Novelty Assessment Report

Paper: OCR-Reasoning Benchmark: Unveiling the True Capabilities of MLLMs in Complex Text-Rich Image Reasoning
PDF URL: https://openreview.net/pdf?id=aH7eyx64pC
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Recent advancements in multimodal slow-thinking systems have demonstrated remarkable performance across various visual reasoning tasks. However, their capabilities in text-rich image reasoning tasks remain understudied due to the absence of a dedicated and systematic benchmark. To address this gap, we propose OCR-Reasoning, a novel benchmark designed to systematically assess Multimodal Large Language Models on text-rich image reasoning tasks. Specifically, OCR-Reasoning comprises 1,069 human-annotated examples spanning 6 core reasoning abilities and 18 practical reasoning tasks in text-rich visual scenarios. Unlike existing text-rich image understanding benchmarks that only provide a final answer, this benchmark additionally provides a detailed step-by-step reasoning process. This dual annotation enables the evaluation of both the models' final answers and their reasoning processes, thereby offering a holistic assessment of text-rich reasoning capabilities. By leveraging this benchmark, we conducted a comprehensive evaluation of the latest MLLMs. Our results demonstrate that even the most advanced MLLMs exhibit substantial difficulties in text-rich image reasoning tasks, with none achieving an accuracy above 50\% on our benchmark, indicating that the challenges of text-rich image reasoning are an urgent issue to be addressed. The dataset and evaluation scripts will be made publicly available.

Core Task Landscape

This paper addresses: **Text-Rich Image Reasoning**

A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Model Architecture and Training Paradigms**
- **Reasoning Mechanisms and Cognitive Strategies**
- **Task-Specific Applications and Domains**
- **Evaluation Benchmarks and Datasets**

Complete Taxonomy Tree

- Text-Rich Image Reasoning Survey Taxonomy
- Model Architecture and Training Paradigms
  - Visual Instruction Tuning for Text-Rich Understanding (2 papers)
  - [1] Llavar: Enhanced visual instruction tuning for text-rich image understanding (Zhang Yan-zhe, 2023) View paper
  - [25] Enhanced visual instruction tuning for text-rich image understanding (Y Zhang, 2023) View paper
  - Synthetic Data Generation and Scaling (4 papers)
  - [4] Scaling text-rich image understanding via code-guided synthetic multimodal data generation (Yue Yang, 2025) View paper
  - [19] VisualWebInstruct: Scaling up Multimodal Instruction Data through Web Search (Yiming Jia, 2025) View paper
  - [32] FLUX-Reason-6M & PRISM-Bench: A Million-Scale Text-to-Image Reasoning Dataset and Comprehensive Benchmark (Fang, 2025) View paper
  - [33] Multimodal self-instruct: Synthetic abstract image and visual reasoning instruction using language model (Wenqi Zhang, 2024) View paper
  - Reinforcement Learning for Multimodal Reasoning (2 papers)
  - [23] DeepEyes: Incentivizing "Thinking with Images" via Reinforcement Learning (Zheng Zi-wei, 2025) View paper
  - [30] WeThink: Toward General-purpose Vision-Language Reasoning via Reinforcement Learning (Yang, 2025) View paper
  - Latent Space and Embedding-Based Reasoning (1 papers)
  - [9] Monet: Reasoning in Latent Visual Space Beyond Images and Language (Qixun Wang, 2025) View paper
- Reasoning Mechanisms and Cognitive Strategies
  - Chain-of-Thought and Multi-Stage Reasoning (5 papers)
  - [6] Textcot: Zoom in for enhanced multimodal text-rich image understanding (Luan, 2024) View paper
  - [11] Joint visual semantic reasoning: Multi-stage decoder for text recognition (Ayan Kumar Bhunia, 2021) View paper
  - [27] Vgr: Visual grounded reasoning (Wang Jiacong, 2025) View paper
  - [28] Reasoning elicitation and multi-granularity contrastive learning for text-rich image understanding in large vision-language models (Jiazhi Xia, 2025) View paper
  - [44] Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination (Zheng Hao-jie, 2024) View paper
  - Visual-Linguistic Fusion and Grounding (3 papers)
  - [18] Multi-modal graph neural network for joint reasoning on vision and scene text (Difei Gao, 2020) View paper
  - [20] Improving visual grounding with visual-linguistic verification and iterative reasoning (Li Yang, 2022) View paper
  - [26] Visual semantics allow for textual reasoning better in scene text recognition (Chen Chen, 2022) View paper

