Novelty Assessment Report

Paper: HSSBench: Benchmarking Humanities and Social Sciences Ability for Multimodal Large Language Models
PDF URL: https://openreview.net/pdf?id=iQsKotob31
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated significant potential to advance a broad range of domains. However, current benchmarks for evaluating MLLMs primarily emphasize general knowledge and vertical step-by-step reasoning typical of STEM disciplines, while overlooking the distinct needs and potential of the Humanities and Social Sciences (HSS). Tasks in the HSS domain require more horizontal, interdisciplinary thinking and a deep integration of knowledge across related fields, which presents unique challenges for MLLMs, particularly in linking abstract concepts with corresponding visual representations. Addressing this gap, we present HSSBench, a dedicated benchmark designed to assess the capabilities of MLLMs on HSS tasks in multiple languages, including the six official languages of the United Nations. We also introduce a novel data generation pipeline tailored for HSS scenarios, in which multiple domain experts and automated agents collaborate to generate and iteratively refine each sample. HSSBench contains over 13,000 meticulously designed samples, covering six key categories. We benchmark more than 20 mainstream MLLMs on HSSBench and demonstrate that it poses significant challenges even for state-of-the-art models. We hope that this benchmark will inspire further research into enhancing the cross-disciplinary reasoning abilities of MLLMs, especially their capacity to internalize and connect knowledge across fields.

Core Task Landscape

This paper addresses: **Evaluating Multimodal Large Language Models on Humanities and Social Sciences Tasks**

A total of **50 papers** were analyzed and organized into a taxonomy with **16 categories**.

Taxonomy Overview

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

- **Domain-Specific Benchmark Development**
- **Cross-Disciplinary Evaluation**
- **Applied Domain Analysis**
- **Methodological Frameworks and Approaches**
- **Model Behavior and Bias Analysis**
- **Theoretical and Interdisciplinary Perspectives**

Complete Taxonomy Tree

- Evaluating Multimodal Large Language Models on Humanities and Social Sciences Tasks Survey Taxonomy
- Domain-Specific Benchmark Development
  - Cultural and Historical Knowledge Assessment (5 papers)
  - [4] Benchmarking Vision-Language and Multimodal Large Language Models in Zero-shot and Few-shot Scenarios: A study on Christian Iconography (Spinaci, 2025) View paper
  - [7] On path to multimodal historical reasoning: Histbench and histagent (Qiu Jia-Hao, 2025) View paper
  - [25] Benchmarking Vision Language Models for Cultural Understanding (Agrawal, 2024) View paper
  - [29] FoodieQA: A Multimodal Dataset for Fine-Grained Understanding of Chinese Food Culture (Elliott, 2024) View paper
  - [40] ZharfSima: Benchmarking Vision Language Models for Persian Language and Culture (Ali Edalat, 2025) View paper
  - Humanities and Social Sciences Task Benchmarks ★ (5 papers)
  - [0] HSSBench: Benchmarking Humanities and Social Sciences Ability for Multimodal Large Language Models (Anon et al., 2026) View paper
  - [5] MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI (Xiang Yue, 2023) View paper
  - [16] Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models (Simeon Hristov, 2024) View paper
  - [17] MMVU: Measuring Expert-Level Multi-Discipline Video Understanding (Yilun Zhao, 2025) View paper
  - [38] CMMMU: A Chinese Massive Multi-discipline Multimodal Understanding Benchmark (Zhang Ge, 2024) View paper
  - Social and Behavioral Understanding Benchmarks (3 papers)
  - [20] Humanibench: A human-centric framework for large multimodal models evaluation (Raza, 2025) View paper
  - [22] D-HUMOR: Dark Humor Understanding via Multimodal Open-ended Reasoning--A Benchmark Dataset and Method (Nagendra, 2025) View paper
  - [45] Evaluating Vision-Language Models on the TriangleCOPA Benchmark (Ankur Chemburkar, 2024) View paper
- Cross-Disciplinary Evaluation
  - Multi-Discipline Expert-Level Assessment (3 papers)
  - [13] Vlind-bench: Measuring language priors in large vision-language models (Kang Il Lee, 2025) View paper
  - [24] Vhelm: A holistic evaluation of vision language models (Tony Lee, 2024) View paper
  - [47] ProBench: Judging Multimodal Foundation Models on Open-ended Multi-domain Expert Tasks (Yan Yang, 2025) View paper

