Novelty Assessment Report

Paper: Kaleidoscope: In-language Exams for Massively Multilingual Vision Evaluation
PDF URL: https://openreview.net/pdf?id=zCYXhSy9UH
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

The evaluation of vision-language models (VLMs) has mainly relied on English-language benchmarks, leaving significant gaps in both multilingual and multicultural coverage. While multilingual benchmarks have expanded, both in size and language, many rely on translations of English datasets, failing to capture cultural nuances. In this work, we propose Kaleidoscope, as the most comprehensive exam benchmark to date for the multilingual evaluation of vision-language models. Kaleidoscope is a large-scale, in-language multimodal benchmark designed to evaluate VLMs across diverse languages and visual inputs. Kaleidoscope covers 18 languages and 14 different subjects, amounting to a total of 20,911 multiple-choice questions. Built through an open science collaboration with a diverse group of researchers worldwide, Kaleidoscope ensures linguistic and cultural authenticity. We evaluate top-performing multilingual vision-language models and find that they perform poorly on low-resource languages and in complex multimodal scenarios. Our results highlight the need for progress on culturally inclusive multimodal evaluation frameworks.

Core Task Landscape

This paper addresses: **multilingual multimodal vision-language model evaluation**

A total of **50 papers** were analyzed and organized into a taxonomy with **14 categories**.

Taxonomy Overview

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

- **Model Architecture and Training Approaches**
- **Cross-Lingual Adaptation Methods**
- **Evaluation Benchmarks and Datasets**
- **Empirical Analysis and Model Behavior Studies**
- **Application-Oriented Studies**
- **Survey and Review Literature**

Complete Taxonomy Tree

- multilingual multimodal vision-language model evaluation Survey Taxonomy
- Model Architecture and Training Approaches
  - Large-Scale Multilingual Vision-Language Models (5 papers)
  - [1] Pali-x: On scaling up a multilingual vision and language model (Chen Xi, 2023) View paper
  - [2] Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features (Tschannen, 2025) View paper
  - [7] On scaling up a multilingual vision and language model (Xi Chen, 2024) View paper
  - [11] Pali: A jointly-scaled multilingual language-image model (Chen Xi, 2022) View paper
  - [37] Aya Vision: Advancing the Frontier of Multilingual Multimodality (Dash, 2025) View paper
  - Cross-Lingual Cross-Modal Pretraining Frameworks (4 papers)
  - [5] Uc2: Universal cross-lingual cross-modal vision-and-language pre-training (Mingyang Zhou, 2021) View paper
  - [23] M2-VLP: Enhancing Multilingual Vision-Language Pre-Training via Multi-Grained Alignment (Ahtamjan Ahmat, 2025) View paper
  - [33] Unifying cross-lingual and cross-modal modeling towards weakly supervised multilingual vision-language pre-training (Zejun Li, 2023) View paper
  - [38] Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training (Zeng Yan, 2023) View paper
  - Efficient and Modular Multilingual Models (5 papers)
  - [29] Jina-VLM: Small Multilingual Vision Language Model (Andreas Koukounas, 2025) View paper
  - [31] Lgvlm-miot: A lightweight generative visual-language model for multilingual iot applications (Yu Weng, 2025) View paper
  - [41] mblip: Efficient bootstrapping of multilingual vision-llms (Geigle, 2024) View paper
  - [42] Distilling Multilingual Vision-Language Models: When Smaller Models Stay Multilingual (Sukrit Sriratanawilai, 2025) View paper
  - [44] uCLIP: Parameter-Efficient Multilingual Extension of Vision-Language Models with Unpaired Data (Dahyun Chung, 2025) View paper
- Cross-Lingual Adaptation Methods
  - Machine Translation-Based Augmentation (2 papers)
  - [20] Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models (Po-Yao Huang, 2021) View paper
  - [21] Large multilingual models pivot zero-shot multimodal learning across languages (Hu Jinyi, 2023) View paper
  - Semantic Alignment and Distillation Approaches (3 papers)

- [6] Cross-Lingual Adaptation for Vision-Language Model via Multimodal Semantic Distillation (Yu Weng, 2025) View paper
- [17] DC-CLIP: Multilingual CLIP Compression via vision-language distillation and vision-language alignment (Wenbo Zhang, 2025) View paper
- [30] Cross-Lingual Semantic Alignment in Large Language Models via Context-Aware Training (Tang, 2025) View paper