- Knowledge-Based and External Reasoning (3 papers)
- [10] Image understanding using vision and reasoning through scene description graph (Somak Aditya, 2018) View paper
- [22] Visual chain-of-thought prompting for knowledge-based visual reasoning (Zhenfang Chen, 2024) View paper
- [39] Reasoning and question answering about image-text multi-modal contexts (Sampat, 2024) View paper
- Spatial and Compositional Reasoning (3 papers)
- [3] Enhancing advanced visual reasoning ability of large language models (Zhiyuan Li, 2024) View paper
- [14] Understand, Think, and Answer: Advancing Visual Reasoning with Large Multimodal Models (Zhan Yufei, 2025) View paper
- [16] Enhancing Spatial Reasoning through Visual and Textual Thinking (Liang Xun, 2025) View paper
- Dual and Mixed Modality Reasoning (2 papers)
- [31] Mixed Signals: Decoding VLMs' Reasoning and Underlying Bias in Vision-Language Conflict (Pouya Pezeshkpour, 2025) View paper
- [45] Think Twice Before You Judge: Mixture of Dual Reasoning Experts for Multimodal Sarcasm Detection (Soumyadeep Jana, 2025) View paper

- Task-Specific Applications and Domains
  - OCR and Text Localization (3 papers)
  - [29] VCR: A Task for Pixel-Level Complex Reasoning in Vision Language Models via Restoring Occluded Text (Zhang Tianyu, 2024) View paper
  - [47] Mme-videoocr: Evaluating ocr-based capabilities of multimodal llms in video scenarios (Shi Yang, 2025) View paper
  - [48] Reasoning-OCR: Can Large Multimodal Models Solve Complex Logical Reasoning Problems from OCR Cues? (He Haibin, 2025) View paper
  - Text-Based Visual Question Answering (3 papers)
  - [42] Weakly-supervised 3d spatial reasoning for text-based visual question answering (Hao Li, 2023) View paper
  - [46] Enhancing sceneâ□□text visual question answering with relational reasoning, attention and dynamic vocabulary integration (Mayank Agrawal, 2024) View paper
  - [50] Vtqa: Visual text question answering via entity alignment and cross-media reasoning (Kang Chen, 2024) View paper
  - Text-to-Image Generation and Reasoning (4 papers)
  - [5] Textual-Visual Logic Challenge: Understanding and Reasoning in Text-to-Image Generation (Peixi Xiong, 2024) View paper
  - [7] T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation (SUN Kaiyue, 2025) View paper
  - [12] Mmmg: A massive, multidisciplinary, multi-tier generation benchmark for text-to-image reasoning (Luo Yuxuan, 2025) View paper
  - [43] WorldGenBench: A World-Knowledge-Integrated Benchmark for Reasoning-Driven Text-to-Image Generation (Zhang, 2025) View paper
  - Cross-Modal Retrieval and Entity Alignment (1 papers)
  - [24] Heterogeneous Prompt-Guided Entity Inferring and Distilling for Scene-Text Aware Cross-Modal Retrieval (Zhiqian Zhao, 2025) View paper
  - Domain-Specific Reasoning Applications (3 papers)
  - [15] Medisee: Reasoning-based pixel-level perception in medical images (Qinyue Tong, 2025) View paper
  - [35] Enhancing Emotion Reasoning for Image Multi-Emotion Prediction (Bingbing Wang, 2025) View paper
  - [36] MathReal: We Keep It Real! A Real Scene Benchmark for Evaluating Math Reasoning in Multimodal Large Language Models (Feng Jun, 2025) View paper
  - Multi-Image and Video Understanding (3 papers)
  - [2] Videochat: Chat-centric video understanding (Li, 2025) View paper
  - [21] Prima: Multi-image vision-language models for reasoning segmentation (Wahed, 2024) View paper
  - [34] Leopard: A vision language model for text-rich multi-image tasks (Jia, 2024) View paper
  - Region-Level and Pixel-Level Understanding (2 papers)
  - [37] Omni-RGPT: Unifying Image and Video Region-level Understanding via Token Marks (Miran Heo, 2025) View paper
  - [40] DiffUHaul: A Training-Free Method for Object Dragging in Images (Omri Avrahami, 2024) View paper
  - Web and Document Understanding (2 papers)
  - [38] Webwatcher: Breaking new frontier of vision-language deep research agent (Geng Xin-yu, 2025) View paper
  - [41] Structextv3: An efficient vision-language model for text-rich image perception, comprehension, and beyond (Lyu, 2024) View paper

- Evaluation Benchmarks and Datasets
  - Comprehensive Multimodal Understanding Benchmarks (2 papers)
  - [8] Multimodal large language models for text-rich image understanding: A comprehensive review (Fu Pei, 2025) View paper
  - [13] MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark (Xiang Yue, 2024) View paper
  - Specialized Task Benchmarks ★ (3 papers)
  - [0] OCR-Reasoning Benchmark: Unveiling the True Capabilities of MLLMs in Complex Text-Rich Image Reasoning (Anon et al., 2026) View paper
  - [17] Contextual: Evaluating context-sensitive text-rich visual reasoning in large multimodal models (Wadhawan, 2024) View paper
  - [49] Mctbench: Multimodal cognition towards text-rich visual scenes benchmark (Shan, 2024) View paper

Narrative

Core task: text-rich image reasoning. This field addresses the challenge of understanding and reasoning over images that contain substantial textual content—such as documents, charts, infographics, web pages, and scene text—where models must integrate visual perception with language comprehension. The taxonomy organizes research into four main branches: Model Architecture and Training Paradigms explores foundational designs and learning strategies (e.g., Llavar[1], Scaling Text-Rich[4]); Reasoning Mechanisms and Cognitive Strategies examines how models perform multi-step inference and leverage structured knowledge (e.g., Textual-Visual Logic[5], Textcot[6]); Task-Specific Applications and Domains targets specialized use cases like medical imaging (Medisee[15]), mathematical reasoning (MathReal[36]), and web navigation (Webwatcher[38]); and Evaluation Benchmarks and Datasets provides resources to measure progress, including general-purpose suites (MMMU-Pro[13]) and specialized task benchmarks that assess targeted capabilities.