    - Language and Cultural Contextualization (2 papers)
    - [26] Measuring Hong Kong Massive Multi-Task Language Understanding (Zhu Zheng-hao, 2025) View paper
    - [27] Pangea: A fully open multilingual multimodal llm for 39 languages (Yue Xiang, 2024) View paper
    - Specialized Task and Reasoning Evaluation (2 papers)
    - [8] MMCR: Advancing Visual Language Model in Multimodal Multi-Turn Contextual Reasoning (Yan Dawei, 2025) View paper
    - [14] A multi-modal assessment framework for comparison of specialized deep learning and general-purpose large language models (Mohammad Nadeem, 2025) View paper
- Applied Domain Analysis
    - Social Media and Online Content Analysis (3 papers)
    - [1] Mm-soc: Benchmarking multimodal large language models in social media platforms (Choi Minje, 2024) View paper
    - [28] From experts to the public: Governing multimodal language models in politically sensitive video analysis (Tanusree Sharma, 2024) View paper
    - [43] GPT-4V(ision) as A Social Media Analysis Engine (Hanjia Lyu, 2024) View paper
    - Urban and Spatial Analysis (4 papers)
    - [10] Interpretable Multimodal Framework for Human-Centered Street Assessment: Integrating Visual-Language Models for Perceptual Urban Diagnostics (Lan, 2025) View paper
    - [11] Generative Multimodal Models for Social Science: An Application with Satellite and Streetscape Imagery (Tina Law, 2025) View paper
    - [15] Perceiving urban inequality from imagery using visual language models with chain-of-thought reasoning (Yunke Zhang, 2025) View paper
    - [19] Evaluating urban visual attractiveness perception using multimodal large language model and street view images (Qianyu Zhou, 2025) View paper
    - Historical Document Processing and Digitization (4 papers)
    - [12] Evaluating LLMs for Historical Document OCR: A Methodological Framework for Digital Humanities (Levchenko, 2025) View paper
    - [18] Multimodal-LLM as A Reliable Tool for Information Extraction from Historical Documents: A Digital Humanities Approach to Swedish Patent Cards (1945-1975) (Luo, 2025) View paper
    - [31] Multimodal LLM-assisted Information Extraction from Historical Documents: The Case of Swedish Patent Cards (1945-1975) and ChatGPT (Yunting Xie, 2025) View paper
    - [50] Handwriting recognition in historical documents with multimodal llm (Lucian Li, 2024) View paper
    - Cultural Heritage and Artifact Analysis (5 papers)
    - [6] XunZi-MLLM: a multimodal large language model for ancient text and image recognition (Dongmei Zhu, 2025) View paper
    - [33] Gallerygpt: Analyzing paintings with large multimodal models (Bin Yi, 2024) View paper
    - [36] How Can Generative Artificial Intelligence Techniques Facilitate Intelligent Research into Ancient Books? (Jiangfeng Liu, 2024) View paper
    - [37] Explainable Search and Discovery of Visual Cultural Heritage Collections with Multimodal Large Language Models (Arnold Taylor, 2024) View paper
    - [46] HistoLens: An LLM-Powered Framework for Multi-Layered Analysis of Historical Texts--A Case Application of Yantie Lun (Zeng, 2024) View paper
- Methodological Frameworks and Approaches
    - Research Automation and Augmentation Frameworks (3 papers)
    - [2] Machine-assisted quantitizing designs: augmenting humanities and social sciences with artificial intelligence (Andres Karjus, 2025) View paper
    - [3] Bridging Technology and Humanities: Evaluating the Impact of Large Language Models on Social Sciences Research with DeepSeek-R1 (Peiran Gu, 2025) View paper
    - [30] A multimodal framework embedding retrieval-augmented generation with mllms for eurobarometer data (George Papageorgiou, 2025) View paper
    - Multi-Agent and Interactive Systems (4 papers)
    - [9] Beyond Static Responses: Multi-Agent LLM Systems as a New Paradigm for Social Science Research (Haase, 2025) View paper
    - [23] Stage Wizard: Enhancing Tangible Storytelling with Multimodal LLMs (Kuntong Han, 2025) View paper
    - [34] Lmagent: A large-scale multimodal agents society for multi-user simulation (Liu Yi-jun, 2024) View paper
    - [35] CAFES: A Collaborative Multi-Agent Framework for Multi-Granular Multimodal Essay Scoring (Su Jiamin, 2025) View paper
    - Corpus Annotation and Multimodal Analysis (1 papers)
    - [44] Can Multimodal Large Language Models Annotate Low- and High-Level Features of Multimodal Artefacts? (Rosa Suviranta, 2025) View paper
- Model Behavior and Bias Analysis
    - Social Bias and Fairness Evaluation (2 papers)
    - [41] VLBiasBench: A Comprehensive Benchmark for Evaluating Bias in Large Vision-Language Model (Wang Sibo, 2024) View paper
    - [42] Data matters most: Auditing social bias in contrastive visionâ□□language models (Sahili, 2025) View paper
    - Trustworthiness and Ethical Alignment (1 papers)
    - [48] Applied with Caution: Extreme-Scenario Testing Reveals Significant Risks in Using LLMs for Humanities and Social Sciences Paper Evaluation (Hua Liu, 2025) View paper
- Theoretical and Interdisciplinary Perspectives (4 papers)
    - [21] Large Language Models: A historical and sociocultural perspective (Eugene Yu Ji, 2024) View paper
    - [32] Multimodal Vision Language Models in Interactive and Physical Environments (L Pereira, 2025) View paper
    - [39] From ChatGPT, DALL-E 3 to Sora: How has Generative AI Changed Digital Humanities Research and Services? (Liu Jiang-feng, 2024) View paper
    - [49] Benchmarking Multimodal Large Language Models in Zero-shot and Few-shot Scenarios: Preliminary Results on Studying Christian Iconography (G Spinaci, 2025) View paper

