W., Yang, C., Fu, D., et al.: Iwr-bench: Can lvlms reconstruct interactive webpage from a user interaction video? arXiv preprint arXiv:2509.24709 (2025)
13. Chen, Y., Shen, Y., Huang, W., Zhou, S., Lin, Q., Cai, X., Yu, Z., Bu, J., Shi, B., Qiao, Y.: Learning only with images: Visual reinforcement learning with reasoning, rendering, and visual feedback. arXiv preprint arXiv:2507.20766 (2025)
14. Chern, E., Hu, Z., Chern, S., Kou, S., Su, J., Ma, Y., Deng, Z., Liu, P.: Thinking with generated images. arXiv preprint arXiv:2505.22525 (2025)
15. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual dialog. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 326–335 (2017)
16. De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., Courville, A.: Guesswhat?! visual object discovery through multi-modal dialogue. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5503–5512 (2017)
17. DeepSeek-AI, Guo, D., Yang, D., Zhang, H., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025), https://arxiv.org/abs/2501.12948

18. Feng, S., Wang, S., Ouyang, S., Kong, L., Song, Z., Zhu, J., Wang, H., Wang, X.: Reasonmap: Towards fine-grained visual reasoning from transit maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 41077–41088 (2026)
19. Fu, L., Kuang, Z., Song, J., Huang, M., Yang, B., Li, Y., Zhu, L., Luo, Q., Wang, X., Lu, H., et al.: Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. arXiv preprint arXiv:2501.00321 (2024)
20. Google: Gemini-2.5-pro. `https://deepmind.google/models/gemini/pro/` (2025)
21. Google: Gemma-3. `https://deepmind.google/models/gemma/gemma-3/` (2025)
22. Google: Gemini-3.1-pro. `https://deepmind.google/models/gemini/pro/` (2026)
23. Google: Gemma-4. `https://deepmind.google/models/gemma/gemma-4/` (2026)
24. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017)
25. Guo, M.H., Xu, J., Zhang, Y., Song, J., Peng, H., Deng, Y.X., Dong, X., Nakayama, K., Geng, Z., Wang, C., et al.: R-bench: Graduate-level multi-disciplinary benchmarks for llm & mllm complex reasoning evaluation. arXiv preprint arXiv:2505.02018 (2025)
26. Hiippala, T., Alikhani, M., Haverinen, J., Kalliokoski, T., Logacheva, E., Orekhova, S., Tuomainen, A., Stone, M., Bateman, J.A.: Ai2d-rst: a multimodal corpus of 1000 primary school science diagrams. Language Resources and Evaluation **55**(3), 661–688 (2021)
27. Huang, X., Dong, Y., Tian, W., Li, B., Feng, R., Liu, Z.: High-resolution visual reasoning via multi-turn grounding-based reinforcement learning. arXiv preprint arXiv:2507.05920 (2025)
28. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. In: Proceedings of the 30th ACM international conference on multimedia. pp. 4083–4091 (2022)
29. Jaume, G., Ekenel, H.K., Thiran, J.P.: Funsd: A dataset for form understanding in noisy scanned documents. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 2, pp. 1–6. IEEE (2019)
30. Jimenez, C.E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., Narasimhan, K.: Swe-bench: Can language models resolve real-world github issues? arXiv preprint arXiv:2310.06770 (2023)
31. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: European Conference on Computer Vision. pp. 498–517. Springer (2022)
32. Kim, G., Hong, T., Yim, M., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Donut: Document understanding transformer without ocr. arXiv preprint arXiv:2111.15664 **7**(15), 2 (2021)
33. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
34. Li, J., Lu, W., Fei, H., Luo, M., Dai, M., Xia, M., Jin, Y., Gan, Z., Qi, D., Fu, C., et al.: A survey on benchmarks of multimodal large language models. arXiv preprint arXiv:2408.08632 (2024)