Within the evaluation landscape, a particularly active line of work focuses on specialized task benchmarks that probe specific reasoning skills beyond generic visual question answering. These benchmarks often emphasize the interplay between OCR accuracy and higher-level inference, testing whether models can extract text and then reason about relationships, logic, or context. OCR-Reasoning Benchmark[0] sits squarely in this cluster, designed to evaluate models on tasks that require both precise text recognition and subsequent reasoning steps. It shares thematic ground with Contextual[17] and Mctbench[49], which similarly target nuanced

comprehension of text-rich content, though each emphasizes different facets—contextual understanding versus multi-choice reasoning formats. This specialization reflects a broader trend: as general-purpose models improve, the community increasingly values fine-grained diagnostics that reveal where text-image integration still falls short.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Contextual: Evaluating context-sensitive text-rich visual reasoning in large multimodal models

Authors: Wadhawan, Rohan, Bansal, Hritik, Rohan Wadhawan, et al. (12 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Many real-world tasks require an agent to reason jointly over text and visual objects, (e.g., navigating in public spaces), which we refer to as context-sensitive text-rich visual reasoning. Specifically, these tasks require an understanding of the context in which the text interacts with visual elements within an image. However, there is a lack of existing datasets to benchmark the state-of-the-art multimodal models' capability on context-sensitive text-rich visual reasoning. In this paper, we ...

Relationship Analysis

Both papers belong to the Specialized Task Benchmarks category, focusing on evaluating MLLMs' reasoning capabilities in text-rich image scenarios. While OCR-Reasoning Benchmark provides 1,069 examples with step-by-step reasoning annotations across 6 core reasoning abilities (spatial, numerical, mathematical, enumerative, logical, and multidisciplinary knowledge), Contextual emphasizes context-sensitive reasoning where text and visual elements must be jointly understood, covering 8 real-world scenarios (navigation, shopping, time reading, etc.) with 506 human-crafted instructions. The key difference is that OCR-Reasoning focuses on complex multi-step reasoning processes with detailed annotations, whereas Contextual specifically targets the interaction between textual and visual context that cannot be solved through OCR alone.

---

2. Mctbench: Multimodal cognition towards text-rich visual scenes benchmark

Authors: Shan, Bin, Fei Xiang, Bin Shan, Shi Wei, et al. (19 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

The comprehension of text-rich visual scenes has become a focal point for evaluating Multi-modal Large Language Models (MLLMs) due to their widespread applications. Current benchmarks tailored to the scenario emphasize perceptual capabilities, while overlooking the assessment of cognitive abilities. To address this limitation, we introduce a Multimodal benchmark towards Text-rich visual scenes, to evaluate the Cognitive capabilities of MLLMs through visual reasoning and content-creation tasks (M...

Relationship Analysis

Both papers belong to the Specialized Task Benchmarks category, focusing on evaluating MLLMs in text-rich image scenarios with curated evaluation protocols. They overlap in assessing reasoning capabilities beyond simple text extraction, with both providing structured evaluation frameworks for text-rich visual understanding. However, the original paper (OCR-Reasoning) emphasizes step-by-step reasoning process annotation across 6 core reasoning abilities with 1,069 examples, while the candidate paper (MCTBench) focuses on evaluating cognitive capabilities through perception, reasoning, and content-creation tasks across 8.5k question-answer pairs with automated evaluation pipelines for open-ended generation.

Contributions Analysis

Overall novelty summary. The paper introduces OCR-Reasoning, a benchmark comprising 1,069 human-annotated examples spanning 6 core reasoning abilities and 18 practical tasks in text-rich visual scenarios. It resides in the 'Specialized Task Benchmarks' leaf alongside two sibling papers (Contextual and Mctbench), indicating a focused but not overcrowded research direction. The taxonomy shows 50 papers across the entire field, with this leaf containing only three works, suggesting that specialized benchmarks for text-rich reasoning remain relatively sparse compared to broader multimodal evaluation efforts.

The taxonomy reveals that OCR-Reasoning sits within the 'Evaluation Benchmarks and Datasets' branch, distinct from the 'Comprehensive Multimodal Understanding Benchmarks' leaf (which houses general-purpose suites like MMMU-Pro). Neighboring branches include 'Task-Specific Applications' (covering OCR, text-based VQA, and domain-specific reasoning) and 'Reasoning Mechanisms' (addressing chain-of-thought and multi-stage inference). The benchmark's emphasis on text-rich scenarios connects it to application-focused work in OCR and text localization, yet its evaluation-centric design keeps it separate from those implementation-oriented papers.