Narrative

Core task: Evaluating multimodal large language models on humanities and social sciences tasks. The field has organized itself around several complementary directions. Domain-Specific Benchmark Development focuses on creating targeted evaluation suites for humanities and social sciences, such as MMMU[5], Exams-v[16], and MMVU[17], which test models on discipline-specific knowledge and reasoning. Cross-Disciplinary Evaluation examines how models perform across multiple domains simultaneously, often revealing gaps in

cultural or contextual understanding. Applied Domain Analysis investigates real-world applications in areas like urban studies, historical document processing, and cultural heritage, with works such as Urban Inequality Imagery[15] and HistBench HistAgent[7] demonstrating practical use cases. Methodological Frameworks and Approaches develop systematic ways to probe model capabilities, while Model Behavior and Bias Analysis scrutinizes fairness and representation issues through benchmarks like VLBiasBench[41]. Theoretical and Interdisciplinary Perspectives bridge computational methods with humanistic inquiry, as seen in Bridging Technology Humanities[3].

A particularly active line of work centers on comprehensive benchmarking that spans multiple humanities and social sciences disciplines, balancing breadth with depth of evaluation. HSSBench[0] exemplifies this approach by providing a broad assessment framework across diverse humanistic tasks, positioning itself alongside other general-purpose benchmarks like MMMU[5] and CMMMU[38] but with a stronger emphasis on social sciences and humanities-specific reasoning. In contrast, works such as Christian Iconography Benchmark[4] or HistBench HistAgent[7] pursue narrower, domain-expert-level evaluations within specialized subfields. The tension between creating widely applicable benchmarks versus deeply specialized assessments remains a central question, as does the challenge of capturing culturally situated knowledge that varies across linguistic and geographic contexts—a concern highlighted by efforts like Pangea[27] and Cultural Understanding Benchmark[25].

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

Authors: Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, et al. (24 authors total) | **Year/Venue**: 2023 • Computer Vision and Pattern Recognition | **URL**: View paper

Abstract

We introduce MMMU: a new benchmark designed to evaluate multimodal models on massive multi-discipline tasks demanding college-level subject knowledge and deliberate reasoning. MMMU includes 11.5K meticulously collected multimodal questions from college exams, quizzes, and text-books, covering six core disciplines: Art & Design, Busi-ness, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering. These questions span 30 subjects and 183 subfields, comprising 30 highly het-e...

Relationship Analysis

Both papers belong to the Humanities and Social Sciences Task Benchmarks category, developing comprehensive multimodal evaluation datasets that assess MLLMs across multiple disciplines including humanities and social sciences domains. While both benchmarks include humanities and social sciences as part of their evaluation scope, HSSBench focuses exclusively and deeply on HSS tasks across 6 categories and 45 types with multilingual support in UN languages, whereas MMMU takes a broader multi-discipline approach covering 6 core disciplines (including but not limited to HSS) with 30 subjects spanning art, business, science, health, and engineering alongside humanities.

---

2. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models

Authors: Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, Preslav Nakov | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â¦ ural science, social science, and other miscellaneous studies, â¦ evaluate the performance of state-of-theart large language â¦ models on massive multi-discipline tasks demanding college-â¦

Relationship Analysis

Both papers belong to the Humanities and Social Sciences Task Benchmarks category, developing comprehensive multimodal evaluation datasets for assessing MLLMs across diverse academic disciplines. They overlap in their focus on creating multilingual, multi-discipline benchmarks that test models' ability to reason over combined text and visual content in educational contexts, including subjects like history, geography, economics, and social sciences. However, HSSBench emphasizes horizontal reasoning and cross-disciplinary knowledge integration specific to HSS domains with a novel expert-agent collaborative data generation pipeline, while EXAMS-V focuses on standardized exam questions with interleaved multimodal elements (text, images, tables embedded together) across 11 languages from actual state examinations.

---

3. MMVU: Measuring Expert-Level Multi-Discipline Video Understanding

Authors: Yilun Zhao, Haowei Zhang, Lujing Xie, Guo Gan, Yitao Long, et al. (21 authors total) | **Year/Venue**: 2025 • Computer Vision and Pattern Recognition | **URL**: View paper

Abstract

We introduce $\color{Blue}{\text{MMVU}}$, a comprehensive expert-level, multi-discipline benchmark for evaluating foundation models in video understanding. $\color{Blue}{\text{MMVU}}$ includes 3,000 expert-annotated questions spanning 27 subjects across four core disciplines: Science, Healthcare, Humanities & Social Sciences, and Engineering. Compared to prior benchmarks, $ \color{Blue}{\text{MMVU}}$ features three key advancements. First, it challenges models to apply domain-specific knowledge a...