- Evaluation Benchmarks and Datasets
  - Comprehensive Multilingual Multimodal Benchmarks ★ (6 papers)
  - [0] Kaleidoscope: In-language Exams for Massively Multilingual Vision Evaluation (Anon et al., 2026) View paper
  - [14] Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models (Das, 2024) View paper
  - [16] M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models (Zhang Wenxuan, 2023) View paper
  - [25] MVL-SIB: A Massively Multilingual Vision-Language Benchmark for Cross-Modal Topical Matching (Schmidt, 2025) View paper
  - [27] M4u: Evaluating multilingual understanding and reasoning for large multimodal models (Wang Hongyu, 2024) View paper
  - [32] M5--A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks (Schneider, 2024) View paper
  - Task-Specific Evaluation Benchmarks (6 papers)
  - [10] A culturally-diverse multilingual multimodal video benchmark & model (Bhuiyan Sanjid Shafique, 2025) View paper
  - [12] Evaluating multimodal large language models for visual question-answering in italian (A Scaiella, 2024) View paper
  - [22] xGQA: Cross-lingual visual question answering (Jonas Pfeiffer, 2022) View paper
  - [34] WorldMedQA-V: a multilingual, multimodal medical examination dataset for multimodal language models evaluation (Matos Joao, 2024) View paper
  - [35] Cross-lingual text-rich visual comprehension: An information theory perspective (Xinmiao Yu, 2025) View paper
  - [40] Multilingual Evaluation of Image-Text Retrieval in Visionâ□□Language Models: A Metric-Based Perspective (Bodhisatta Maiti, 2025) View paper
  - Cultural and Linguistic Diversity Benchmarks (5 papers)
  - [15] Chitrarth: Bridging vision and language for a billion people (Shaharukh Khan, 2025) View paper
  - [24] Alignmmbench: Evaluating chinese multimodal alignment in large vision-language models (Wu, 2025) View paper
  - [47] CVLUE: A New Benchmark Dataset for Chinese Vision-Language Understanding Evaluation (Yuxuan Wang, 2025) View paper
  - [49] EverydayMMQA: A Multilingual and Multimodal Framework for Culturally Grounded Spoken Visual QA (Alam, 2025) View paper
  - [50] LinguaMark: Do Multimodal Models Speak Fairly? A Benchmark-Based Evaluation (Narayanan Aravind, 2025) View paper
- Empirical Analysis and Model Behavior Studies
  - Cross-Lingual Transfer and Generalization Analysis (4 papers)
  - [3] Multilingual diversity improves vision-language representations (Nguyen Thao, 2024) View paper
  - [8] Centurio: On drivers of multilingual ability of large vision-language model (Geigle, 2025) View paper
  - [18] TowerVision: Understanding and Improving Multilinguality in Vision-Language Models (Fernandes, 2025) View paper
  - [28] Improving the cross-lingual generalisation in visual question answering (Nooralahzadeh, 2023) View paper
  - Model Robustness and Failure Analysis (2 papers)
  - [4] Towards cross-lingual explanation of artwork in large-scale vision language models (Hayashi Kazuki, 2025) View paper
  - [9] Mitigating multilingual hallucination in large vision-language models (Qu, 2024) View paper
- Application-Oriented Studies
  - Multimodal Machine Translation and Captioning (2 papers)
  - [39] Multimodal machine translation through visuals and speech (U. Sulubacak, 2020) View paper
  - [45] Rethinking Multilingual Vision-Language Translation: Dataset, Evaluation, and Adaptation (Wang, 2025) View paper
  - Navigation and Assistive Applications (2 papers)
  - [26] Cross-lingual vision-language navigation (Yan, 2019) View paper
  - [43] Evaluating Multimodal Language Models as Visual Assistants for Visually Impaired Users (Karamolegkou, 2025) View paper
  - Affective Analysis and Stance Detection (2 papers)
  - [36] Exploring vision language models for multimodal and multilingual stance detection (Vasilakes, 2025) View paper
  - [46] MMAFFBen: A Multilingual and Multimodal Affective Analysis Benchmark for Evaluating LLMs and VLMs (Liu ZhiWei, 2025) View paper
- Survey and Review Literature (3 papers)
  - [13] A Survey of State of the Art Large Vision Language Models: Benchmark Evaluations and Challenges (Z Li, 2025) View paper
  - [19] Multilingual Vision-Language Models, A Survey (LibovickÃ½, 2025) View paper
  - [48] Towards Multilingual Vision-Language Models (Huang, 2022) View paper