Among 30 candidates examined, the dual annotation scheme (reasoning processes plus final answers) encountered 2 refutable candidates, while the systematic definition of text-rich reasoning abilities found 1 refutable candidate. The core benchmark contribution itself showed no clear refutations across 10 examined papers. These statistics suggest that while the overall benchmark concept appears relatively novel within the limited search scope, the dual annotation approach and systematic ability taxonomy have more substantial prior work. The analysis does not claim exhaustive coverage; it reflects patterns among top-30 semantic matches and their citations.

Based on the limited literature search, the benchmark appears to occupy a moderately novel position in specialized text-rich evaluation. The taxonomy structure indicates this is an emerging rather than saturated area, though the dual annotation and systematic ability frameworks show partial overlap with existing work. The analysis covers top-30 candidates and does not account for potentially relevant papers outside this scope or in adjacent subfields.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: OCR-Reasoning benchmark for text-rich image reasoning

Description: The authors introduce OCR-Reasoning, a benchmark containing 1,069 human-annotated examples spanning 6 core reasoning abilities and 18 practical reasoning tasks in text-rich visual scenarios. Unlike existing benchmarks that only provide final answers, this benchmark additionally provides detailed step-by-step reasoning processes for holistic assessment.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. A token-level text image foundation model for document understanding

URL: View paper

Brief Assessment

Token-Level Foundation[52] focuses on building a token-level visual foundation model (TokenFD) and dataset (TokenIT) for document understanding tasks, not on creating a reasoning benchmark with step-by-step annotations for evaluating MLLMs' reasoning capabilities in text-rich scenarios.

---

2. Visuriddles: Fine-grained perception is a primary bottleneck for multimodal large language models in abstract visual reasoning

URL: View paper

Brief Assessment

Visuriddles[57] focuses on abstract visual reasoning tasks (e.g., pattern recognition, spatial reasoning) rather than text-rich image understanding. The candidate addresses fundamentally different reasoning challenges involving abstract graphics, not OCR or text extraction from documents.

---

3. Bliva: A simple multimodal llm for better handling of text-rich visual questions

URL: View paper

Brief Assessment

Bliva[56] focuses on improving multimodal LLM architecture for text-rich visual questions through query embeddings and patch embeddings, not on creating reasoning benchmarks with step-by-step annotations. The paper evaluates on existing benchmarks but does not propose a new benchmark with detailed reasoning processes.

---

4. Ocrbench: on the hidden mystery of ocr in large multimodal models

URL: View paper

Brief Assessment

Ocrbench[60] focuses on evaluating OCR capabilities (text recognition, scene text VQA, document VQA, KIE, HMER) without requiring complex reasoning processes. The original paper's OCR-Reasoning benchmark specifically targets reasoning abilities with step-by-step annotations, representing a distinct evaluation focus.

---

5. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning

URL: View paper

Brief Assessment

Ocrbench v2[54] focuses on evaluating OCR capabilities across diverse text-oriented tasks (recognition, localization, parsing) rather than specifically assessing step-by-step reasoning processes in text-rich scenarios. The original paper's novelty lies in providing detailed reasoning annotations alongside answers for holistic reasoning assessment, which is not the primary focus of Ocrbench v2.

---

6. MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding

URL: View paper

Brief Assessment

MedXpertQA[58] focuses on medical knowledge evaluation with clinical images and patient records, not general text-rich image reasoning tasks like financial reports or invoices that OCR-Reasoning addresses.

---

7. mPLUG-DocOwl: Modularized Multimodal Large Language Model for Document Understanding

URL: View paper

Brief Assessment

mPLUG-DocOwl[53] focuses on document understanding tasks (DocVQA, InfoVQA, table extraction) without providing step-by-step reasoning annotations. The original paper's OCR-Reasoning benchmark uniquely provides detailed reasoning processes for 6 core reasoning abilities across 18 tasks, which is not present in mPLUG-DocOwl[53].

---

8. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning

URL: View paper

Brief Assessment

Multimodal Reasoning Survey[59] is a comprehensive survey paper that reviews existing evaluation protocols and benchmarks for multimodal reasoning. It does not introduce a new benchmark for text-rich image reasoning tasks with step-by-step reasoning annotations.

---

9. Colpali: Efficient document retrieval with vision language models

URL: View paper

Brief Assessment

Colpali[55] focuses on document retrieval using vision-language models for matching queries to document pages, not on evaluating reasoning capabilities through step-by-step processes in text-rich images.

---

10. A survey on benchmarks of multimodal large language models

URL: View paper

Brief Assessment

MLLM Benchmarks Survey[51] is a comprehensive survey paper that reviews existing benchmarks across multiple domains. While it discusses text-rich VQA benchmarks in Section 5.1, it does not present a novel benchmark that would refute the originality of OCR-Reasoning's specific contributions (1,069 human-annotated examples with step-by-step reasoning processes across 6 core abilities and 18 tasks).