Relationship Analysis

Both papers belong to the Humanities and Social Sciences Task Benchmarks category, developing comprehensive evaluation datasets for multimodal models in specialized domains. While HSSBench focuses exclusively on humanities and social sciences tasks (geography, art, culture, history, economics, social sciences) with 13,152 samples across 45 types in six languages, MMVU takes a broader multi-discipline approach covering science, healthcare, humanities & social sciences, and engineering with 3,000 expert-annotated video understanding questions across 27 subjects. The key distinction is that HSSBench emphasizes static image-based VQA for HSS domains with horizontal reasoning requirements, whereas MMVU focuses on expert-level video understanding across multiple disciplines including but not limited to humanities and social sciences.

---

4. CMMMU: A Chinese Massive Multi-discipline Multimodal Understanding Benchmark

Authors: Zhang Ge, Du Xinrun, Ge Zhang, Chen Bei, Xinrun Du, et al. (47 authors total) | **Year/Venue**: 2024 • arXiv.org | **URL**: View paper

Abstract

As the capabilities of large multimodal models (LMMs) continue to advance, evaluating the performance of LMMs emerges as an increasing need. Additionally, there is an even larger gap in evaluating the advanced knowledge and reasoning abilities of LMMs in non-English contexts such as Chinese. We introduce CMMMU, a new Chinese Massive Multi-discipline Multimodal Understanding benchmark designed to evaluate LMMs on tasks demanding college-level subject knowledge and deliberate reasoning in a Chines...

Relationship Analysis

Both papers belong to the Humanities and Social Sciences Task Benchmarks category, developing comprehensive multimodal evaluation datasets for assessing LMMs on HSS-related tasks. While HSSBench focuses on six HSS categories (geography, art, culture, social sciences, history, economics) with multilingual support across UN languages and emphasizes cross-modal knowledge transfer challenges, CMMMU specifically targets Chinese-language evaluation across six disciplines including Art&Design, Business, Science, Health&Medicine, Humanities&Social Science, and Tech&Engineering, with a broader scope beyond pure HSS to include STEM domains in a Chinese cultural context.

Contributions Analysis

Overall novelty summary. The paper introduces HSSBench, a benchmark evaluating multimodal large language models on humanities and social sciences tasks across multiple languages. It resides in the 'Humanities and Social Sciences Task Benchmarks' leaf, which contains five papers total, including the original work. This leaf sits within the broader 'Domain-Specific Benchmark Development' branch, indicating a moderately populated research direction focused on creating specialized evaluation suites. The taxonomy reveals that while domain-specific benchmarking is active, this particular leaf represents a concentrated effort to address HSS-specific evaluation needs rather than a crowded subfield.

The taxonomy structure shows that HSSBench's leaf neighbors include 'Cultural and Historical Knowledge Assessment' (five papers) and 'Social and Behavioral Understanding Benchmarks' (three papers), both emphasizing narrower aspects of humanistic evaluation. Nearby branches like 'Cross-Disciplinary Evaluation' contain multi-domain benchmarks (e.g., MMMU, CMMMU) that span STEM and humanities but lack HSS-specific depth. The taxonomy's scope notes clarify that HSSBench's comprehensive HSS focus distinguishes it from purely cultural heritage benchmarks or general expert-level assessments, positioning it at the intersection of breadth and domain specialization within the humanities evaluation landscape.

Among thirty candidates examined, the 'HSSBench benchmark' contribution shows one refutable candidate out of ten examined, suggesting some prior work in comprehensive HSS benchmarking exists but is limited. The 'VQA Generation Pipeline for HSS scenarios' contribution found no refutable candidates among ten examined, indicating potential novelty in data generation methodology. The 'multilingual evaluation' contribution similarly found no refutable candidates among ten examined. These statistics reflect a focused search scope rather than exhaustive coverage, with the single refutable instance likely representing overlap with existing multi-domain benchmarks that include HSS components.

Based on the limited search of thirty candidates, the work appears to occupy a relatively underexplored niche combining comprehensive HSS coverage with multilingual evaluation. The taxonomy context suggests that while related benchmarks exist, few target the specific intersection of broad humanities disciplines, social sciences reasoning, and multilingual assessment. The analysis does not capture potential work outside the top-thirty semantic matches or recent preprints, so the novelty assessment remains provisional pending broader literature review.

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

Contribution 1: HSSBench benchmark for Humanities and Social Sciences

Description: The authors introduce HSSBench, a large-scale multilingual benchmark containing over 13,000 samples across six categories and 45 types, specifically designed to evaluate multimodal large language models on Humanities and Social Sciences tasks that require horizontal, interdisciplinary reasoning rather than vertical STEM-style reasoning.

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

Prior Art Analysis

MMMU[5] demonstrates that a large-scale multimodal benchmark covering Humanities & Social Science tasks already existed prior to HSSBench. MMMU[5] includes 11.5k questions spanning six core disciplines including 'humanities & social science' as one of its major categories, covering 30 subjects and 183 subfields with diverse image types. This shows that the concept of a comprehensive multimodal benchmark specifically evaluating models on HSS tasks was not novel to the original paper's authors.