Narrative

Core task: multilingual multimodal vision-language model evaluation. The field has evolved around several interconnected branches that together address how vision-language models perform across diverse languages and modalities. Model Architecture and Training Approaches explore foundational designs—ranging from early multilingual pretraining frameworks like UC2[5] and Pali[11] to more recent large-scale efforts such as Pali-X[1] and Siglip 2[2]—that aim to build robust cross-lingual representations from scratch or via distillation methods like Cross-Lingual Multimodal Distillation[6]. Cross-Lingual Adaptation Methods focus on transfer techniques, including pivoting strategies and weakly supervised alignment, to extend English-centric models to lower-resource languages. Evaluation Benchmarks and Datasets form a dense branch, introducing comprehensive test suites like Exams-V[14], M3exam[16], and M4u[27] that span multiple languages and task types, while Empirical Analysis and Model Behavior Studies investigate phenomena such as hallucination mitigation and semantic alignment. Application-Oriented Studies apply these models to domains like medical QA, artwork explanation, and navigation, and Survey and Review Literature synthesizes progress across the taxonomy.

Within the evaluation landscape, a particularly active line of work centers on comprehensive multilingual multimodal benchmarks that stress-test models on diverse reasoning and perception tasks. Kaleidoscope[0] sits squarely in this cluster, offering a broad assessment framework that complements neighboring efforts like Exams-V[14] and M3exam[16], which emphasize academic exam-style questions, and M4u[27], which targets understanding across varied modalities. While Exams-V[14] and M3exam[16] prioritize structured knowledge evaluation in educational contexts, Kaleidoscope[0] appears to adopt a wider lens, potentially incorporating richer task diversity or cultural variation akin to what Culturally-Diverse Video Benchmark[10] explores for video data. This positioning reflects an ongoing tension in the field: balancing depth in specific reasoning domains against breadth in language and task coverage, a trade-off that remains central as researchers seek benchmarks capable of revealing both cross-lingual transfer gaps and fine-grained model behaviors.

Related Works in Same Category

The following **5 sibling papers** share the same taxonomy leaf node with the original paper:

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

Authors: Das, Rocktim Jyoti, Rocktim Jyoti Das, Li, Haonan, et al. (14 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

We introduce EXAMS-V, a new challenging multi-discipline multimodal multilingual exam benchmark for evaluating vision language models. It consists of 20,932 multiple-choice questions across 20 school disciplines covering natural science, social science, and other miscellaneous studies, e.g., religion, fine arts, business, etc. EXAMS-V includes a variety of multimodal features such as text, images, tables, figures, diagrams, maps, scientific symbols, and equations. The questions come in 11 langua...

Relationship Analysis

Both papers belong to the Comprehensive Multilingual Multimodal Benchmarks category, focusing on large-scale exam-based evaluation of vision-language models across diverse languages and subjects. They overlap significantly in their approach of using real-world, in-language exam questions with multimodal content (images, tables, diagrams) to assess VLM performance across multiple languages and educational domains. The key differences are that Kaleidoscope covers 18 languages with 20,911 questions emphasizing vision-grounded reasoning and cultural authenticity through open science collaboration, while EXAMS-V covers 11 languages with 20,932 questions and uniquely presents entire question snapshots (interleaved text and visual elements) rather than separating them, requiring models to perform text extraction alongside visual reasoning.

---

2. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models

Authors: Zhang Wenxuan, Wenxuan Zhang, Aljunied, Sharifah Mahani, Sharifah Mahani Aljunied, et al. (13 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Despite the existence of various benchmarks for evaluating natural language processing models, we argue that human exams are a more suitable means of evaluating general intelligence for large language models (LLMs), as they inherently demand a much wider range of abilities such as language understanding, domain knowledge, and problem-solving skills. To this end, we introduce M3Exam, a novel benchmark sourced from real and official human exam questions for evaluating LLMs in a multilingual, multi...

Relationship Analysis

Both papers belong to the comprehensive multilingual multimodal benchmarks category, focusing on large-scale evaluation of vision-language models across diverse languages and modalities using exam-based questions. They overlap significantly in their approach of collecting real-world, in-language exam questions with multimodal components (images) to assess VLM capabilities across multiple languages and educational contexts. The key differences are in scale and scope: Kaleidoscope covers 18 languages with 20,911 questions (55% multimodal) across 14 subjects, while M3Exam covers 9 languages with 12,317 questions (23% multimodal) organized by three educational levels, with M3Exam emphasizing the multilevel structure more explicitly and Kaleidoscope providing broader language and subject coverage.