Contribution 2: Dual annotation scheme with reasoning processes and final answers

Description: The benchmark provides annotations for both final answers and step-by-step reasoning processes, enabling comprehensive evaluation of MLLMs' reasoning capabilities rather than just answer accuracy. This distinguishes it from existing text-rich image understanding benchmarks that only annotate final answers.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark

URL: View paper

Brief Assessment

MLLM-as-a-Judge[71] focuses on evaluating MLLMs' ability to judge and score other models' responses across multiple evaluation settings (scoring, pair comparison, batch ranking), not on annotating reasoning processes for benchmark construction. The candidate does not address systematic annotation of step-by-step reasoning for vision-language understanding tasks.

---

2. Insight-v: Exploring long-chain visual reasoning with multimodal large language models

URL: View paper

Brief Assessment

Insight-V[66] focuses on generating long-chain reasoning data for multi-modal tasks through a progressive generation pipeline, not on benchmarking with dual annotations for evaluation purposes like the original paper's OCR-Reasoning benchmark.

---

3. Reasoning grasping via multimodal large language model

URL: View paper

Brief Assessment

Reasoning Grasping[77] focuses on robotic grasping tasks with implicit instructions, not on evaluating reasoning processes in vision-language models through dual annotation schemes for benchmarking purposes.

---

4. Llava-cot: Let vision language models reason step-by-step

URL: View paper

Brief Assessment

Llava-COT[76] focuses on training vision-language models to autonomously generate structured reasoning in four stages (summary, caption, reasoning, conclusion), rather than providing benchmark annotations with both reasoning processes and final answers for evaluation purposes.

---

5. Measuring and improving chain-of-thought reasoning in vision-language models

URL: View paper

Prior Art Analysis

Chain-of-Thought Reasoning[69] demonstrates prior work that provides annotations for both final answers and step-by-step reasoning processes in vision-language evaluation. The candidate paper explicitly describes creating a benchmark (CURE) that contains 'cot reasoning chains' consisting of 'progressive subquestions' to evaluate both the final inference and intermediate reasoning steps. This dual annotation approach predates the original paper's claim of being the first to provide such comprehensive annotations.

Evidence

Evidence 1 - **Rationale**: Both papers provide annotations for final answers and step-by-step reasoning chains. The candidate explicitly describes creating 'cot reasoning chains' with 'progressive subquestions' for evaluation, which directly parallels the original's claim of annotating both 'final answers and the step-by-step reasoning process.' - **Original**: unlike existing text-rich image understanding benchmarks that only annotate the final answers, ocr-reasoning provides annotations for both the final answers and the step-by-step reasoning process. this comprehensive annotation scheme facilitates a more in-depth evaluation of mllms' reasoning capabil... - **Candidate**: based on an existing coarse-grained visual inference dataset sherlock (hessel et al., 2022), we establish a benchmark cure for chain-ofthought visual reasoning evaluation. it contains 1,622 human-verified samples of high-level visual inference and corresponding cot reasoning chains, intended for zer...

Evidence 2 - **Rationale**: The candidate paper explicitly describes adding annotations for both reasoning chains (step-by-step process) and final answers (candidate answers), demonstrating that this dual annotation approach existed prior to the original paper's publication. - **Original**: furthermore, unlike other text-rich image understanding benchmarks that only annotate the final answers, ocr-reasoning provides annotations for both the final answers and the step-by-step reasoning process. this comprehensive annotation scheme facilitates a more in-depth evaluation of mllms' reasoni... - **Candidate**: we thus add two new annotations to enable this: (1) reasoning chains: we provide fine-grained and precise cot reasoning containing coherent subquestions that can be chained together to derive the high-level inference provided by sherlock. (2) candidate answers: to avoid the long-standing issues in t...

---

6. End-to-end chart summarization via visual chain-of-thought in vision-language models

URL: View paper

Brief Assessment

Chart Summarization[74] focuses on generating textual summaries from chart images using visual chain-of-thought, not on evaluating reasoning processes versus final answers in MLLMs. The candidate addresses chart-to-text generation, while the original paper evaluates MLLM reasoning capabilities on text-rich images with dual annotations.

---

7. Commonsense reasoning for legged robot adaptation with vision-language models

URL: View paper

Brief Assessment

Commonsense Legged Robot[73] focuses on using vision-language models for legged robot locomotion and behavior selection, not on benchmarking or evaluating reasoning processes in MLLMs with dual annotations.

---

8. Vision-r1: Incentivizing reasoning capability in multimodal large language models

URL: View paper

Brief Assessment

Vision-R1[70] focuses on constructing multimodal CoT datasets for training reasoning MLLMs via RL, not on benchmark evaluation methodology. The paper does not propose a benchmark with dual annotations for evaluating both reasoning processes and final answers.

---

9. Visual cognition in multimodal large language models

URL: View paper

Brief Assessment

Visual Cognition[72] evaluates multimodal models on cognitive tasks but does not provide dual annotations of reasoning processes and final answers. The paper focuses on comparing model outputs to human judgments in cognitive domains, not on benchmark annotation methodology.