Evidence

Evidence 1 - **Rationale**: Both papers present large-scale multimodal benchmarks with meticulously collected/designed samples covering multiple disciplines including humanities & social science. MMMU[5]'s inclusion of 'humanities & social science' as a core discipline with 11.5k questions demonstrates prior work in this area. - **Original**: we present hssbench, a dedicated benchmark designed to assess the capabilities of mllms on hss tasks in multiple languages, including the six official languages of the united nations. we also introduce a novel data generation pipeline tailored for hss scenarios, in which multiple domain experts and ... - **Candidate**: we introduce mmmu: a new benchmark designed to evaluate multimodal models on massive multi-discipline tasks demanding college-level subject knowledge and deliberate reasoning. mmmu includes 11.5k meticulously collected multimodal questions from college exams, quizzes, and text-books, covering six co...

Evidence 2 - **Rationale**: MMMU[5] already emphasized advanced perception and reasoning with domain-specific knowledge across diverse image types, including those relevant to HSS tasks (maps, charts, diagrams), demonstrating that the challenge of linking visual representations with abstract domain knowledge was already being addressed. - **Original**: Tasks in the hss domain require more horizontal, interdisciplinary thinking and a deep integration of knowledge across related fields, which presents unique challenges for mllms, particularly in linking abstract concepts with corresponding visual representations. - **Candidate**: these questions span 30 subjects and 183 subfields, comprising 30 highly het-erogeneous image types, such as charts, diagrams, maps, tables, music sheets, and chemical structures. unlike existing benchmarks, mmmu focuses on advanced perception and reasoning with domain-specific knowledge, challengin...

---

2. Digital multimodal composing as translanguaging assessment in CLIL classrooms

URL: View paper

Brief Assessment

Translanguaging Assessment CLIL[51] focuses on digital multimodal composing as assessment in CLIL (Content and Language Integrated Learning) classrooms, not on creating multilingual benchmarks for evaluating multimodal large language models on Humanities and Social Sciences tasks.

---

3. Multilingual and multimodal composition at school: ScribJab in action

URL: View paper

Brief Assessment

ScribJab[54] focuses on multilingual and multimodal composition in educational settings at schools, not on benchmarking multimodal large language models for Humanities and Social Sciences tasks. No full text context was provided for the candidate paper to enable detailed comparison.

---

4. Multimodal composing in multilingual learning and teaching contexts

URL: View paper

Brief Assessment

Multilingual Learning Teaching[53] focuses on multimodal composing and assessment tasks in educational contexts, not on creating benchmarks for evaluating multimodal large language models on Humanities and Social Sciences reasoning tasks.

---

5. Placing multi-modal, and multi-lingual Data in the Humanities Domain on the Map: the Mythotopia Geo-tagged Corpus

URL: View paper

Brief Assessment

Mythotopia Corpus[58] focuses on geotagged multimodal data for cultural heritage and mythology in Northern Greece, not on evaluating multimodal LLMs' reasoning abilities across diverse HSS tasks as HSSBench does.

---

6. Designing a multilingual, multimodal and collaborative platform of resources for higher education

URL: View paper

Brief Assessment

Collaborative Platform Resources[52] focuses on designing a multilingual, multimodal platform for higher education resources, not on creating a benchmark for evaluating multimodal large language models on Humanities and Social Sciences tasks with horizontal reasoning requirements.

---

7. Multilingualism and multimodality in the CLIL/EMI classroom

URL: View paper

Brief Assessment

CLIL EMI Classroom[55] focuses on multilingualism and multimodality in educational contexts (CLIL/EMI classrooms), not on benchmarking multimodal large language models for Humanities and Social Sciences tasks. The domains and objectives are fundamentally different.

---

8. Mlm: a benchmark dataset for multitask learning with multiple languages and modalities

URL: View paper

Brief Assessment

MLM Benchmark[56] focuses on multitask learning systems for human settlements across multiple modalities (text, images, coordinates) in three languages, targeting digital humanities applications like location estimation and entity retrieval. In contrast, HSSBench specifically evaluates multimodal large language models on Humanities and Social Sciences reasoning tasks requiring interdisciplinary knowledge integration across six categories (geography, art, culture, social sciences, history, economy) in six UN languages. The datasets serve fundamentally different purposes and evaluation objectives.

---

9. MMReview: A Multidisciplinary and Multimodal Benchmark for LLM-Based Peer Review Automation

URL: View paper

Brief Assessment

MMReview[59] focuses on peer review automation across multiple academic disciplines including social sciences, but does not specifically target Humanities and Social Sciences evaluation with horizontal, interdisciplinary reasoning tasks as HSSBench does.

---

10. EverydayMMQA: A Multilingual and Multimodal Framework for Culturally Grounded Spoken Visual QA

URL: View paper

Brief Assessment

EverydayMMQA[57] focuses on culturally-grounded everyday knowledge in low-resource languages for spoken visual QA, not on Humanities and Social Sciences academic tasks requiring interdisciplinary reasoning across domains like geography, art, history, and economics.

Contribution 2: VQA Generation Pipeline for HSS scenarios

Description: The authors develop a three-stage data construction pipeline (Dataset Preparation, Dataset Construction, and Validation) that combines domain expert annotation with a multi-agent framework to efficiently generate high-quality visual question answering data tailored to the unique requirements of Humanities and Social Sciences domains.