---

3. MVL-SIB: A Massively Multilingual Vision-Language Benchmark for Cross-Modal Topical Matching

Authors: Schmidt, Fabian David, Schneider, Florian, Fabian David Schmidt, et al. (12 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Existing multilingual vision-language (VL) benchmarks often only cover a handful of languages. Consequently, evaluations of large vision-language models (LVLMs) predominantly target high-resource languages, underscoring the need for evaluation data for low-resource languages. To address this limitation, we introduce MVL-SIB, a massively multilingual vision-language benchmark that evaluates both cross-modal and text-only topical matching across 205 languages -- over 100 more than the most multili...

Relationship Analysis

Both papers belong to the Comprehensive Multilingual Multimodal Benchmarks category, focusing on large-scale evaluation of vision-language models across diverse languages and modalities. They overlap in their goal of assessing VLMs on multilingual multimodal tasks using multiple-choice question formats, with both emphasizing the importance of in-language evaluation beyond English-centric benchmarks. However, Kaleidoscope focuses on real-world exam questions across 18 languages with 14 subjects (20,911 questions), emphasizing educational assessment and cultural authenticity through manual collection, while MVL-SIB covers 205 languages with cross-modal topic matching tasks (images-to-sentence and sentences-to-image), leveraging professionally translated parallel texts from SIB-200 to enable comparative evaluation across a much broader linguistic range.

---

4. M4u: Evaluating multilingual understanding and reasoning for large multimodal models

Authors: Wang Hongyu, Xu Jiayu, Hongyu Wang, Jiayu Xu, Wang Rui-ping, et al. (16 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Multilingual capability is an essential aspect for large multimodal models, since they are usually deployed across various countries and languages. However, most existing benchmarks for multilingual multimodal reasoning struggle to differentiate between models of varying performance; even language models without visual capabilities can easily achieve high scores. This leaves a comprehensive evaluation of leading multilingual multimodal models largely unexplored. In this work, we introduce M4U, a...

Relationship Analysis

Both papers belong to the Comprehensive Multilingual Multimodal Benchmarks category, focusing on large-scale evaluation of vision-language models across diverse languages and task types. They overlap in their emphasis on multilingual multimodal evaluation using exam-style multiple-choice questions, covering diverse subjects including STEM and humanities, and both highlight performance disparities across languages and modalities. However, Kaleidoscope covers 18 languages with 20,911 questions emphasizing in-language authenticity through global collaboration, while M4U focuses on 6 languages (with a mini version extending to 3 more) with 10,005 questions primarily translated from Chinese, and includes more detailed analysis of cross-lingual multimodal reasoning where images and text are in different languages.

---

5. M5--A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks

Authors: Schneider, Florian, Florian Schneider, Sitaram, Sunayana, et al. (6 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Since the release of ChatGPT, the field of Natural Language Processing has experienced rapid advancements, particularly in Large Language Models (LLMs) and their multimodal counterparts, Large Multimodal Models (LMMs). Despite their impressive capabilities, LLMs often exhibit significant performance disparities across different languages and cultural contexts, as demonstrated by various text-only benchmarks. However, current research lacks such benchmarks for multimodal visio-linguistic settings...

Relationship Analysis

Both papers belong to the Comprehensive Multilingual Multimodal Benchmarks category, focusing on large-scale evaluation of vision-language models across diverse languages and tasks. They overlap in evaluating VLMs on multilingual multimodal question-answering tasks with culturally diverse content, covering multiple languages (18 in Kaleidoscope vs. 41 in M5) and various visual reasoning challenges. The key difference is that Kaleidoscope focuses exclusively on real-world exam questions (20,911 MCQs from 14 subjects) with in-language curation by native speakers, while M5 aggregates eight existing datasets plus two novel ones across five different task types (VQA, VGR, VNLI, VLOD, IC) with emphasis on underrepresented African and Asian languages.

Contributions Analysis

Overall novelty summary. The paper introduces Kaleidoscope, a large-scale multilingual multimodal benchmark spanning 18 languages and 14 subjects with over 20,000 multiple-choice questions. It resides in the Comprehensive Multilingual Multimodal Benchmarks leaf, which contains six papers including Exams-V, M3exam, and M4u. This leaf represents one of the most active research directions in the taxonomy, reflecting sustained community interest in holistic evaluation frameworks that stress-test vision-language models across diverse linguistic and task contexts rather than narrow domain-specific assessments.