---

10. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models

URL: View paper

Prior Art Analysis

Visulogic[75] demonstrates that providing dual annotations (reasoning processes and final answers) for visual reasoning benchmarks was already established prior to the ORIGINAL paper. The candidate paper explicitly states that their benchmark includes both step-by-step

reasoning processes and final answers for comprehensive evaluation, using the same dual annotation approach claimed as novel by the ORIGINAL paper. Both papers use this annotation scheme to enable evaluation beyond just answer accuracy, making the ORIGINAL paper's novelty claim questionable.

Evidence

Evidence 1 - **Rationale**: This establishes that Visulogic[75] is focused on reasoning evaluation with detailed processes, setting the foundation for their dual annotation approach. - **Original**: unlike existing text-rich image understanding benchmarks that only provide a final answer, this benchmark additionally provides a detailed step-by-step reasoning process. this dual annotation enables the evaluation of both the models' final answers and their reasoning processes, thereby offering a h... - **Candidate**: reasoning, as fundamental component of human intelligence, has become a critical criterion in evaluating progress toward artificial general intelligence (agi) [26, 74]. recent advancements in large language models (llms) have demonstrated substantial improvements in reasoning capabilities across co...

Evidence 2 - **Rationale**: Visulogic[75] describes their benchmark as evaluating visual reasoning with comprehensive task construction, implying detailed evaluation beyond just final answers. - **Original**: furthermore, unlike other text-rich image understanding benchmarks that only annotate the final answers, ocr-reasoning provides annotations for both the final answers and the step-by-step reasoning process. this comprehensive annotation scheme facilitates a more in-depth evaluation of mllms' reasoni... - **Candidate**: we propose visulogic, a novel benchmark specifically designed to evaluate visual reasoning abilities in multimodal models without mixing them with purely text-based reasoning (see figure 3). visulogic comprises carefully constructed tasks that span multiple reasoning categories (see figure 1). as sh...

Evidence 3 - **Rationale**: This demonstrates that Visulogic[75] evaluates both reasoning processes and final answers, as they assess whether models can solve tasks through proper reasoning rather than shortcuts. - **Original**: notably, while existing benchmarks (mathew et al., 2022; masry et al., 2022; liu et al., 2024d) focus solely on final answers, ocr-reasoning provides annotations for both the final answers and the - **Candidate**: we conducted a comprehensive evaluation and systematic analysis to assess current models' visual reasoning capabilities. when leading text-only llms were supplied with detailed descriptions in place of raw images, their accuracy-doubao-1.5-pro (26.6%), claude-3.7-sonnet (25.9%) and qwen2.5-72b-instr...

Contribution 3: Systematic definition and evaluation of text-rich image reasoning abilities

Description: The authors claim to be the first to concretely define various core sub-abilities (6 core reasoning abilities across 18 tasks) for text-rich image reasoning and provide a systematic evaluation framework. This addresses the gap in existing benchmarks that lack systematic assessment of reasoning capabilities in text-rich visual contexts.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi

URL: View paper

Brief Assessment

MMMU[67] focuses on college-level multimodal understanding across diverse disciplines with heterogeneous image types, not specifically on systematically defining core sub-abilities for text-rich image reasoning tasks as claimed by the original paper.

---

2. Insight-v: Exploring long-chain visual reasoning with multimodal large language models

URL: View paper

Brief Assessment

Insight-V[66] addresses general visual reasoning across diverse multi-modal tasks, not specifically text-rich image reasoning with defined sub-abilities (6 core reasoning abilities across 18 tasks) as claimed in the original paper.

---

3. Measuring multimodal mathematical reasoning with math-vision dataset

URL: View paper

Brief Assessment

Math-Vision[63] focuses on mathematical reasoning with visual contexts across 16 mathematical disciplines, not text-rich image reasoning. The candidate addresses multimodal mathematical problems from competitions, while the original paper targets text-rich scenarios like financial reports and invoices.

---

4. Llavar: Enhanced visual instruction tuning for text-rich image understanding

URL: View paper

Brief Assessment

Llavar[1] focuses on enhancing visual instruction tuning for text-rich images through data collection and model training, not on systematically defining core reasoning sub-abilities or creating evaluation frameworks for reasoning capabilities.

---

5. Spatialrgpt: Grounded spatial reasoning in vision-language models

URL: View paper

Brief Assessment

SpatialRGPT[61] focuses on spatial reasoning in vision-language models with 3D scene understanding, not on defining and evaluating reasoning abilities specifically for text-rich images. The candidate addresses spatial arrangements and geometric relationships rather than text-rich visual contexts like documents, charts, or invoices that the original paper targets.

---

6. Measuring and improving chain-of-thought reasoning in vision-language models

URL: View paper

Brief Assessment

Chain-of-Thought Reasoning[69] focuses on general vision-language reasoning consistency across diverse visual scenarios, not specifically on text-rich image reasoning with defined sub-abilities across practical tasks like financial analysis or document understanding.