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. VizGenie: Toward Self-Refining, Domain-Aware Workflows for Next-Generation Scientific Visualization

URL: View paper

Brief Assessment

VizGenie[70] focuses on scientific visualization workflows using VQA for querying generated visualizations, not on generating VQA datasets for humanities and social sciences domains through expert-agent collaboration.

---

2. A question-type guided and progressive self-attention network for remote sensing visual question answering

URL: View paper

Brief Assessment

Remote Sensing VQA[69] focuses on remote sensing visual question answering with a question-type guided prediction pipeline, not on HSS domain data generation combining domain experts with multi-agent frameworks for humanities and social sciences scenarios.

---

3. An AI-Assisted Bridge Inspection System: Ontology-Based Visual Question-Answering Methodology Using Large Language Models

URL: View paper

Brief Assessment

Bridge Inspection System[76] focuses on bridge infrastructure inspection using ontology-based VQA for civil engineering, not humanities and social sciences domains. The technical context and application domain differ fundamentally from the HSS-focused multi-agent pipeline.

---

4. Eagle: Expert-guided self-enhancement for preference alignment in pathology large vision-language model

URL: View paper

Brief Assessment

Eagle[71] focuses on pathology-specific preference alignment for large vision-language models in medical diagnosis, not on VQA data generation for Humanities and Social Sciences domains. The technical approaches and application domains are fundamentally different.

---

5. Toolvqa: A dataset for multi-step reasoning vqa with external tools

URL: View paper

Brief Assessment

ToolVQA[73] focuses on multi-step reasoning VQA with external tools for general domains, not specifically HSS scenarios. Their pipeline uses image-guided DFS with LCS-based matching for tool-use trajectories, while the original paper's pipeline combines domain expert annotation with multi-agent frameworks specifically tailored to HSS requirements.

---

6. MedFrameQA: A Multi-Image Medical VQA Benchmark for Clinical Reasoning

URL: View paper

Brief Assessment

MedFrameQA[72] focuses on multi-image medical VQA with automated frame extraction from videos, while the original paper addresses humanities and social sciences domains with a multi-agent framework combining domain experts and automated agents for data construction.

---

7. VilBias: A Study of Bias Detection through Linguistic and Visual Cues, presenting Annotation Strategies, Evaluation, and Key Challenges

URL: View paper

Brief Assessment

VilBias[74] focuses on bias detection through linguistic and visual cues in multimodal content, not on generating VQA data for Humanities and Social Sciences domains using domain experts and multi-agent frameworks.

---

8. Towards Human-Level Understanding of Complex Process Engineering Schematics: A Pedagogical, Introspective Multi-Agent Framework for Open-Domain Question â¦

URL: View paper

Brief Assessment

Process Engineering Schematics[75] focuses on technical engineering diagrams (P&IDs, PFDs) with VQA and OCR tasks for industrial process understanding, not on humanities and social sciences domains requiring cultural and historical reasoning.

---

9. Drivelm: Driving with graph visual question answering

URL: View paper

Brief Assessment

DriveLM[68] focuses on autonomous driving scenarios with graph-structured visual question answering, not on Humanities and Social Sciences domains. The domain-specific requirements and knowledge types differ fundamentally.

---

10. Explaining CLIP's performance disparities on data from blind/low vision users

URL: View paper

Brief Assessment

CLIP Blind Vision[77] focuses on evaluating CLIP's performance on data from blind/low vision users, not on developing data generation pipelines for visual question answering. The paper addresses fairness and quality-of-service issues in existing models rather than proposing novel data construction methodologies.

Contribution 3: Comprehensive multilingual evaluation of MLLMs on HSS tasks

Description: The authors conduct extensive evaluations of over 20 mainstream multimodal large language models across six languages, revealing that current state-of-the-art models struggle with HSS tasks and demonstrating the benchmark's effectiveness in identifying limitations in cross-disciplinary reasoning and cross-modal knowledge transfer.

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. Language Is Not All You Need: Aligning Perception with Language Models

URL: View paper

Brief Assessment

Aligning Perception Language[63] focuses on training and evaluating a multimodal large language model (KOSMOS-1) on general tasks including language understanding, perception-language tasks, and vision tasks, but does not specifically address humanities and social sciences (HSS) evaluation or multilingual assessment of cross-disciplinary reasoning in HSS domains.

---

2. MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

URL: View paper

Brief Assessment

MMMU[5] evaluated many models but does not explicitly describe multilingual evaluation across six languages or detailed analysis of cross-disciplinary reasoning and cross-modal knowledge transfer as distinct evaluation contributions.

---

3. Position: Multimodal large language models can significantly advance scientific reasoning

URL: View paper

Brief Assessment

Scientific Reasoning Position[61] focuses on scientific reasoning across STEM disciplines (mathematics, physics, chemistry, biology), not humanities and social sciences (HSS) tasks. The domains, evaluation scope, and research objectives are fundamentally different.

---

4. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models

URL: View paper

Brief Assessment

Ddcot[67] focuses on multimodal chain-of-thought prompting for science question answering, not on comprehensive multilingual evaluation of MLLMs across humanities and social sciences disciplines.