35. Li, Z., Grossman, M., Kulkarni, M., Chen, M., Chiang, Y.Y., et al.: Mapqa: Open-domain geospatial question answering on map data. arXiv preprint arXiv:2503.07871 (2025)
36. Liu, F., Eisenschlos, J., Piccinno, F., Krichene, S., Pang, C., Lee, K., Joshi, M., Chen, W., Collier, N., Altun, Y.: Deplot: One-shot visual language reasoning by plot-to-table translation. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 10381–10399 (2023)
37. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
38. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
39. Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.C., Liu, C.L., Jin, L., Bai, X.: Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences **67**(12), 220102 (2024)
40. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
41. Luo, C., Shen, Y., Zhu, Z., Zheng, Q., Yu, Z., Yao, C.: Layoutllm: Layout instruction tuning with large language models for document understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15630–15640 (2024)
42. Luo, S., Zhu, Z., Yuan, Y., Yang, Y., Shan, L., Wu, Y.: Geogrambench: Benchmarking the geometric program reasoning in modern llms. arXiv preprint arXiv:2505.17653 (2025)
43. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 3195–3204 (2019)
44. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244 (2022)
45. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1697–1706 (2022)
46. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021)
47. OpenAI: GPT-4o. https://openai.com/zh-Hans-CN/index/hello-gpt-4o/ (2024)
48. OpenAI: OpenAIo1. https://openai.com/zh-Hans-CN/o1/ (2024)
49. OpenAI: GPT-4-1. https://openai.com/index/gpt-4-1/ (2025)
50. OpenAI: GPT-4V-system-card. https://openai.com/index/gpt-4v-system-card/ (2025)
51. OpenAI: Introducing-GPT-5. https://openai.com/zh-Hans-CN/index/introducing-gpt-5/ (2025)
52. OpenAI: OpenAI-o3 . https://openai.com/zh-Hans-CN/index/introducing-o3-and-o4-mini/ (2025)
53. OpenAI: Thinking-with-Images. https://openai.com/zh-Hans-CN/index/thinking-with-images/ (2025)

54. Park, J.S., Bhagavatula, C., Mottaghi, R., Farhadi, A., Choi, Y.: Visualcomet: Reasoning about the dynamic context of a still image. In: European Conference on Computer Vision. pp. 508–524. Springer (2020)
55. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
56. QwenTeam: Qwen3-VL. https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef&from=research.latest-advancements-list (2025)
57. QwenTeam: Qwen3.5. https://qwen.ai/blog?id=qwen3.5 (2026)
58. Saikh, T., Ghosal, T., Mittal, A., Ekbal, A., Bhattacharyya, P.: Scienceqa: A novel resource for question answering on scholarly articles. International Journal on Digital Libraries **23**(3), 289–301 (2022)
59. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. Advances in Neural Information Processing Systems **37**, 8612–8642 (2024)
60. Shen, H., Zhao, K., Zhao, T., Xu, R., Zhang, Z., Zhu, M., Yin, J.: Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. arXiv preprint arXiv:2411.16044 (2024)
61. Shen, Y., Luo, C., Zhu, Z., Chen, Y., Zheng, Q., Yu, Z., Bu, J., Yao, C.: Proctag: Process tagging for assessing the efficacy of document instruction data. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6851–6859 (2025)
62. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
63. Stanisławek, T., Graliński, F., Wróblewska, A., Lipiński, D., Kaliska, A., Rosalska, P., Topolski, B., Biecek, P.: Kleister: key information extraction datasets involving long documents with complex layouts. In: International Conference on Document Analysis and Recognition. pp. 564–579. Springer (2021)
64. Su, Z., Li, L., Song, M., Hao, Y., Yang, Z., Zhang, J., Chen, G., Gu, J., Li, J., Qu, X., et al.: Openthinkimg: Learning to think with images via visual tool reinforcement learning. arXiv preprint arXiv:2505.08617 (2025)
65. Su, Z., Xia, P., Guo, H., Liu, Z., Ma, Y., Qu, X., Liu, J., Li, Y., Zeng, K., Yang, Z., et al.: Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. arXiv preprint arXiv:2506.23918 (2025)
66. 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)
67. Team, C., Yue, Z., Lin, Z., Song, Y., Wang, W., Ren, S., Gu, S., Li, S., Li, P., Zhao, L., et al.: Mimo-vl technical report (2025), https://arxiv.org/abs/2506.03569
68. Team, K., Bai, T., Bai, Y., Bao, Y., Cai, S.H., Cao, Y., Charles, Y., Che, H.S., Chen, C., Chen, G., Chen, H., et al.: Kimi k2.5: Visual agentic intelligence (2026), https://arxiv.org/abs/2602.02276
69. Team, K., Du, A., Gao, B., Xing, B., Jiang, C., Chen, C., Li, C., Xiao, C., Du, C., Liao, C., et al.: Kimi k1. 5: Scaling reinforcement learning with llms. arXiv preprint arXiv:2501.12599 (2025)

70. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., Wang, C., Zhang, D., Du, D., et al.: Kimi-vl technical report (2025), https://arxiv.org/abs/2504.07491
71. Team, V., Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., Duan, S., et al.: Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning (2025), https://arxiv.org/abs/2507.01006
72. Wang, K., Pan, J., Wei, L., Zhou, A., Shi, W., Lu, Z., Xiao, H., Yang, Y., Ren, H., Zhan, M., et al.: Mathcoder-vl: Bridging vision and code for enhanced multimodal mathematical reasoning. arXiv preprint arXiv:2505.10557 (2025)
73. 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 (2025), https://arxiv.org/abs/2508.18265
74. Wang, Y., Wu, S., Zhang, Y., Yan, S., Liu, Z., Luo, J., Fei, H.: Multimodal chain-of-thought reasoning: A comprehensive survey, 2025. URL https://arxiv.org/abs/2503.12605 (2025)
75. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2023), https://arxiv.org/abs/2201.11903
76. Wu, J., Guan, J., Feng, K., Liu, Q., Wu, S., Wang, L., Wu, W., Tan, T.: Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. arXiv preprint arXiv:2506.09965 (2025)
77. Wu, M., Yang, J., Jiang, J., Li, M., Yan, K., Yu, H., Zhang, M., Zhai, C., Nahrstedt, K.: Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use. arXiv preprint arXiv:2505.19255 (2025)
78. Xing, S., Sun, Z., Xie, S., Chen, K., Huang, Y., Wang, Y., Li, J., Song, D., Tu, Z.: Can large vision language models read maps like a human? arXiv preprint arXiv:2503.14607 (2025)
79. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., et al.: Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. arXiv preprint arXiv:2012.14740 (2020)
80. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 1192–1200 (2020)
81. Yang, C., Shi, C., Liu, Y., Shui, B., Wang, J., Jing, M., Xu, L., Zhu, X., Li, S., Zhang, Y., et al.: Chartmimic: Evaluating lmm's cross-modal reasoning capability via chart-to-code generation. arXiv preprint arXiv:2406.09961 (2024)
82. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12), nwae403 (2024)
83. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European conference on computer vision. pp. 69–85. Springer (2016)
84. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., 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. pp. 9556–9567 (2024)
85. Yue, X., Zheng, T., Ni, Y., Wang, Y., Zhang, K., Tong, S., Sun, Y., Yu, B., Zhang, G., Sun, H., et al.: Mmmu-pro: A more robust multi-discipline multimodal under-

standing benchmark. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 15134–15186 (2025)
86. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6720–6731 (2019)
87. Zhang, H., Li, C., Wu, W., Mao, S., Zhang, Y., Tian, H., Vulić, I., Zhang, Z., Wang, L., Tan, T., et al.: Scaling and beyond: Advancing spatial reasoning in mllms requires new recipes. arXiv preprint arXiv:2504.15037 (2025)
88. Zhang, X., Gao, Z., Zhang, B., Li, P., Zhang, X., Liu, Y., Yuan, T., Wu, Y., Jia, Y., Zhu, S.C., et al.: Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. arXiv preprint arXiv:2505.15436 (2025)
89. Zhao, X., Luo, X., Shi, Q., Chen, C., Wang, S., Liu, Z., Sun, M.: Chartcoder: Advancing multimodal large language model for chart-to-code generation. arXiv preprint arXiv:2501.06598 (2025)
90. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing" thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025)
91. Zhou, P., Zhang, F., Peng, X., Xu, Z., Ai, J., Qiu, Y., Li, C., Li, Z., Li, M., Feng, Y., et al.: Mdk12-bench: A multi-discipline benchmark for evaluating reasoning in multimodal large language models. arXiv preprint arXiv:2504.05782 (2025)