---

7. Vla-r1: Enhancing reasoning in vision-language-action models

URL: View paper

Brief Assessment

VLA-R1[68] focuses on vision-language-action models for robotic manipulation tasks, not text-rich image reasoning. The candidate addresses affordance perception and trajectory prediction in embodied AI, which is a fundamentally different domain from evaluating reasoning capabilities in text-rich visual contexts.

---

8. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models

URL: View paper

Brief Assessment

Cot-VLA[62] focuses on vision-language-action models for robotic manipulation tasks, not on defining and evaluating reasoning abilities in text-rich visual understanding. The candidate addresses visual chain-of-thought reasoning for robot control, which is a fundamentally different domain from text-rich image reasoning benchmarks.

---

9. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts

URL: View paper

Prior Art Analysis

MathVista[65] demonstrates prior work that systematically defined and evaluated reasoning abilities in visual contexts. The candidate paper explicitly states it identifies seven types of mathematical reasoning abilities and provides a comprehensive evaluation framework across multiple tasks and visual contexts. Both papers claim to be first in systematically defining core reasoning abilities and providing structured evaluation frameworks for visual reasoning tasks, with MathVista[65] published in 2024 at ICLR, predating the original paper's submission.

Evidence

Evidence 1 - **Rationale**: Both papers claim to be first in systematically defining core reasoning abilities. MathVista[65] explicitly identifies seven mathematical reasoning types, demonstrating prior systematic categorization of reasoning abilities in visual contexts. - **Original**: to the best of our knowledge, we are the first to concretely define various core sub-abilities for text-rich image reasoning and conduct systematic evaluations. - **Candidate**: we identify seven types of mathematical reasoning: algebraic reasoning, arithmetic reasoning, geometry reasoning, logical reasoning, numeric common sense, scientific reasoning, and statistical reasoning

Evidence 2 - **Rationale**: Both papers present comprehensive benchmarks with systematically categorized reasoning abilities. MathVista[65] provides a larger-scale systematic evaluation framework covering multiple reasoning types and tasks. - **Original**: ocr-reasoning comprises 1,069 human-annotated examples spanning 6 core reasoning abilities and 18 practical reasoning tasks in text-rich visual scenarios. - **Candidate**: mathvista consists of 6,141 examples, derived from 28 existing multimodal datasets involving mathematics and 3 newly created datasets

Evidence 3 - **Rationale**: Both papers present benchmarks designed for systematic assessment of reasoning capabilities in visual contexts, with MathVista[65] establishing this approach in 2024. - **Original**: we propose ocr-reasoning, a novel benchmark designed to systematically assess multimodal large language models on text-rich image reasoning tasks. - **Candidate**: we present mathvista, a benchmark designed to combine challenges from diverse mathematical and visual tasks... completing these tasks requires fine-grained, deep visual understanding and compositional reasoning

---

10. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models

URL: View paper

Brief Assessment

NavGPT-2[64] focuses on vision-and-language navigation in physical environments, not text-rich image reasoning. The candidate addresses navigational reasoning in 3D spaces with instruction-following, while the original paper evaluates reasoning capabilities specifically for text-rich visual content across multiple document types and scenarios.

Appendix: Text Similarity Detection

Textual similarity detection checked 30 papers and found 5 similarity segment(s) across 3 paper(s).

The following **3 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts

Detected in: Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

---

2. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models

Detected in: Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

---

3. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi

Detected in: Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] OCR-Reasoning Benchmark: Unveiling the True Capabilities of MLLMs in Complex Text-Rich Image Reasoning View paper
- [1] Llavar: Enhanced visual instruction tuning for text-rich image understanding View paper
- [2] Videochat: Chat-centric video understanding View paper
- [3] Enhancing advanced visual reasoning ability of large language models View paper
- [4] Scaling text-rich image understanding via code-guided synthetic multimodal data generation View paper
- [5] Textual-Visual Logic Challenge: Understanding and Reasoning in Text-to-Image Generation View paper
- [6] Textcot: Zoom in for enhanced multimodal text-rich image understanding View paper
- [7] T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation View paper
- [8] Multimodal large language models for text-rich image understanding: A comprehensive review View paper
- [9] Monet: Reasoning in Latent Visual Space Beyond Images and Language View paper
- [10] Image understanding using vision and reasoning through scene description graph View paper
- [11] Joint visual semantic reasoning: Multi-stage decoder for text recognition View paper
- [12] Mmmg: A massive, multidisciplinary, multi-tier generation benchmark for text-to-image reasoning View paper