---

5. On path to multimodal historical reasoning: Histbench and histagent

URL: View paper

Brief Assessment

HistBench HistAgent[7] focuses specifically on historical reasoning tasks with 414 questions across 29 languages, while the original paper evaluates over 20 MLLMs on a broader HSS benchmark (13,152 samples) covering six categories including geography, art, culture, social sciences, history, and economics. The candidate's narrower historical focus does not challenge the original's comprehensive cross-disciplinary HSS evaluation.

---

6. Multimodal reasoning with multimodal knowledge graph

URL: View paper

Brief Assessment

Multimodal Knowledge Graph[62] focuses on enhancing multimodal reasoning through knowledge graphs for question answering and analogy reasoning tasks, not on comprehensive multilingual evaluation of MLLMs across humanities and social sciences disciplines.

---

7. Multimodal chain-of-thought reasoning: A comprehensive survey

URL: View paper

Brief Assessment

Multimodal Chain-of-Thought Survey[60] is a survey paper focused on chain-of-thought reasoning methodologies across multimodal contexts, not an evaluation benchmark for humanities and social sciences tasks across multiple languages.

---

8. A survey of scientific large language models: From data foundations to agent frontiers

URL: View paper

Brief Assessment

Scientific LLMs Survey[65] focuses on scientific domain LLMs and their data foundations, not on evaluating multimodal models for humanities and social sciences tasks across multiple languages.

---

9. Skywork R1V: Pioneering Multimodal Reasoning with Chain-of-Thought

URL: View paper

Brief Assessment

Skywork R1V[64] focuses on multimodal reasoning model development and evaluation on STEM-oriented benchmarks (MMMU, MathVista, AIME, MATH500), not on humanities and social sciences tasks or multilingual evaluation frameworks.

---

10. Interdisciplinary-QG: An LLM-Based Framework for Generating High-Quality Interdisciplinary Test Questions with Knowledge Graphs and Chain-of-Thought â¦

URL: View paper

Brief Assessment

Interdisciplinary-QG[66] focuses on generating interdisciplinary test questions using knowledge graphs and chain-of-thought prompting, not on evaluating multimodal large language models across languages or assessing their performance on humanities and social sciences tasks.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] HSSBench: Benchmarking Humanities and Social Sciences Ability for Multimodal Large Language Models View paper
- [1] Mm-soc: Benchmarking multimodal large language models in social media platforms View paper
- [2] Machine-assisted quantitizing designs: augmenting humanities and social sciences with artificial intelligence View paper
- [3] Bridging Technology and Humanities: Evaluating the Impact of Large Language Models on Social Sciences Research with DeepSeek-R1 View paper
- [4] Benchmarking Vision-Language and Multimodal Large Language Models in Zero-shot and Few-shot Scenarios: A study on Christian Iconography View paper
- [5] MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI View paper
- [6] XunZi-MLLM: a multimodal large language model for ancient text and image recognition View paper
- [7] On path to multimodal historical reasoning: Histbench and histagent View paper
- [8] MMCR: Advancing Visual Language Model in Multimodal Multi-Turn Contextual Reasoning View paper
- [9] Beyond Static Responses: Multi-Agent LLM Systems as a New Paradigm for Social Science Research View paper
- [10] Interpretable Multimodal Framework for Human-Centered Street Assessment: Integrating Visual-Language Models for Perceptual Urban Diagnostics View paper
- [11] Generative Multimodal Models for Social Science: An Application with Satellite and Streetscape Imagery View paper
- [12] Evaluating LLMs for Historical Document OCR: A Methodological Framework for Digital Humanities View paper
- [13] Vlind-bench: Measuring language priors in large vision-language models View paper
- [14] A multi-modal assessment framework for comparison of specialized deep learning and general-purpose large language models View paper
- [15] Perceiving urban inequality from imagery using visual language models with chain-of-thought reasoning View paper
- [16] Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models View paper
- [17] MMVU: Measuring Expert-Level Multi-Discipline Video Understanding View paper
- [18] Multimodal-LLM as A Reliable Tool for Information Extraction from Historical Documents: A Digital Humanities Approach to Swedish Patent Cards (1945-1975) View paper
- [19] Evaluating urban visual attractiveness perception using multimodal large language model and street view images View paper

- [20] Humanibench: A human-centric framework for large multimodal models evaluation View paper
- [21] Large Language Models: A historical and sociocultural perspective View paper
- [22] D-HUMOR: Dark Humor Understanding via Multimodal Open-ended Reasoning--A Benchmark Dataset and Method View paper
- [23] Stage Wizard: Enhancing Tangible Storytelling with Multimodal LLMs View paper
- [24] Vhelm: A holistic evaluation of vision language models View paper
- [25] Benchmarking Vision Language Models for Cultural Understanding View paper
- [26] Measuring Hong Kong Massive Multi-Task Language Understanding View paper
- [27] Pangea: A fully open multilingual multimodal llm for 39 languages View paper
- [28] From experts to the public: Governing multimodal language models in politically sensitive video analysis View paper
- [29] FoodieQA: A Multimodal Dataset for Fine-Grained Understanding of Chinese Food Culture View paper
- [30] A multimodal framework embedding retrieval-augmented generation with mllms for eurobarometer data View paper
- [31] Multimodal LLM-assisted Information Extraction from Historical Documents: The Case of Swedish Patent Cards (1945-1975) and ChatGPT View paper
- [32] Multimodal Vision Language Models in Interactive and Physical Environments View paper
- [33] Gallerygpt: Analyzing paintings with large multimodal models View paper
- [34] Lmagent: A large-scale multimodal agents society for multi-user simulation View paper
- [35] CAFES: A Collaborative Multi-Agent Framework for Multi-Granular Multimodal Essay Scoring View paper
- [36] How Can Generative Artificial Intelligence Techniques Facilitate Intelligent Research into Ancient Books? View paper
- [37] Explainable Search and Discovery of Visual Cultural Heritage Collections with Multimodal Large Language Models View paper
- [38] CMMMU: A Chinese Massive Multi-discipline Multimodal Understanding Benchmark View paper
- [39] From ChatGPT, DALL-E 3 to Sora: How has Generative AI Changed Digital Humanities Research and Services? View paper
- [40] ZharfSima: Benchmarking Vision Language Models for Persian Language and Culture View paper
- [41] VLBiasBench: A Comprehensive Benchmark for Evaluating Bias in Large Vision-Language Model View paper
- [42] Data matters most: Auditing social bias in contrastive visionâ□□language models View paper
- [43] GPT-4V(ision) as A Social Media Analysis Engine View paper
- [44] Can Multimodal Large Language Models Annotate Low- and High-Level Features of Multimodal Artefacts? View paper
- [45] Evaluating Vision-Language Models on the TriangleCOPA Benchmark View paper
- [46] HistoLens: An LLM-Powered Framework for Multi-Layered Analysis of Historical Texts--A Case Application of Yantie Lun View paper
- [47] ProBench: Judging Multimodal Foundation Models on Open-ended Multi-domain Expert Tasks View paper
- [48] Applied with Caution: Extreme-Scenario Testing Reveals Significant Risks in Using LLMs for Humanities and Social Sciences Paper Evaluation View paper
- [49] Benchmarking Multimodal Large Language Models in Zero-shot and Few-shot Scenarios: Preliminary Results on Studying Christian Iconography View paper
- [50] Handwriting recognition in historical documents with multimodal llm View paper
- [51] Digital multimodal composing as translanguaging assessment in CLIL classrooms View paper
- [52] Designing a multilingual, multimodal and collaborative platform of resources for higher education View paper
- [53] Multimodal composing in multilingual learning and teaching contexts View paper
- [54] Multilingual and multimodal composition at school: ScribJab in action View paper
- [55] Multilingualism and multimodality in the CLIL/EMI classroom View paper
- [56] Mlm: a benchmark dataset for multitask learning with multiple languages and modalities View paper
- [57] EverydayMMQA: A Multilingual and Multimodal Framework for Culturally Grounded Spoken Visual QA View paper
- [58] Placing multi-modal, and multi-lingual Data in the Humanities Domain on the Map: the Mythotopia Geo-tagged Corpus View paper
- [59] MMReview: A Multidisciplinary and Multimodal Benchmark for LLM-Based Peer Review Automation View paper
- [60] Multimodal chain-of-thought reasoning: A comprehensive survey View paper
- [61] Position: Multimodal large language models can significantly advance scientific reasoning View paper
- [62] Multimodal reasoning with multimodal knowledge graph View paper
- [63] Language Is Not All You Need: Aligning Perception with Language Models View paper
- [64] Skywork R1V: Pioneering Multimodal Reasoning with Chain-of-Thought View paper
- [65] A survey of scientific large language models: From data foundations to agent frontiers View paper
- [66] Interdisciplinary-QG: An LLM-Based Framework for Generating High-Quality Interdisciplinary Test Questions with Knowledge Graphs and Chain-of-Thought â□¦ View paper
- [67] Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models View paper
- [68] Drivelm: Driving with graph visual question answering View paper
- [69] A question-type guided and progressive self-attention network for remote sensing visual question answering View paper
- [70] VizGenie: Toward Self-Refining, Domain-Aware Workflows for Next-Generation Scientific Visualization View paper
- [71] Eagle: Expert-guided self-enhancement for preference alignment in pathology large vision-language model View paper
- [72] MedFrameQA: A Multi-Image Medical VQA Benchmark for Clinical Reasoning View paper
- [73] Toolvqa: A dataset for multi-step reasoning vqa with external tools View paper
- [74] VilBias: A Study of Bias Detection through Linguistic and Visual Cues, presenting Annotation Strategies, Evaluation, and Key Challenges View paper
- [75] Towards Human-Level Understanding of Complex Process Engineering Schematics: A Pedagogical, Introspective Multi-Agent Framework for Open-Domain Question â□¦ View paper
- [76] An AI-Assisted Bridge Inspection System: Ontology-Based Visual Question-Answering Methodology Using Large Language Models View paper
- [77] Explaining CLIP's performance disparities on data from blind/low vision users View paper