- [13] MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark View paper
- [14] Understand, Think, and Answer: Advancing Visual Reasoning with Large Multimodal Models View paper
- [15] Medisee: Reasoning-based pixel-level perception in medical images View paper
- [16] Enhancing Spatial Reasoning through Visual and Textual Thinking View paper
- [17] Contextual: Evaluating context-sensitive text-rich visual reasoning in large multimodal models View paper
- [18] Multi-modal graph neural network for joint reasoning on vision and scene text View paper
- [19] VisualWebInstruct: Scaling up Multimodal Instruction Data through Web Search View paper
- [20] Improving visual grounding with visual-linguistic verification and iterative reasoning View paper
- [21] Prima: Multi-image vision-language models for reasoning segmentation View paper
- [22] Visual chain-of-thought prompting for knowledge-based visual reasoning View paper
- [23] DeepEyes: Incentivizing "Thinking with Images" via Reinforcement Learning View paper
- [24] Heterogeneous Prompt-Guided Entity Inferring and Distilling for Scene-Text Aware Cross-Modal Retrieval View paper
- [25] Enhanced visual instruction tuning for text-rich image understanding View paper
- [26] Visual semantics allow for textual reasoning better in scene text recognition View paper
- [27] Vgr: Visual grounded reasoning View paper
- [28] Reasoning elicitation and multi-granularity contrastive learning for text-rich image understanding in large vision-language models View paper
- [29] VCR: A Task for Pixel-Level Complex Reasoning in Vision Language Models via Restoring Occluded Text View paper
- [30] WeThink: Toward General-purpose Vision-Language Reasoning via Reinforcement Learning View paper
- [31] Mixed Signals: Decoding VLMs' Reasoning and Underlying Bias in Vision-Language Conflict View paper
- [32] FLUX-Reason-6M & PRISM-Bench: A Million-Scale Text-to-Image Reasoning Dataset and Comprehensive Benchmark View paper
- [33] Multimodal self-instruct: Synthetic abstract image and visual reasoning instruction using language model View paper
- [34] Leopard: A vision language model for text-rich multi-image tasks View paper
- [35] Enhancing Emotion Reasoning for Image Multi-Emotion Prediction View paper
- [36] MathReal: We Keep It Real! A Real Scene Benchmark for Evaluating Math Reasoning in Multimodal Large Language Models View paper
- [37] Omni-RGPT: Unifying Image and Video Region-level Understanding via Token Marks View paper
- [38] Webwatcher: Breaking new frontier of vision-language deep research agent View paper
- [39] Reasoning and question answering about image-text multi-modal contexts View paper
- [40] DiffUHaul: A Training-Free Method for Object Dragging in Images View paper
- [41] Structextv3: An efficient vision-language model for text-rich image perception, comprehension, and beyond View paper
- [42] Weakly-supervised 3d spatial reasoning for text-based visual question answering View paper
- [43] WorldGenBench: A World-Knowledge-Integrated Benchmark for Reasoning-Driven Text-to-Image Generation View paper
- [44] Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination View paper
- [45] Think Twice Before You Judge: Mixture of Dual Reasoning Experts for Multimodal Sarcasm Detection View paper
- [46] Enhancing sceneâ□□text visual question answering with relational reasoning, attention and dynamic vocabulary integration View paper
- [47] Mme-videoocr: Evaluating ocr-based capabilities of multimodal llms in video scenarios View paper
- [48] Reasoning-OCR: Can Large Multimodal Models Solve Complex Logical Reasoning Problems from OCR Cues? View paper
- [49] Mctbench: Multimodal cognition towards text-rich visual scenes benchmark View paper
- [50] Vtqa: Visual text question answering via entity alignment and cross-media reasoning View paper
- [51] A survey on benchmarks of multimodal large language models View paper
- [52] A token-level text image foundation model for document understanding View paper
- [53] mPLUG-DocOwl: Modularized Multimodal Large Language Model for Document Understanding View paper
- [54] Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning View paper
- [55] Colpali: Efficient document retrieval with vision language models View paper
- [56] Bliva: A simple multimodal llm for better handling of text-rich visual questions View paper
- [57] Visuriddles: Fine-grained perception is a primary bottleneck for multimodal large language models in abstract visual reasoning View paper
- [58] MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding View paper
- [59] Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning View paper
- [60] Ocrbench: on the hidden mystery of ocr in large multimodal models View paper
- [61] Spatialrgpt: Grounded spatial reasoning in vision-language models View paper
- [62] Cot-vla: Visual chain-of-thought reasoning for vision-language-action models View paper
- [63] Measuring multimodal mathematical reasoning with math-vision dataset View paper
- [64] Navgpt-2: Unleashing navigational reasoning capability for large vision-language models View paper
- [65] Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts View paper
- [66] Insight-v: Exploring long-chain visual reasoning with multimodal large language models View paper
- [67] Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi View paper
- [68] Vla-r1: Enhancing reasoning in vision-language-action models View paper
- [69] Measuring and improving chain-of-thought reasoning in vision-language models View paper
- [70] Vision-r1: Incentivizing reasoning capability in multimodal large language models View paper
- [71] Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark View paper
- [72] Visual cognition in multimodal large language models View paper
- [73] Commonsense reasoning for legged robot adaptation with vision-language models View paper
- [74] End-to-end chart summarization via visual chain-of-thought in vision-language models View paper
- [75] Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models View paper
- [76] Llava-cot: Let vision language models reason step-by-step View paper
- [77] Reasoning grasping via multimodal large language model View paper