The taxonomy reveals that Kaleidoscope sits within the broader Evaluation Benchmarks and Datasets branch, which also includes Task-Specific Evaluation Benchmarks (focused on VQA, retrieval, or document comprehension) and Cultural and Linguistic Diversity Benchmarks (emphasizing region-specific visual contexts). Neighboring branches address Model Architecture and Training Approaches and Cross-Lingual Adaptation Methods, indicating that the field balances benchmark creation with model development. Kaleidoscope's emphasis on in-language data collection and cultural authenticity aligns it more closely with the Cultural and Linguistic Diversity leaf than with translation-based benchmarks, though it remains classified under comprehensive evaluation due to its multi-subject scope.

Among 30 candidates examined, the KALEIDOSCOPE benchmark contribution shows one refutable candidate out of ten examined, suggesting substantial prior work in comprehensive multilingual evaluation. The open science collaboration contribution faces stronger overlap, with six refutable candidates among ten examined, indicating that collaborative data collection methods are well-established in the field. The evaluation revealing performance disparities shows one refutable candidate out of ten, implying that while empirical findings on cross-lingual gaps are documented, the specific modality-language interaction patterns may offer incremental insights. The limited search scope means these statistics reflect top-K semantic matches rather than exhaustive coverage.

Given the search examined 30 candidates and found eight refutable pairs across three contributions, the work appears to build on a moderately crowded research area. The benchmark's scale and language coverage may differentiate it from siblings like Exams-V or M3exam, but the analysis cannot confirm whether these differences constitute substantial novelty without deeper comparison. The collaborative methodology and performance findings align with established patterns in multilingual evaluation research, though the specific combination of scale, authenticity, and task diversity may offer value to practitioners seeking comprehensive assessment tools.

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

Contribution 1: KALEIDOSCOPE benchmark for multilingual multimodal evaluation

Description: The authors introduce KALEIDOSCOPE, a large-scale benchmark containing 20,911 multiple-choice questions across 18 languages and 14 subjects. The benchmark is designed to evaluate vision-language models using in-language, culturally authentic exam questions, with 55% requiring image understanding for accurate resolution.

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

URL: View paper

Prior Art Analysis

Exams-V[14] demonstrates that a similar multilingual multimodal exam benchmark was published prior to KALEIDOSCOPE. Both benchmarks share core characteristics: they collect real-world exam questions across multiple languages, include multimodal content (images, tables, diagrams), use multiple-choice question format, and aim to evaluate vision-language models on culturally authentic content. Exams-V[14] contains 20,932 questions across 11 languages and 20 subjects, with 5,086 multimodal samples (24.3%), while KALEIDOSCOPE contains 20,911 questions across 18 languages and 14 subjects, with 11,459 multimodal samples (55%). The key methodological overlap is that both benchmarks collect in-language exam questions from official sources rather than translations, and both emphasize the importance of culturally grounded evaluation.

Evidence

Evidence 1 - **Rationale**: Both papers claim to introduce comprehensive multilingual multimodal exam benchmarks with similar scale (20,911 vs 20,932 questions), multiple-choice format, and coverage of diverse subjects and languages. This demonstrates prior work exists in this space. - **Original**: we propose kaleidoscope, as the most comprehensive exam benchmark to date for the multilingual evaluation of vision-language models. kaleidoscopeis a large-scale, in-language multimodal benchmark designed to evaluate vlms across diverse languages and visual inputs. kaleidoscope covers 18 languages a... - **Candidate**: we introduce exams-v, a new challenging multi-discipline multimodal multilingual exam benchmark for evaluating vision language models. it consists of 20,932 multiple-choice questions across 20 school disciplines covering natural science, social science, and other miscellaneous studies, e.g., religio...

Evidence 2 - **Rationale**: Both benchmarks collect real-world exam questions from official sources. Exams-V[14] explicitly states it collects from 'official state examinations crafted by the ministries of education', similar to KALEIDOSCOPE's claim of using 'real-world, in-language exam questions'. - **Original**: in this work, we introduce the largest benchmark of real-world, in-language exam questions blending image and text modalities. our dataset pushes beyond simple captioning, challenging models to reason about visual content in various topics, the way humans are evaluated in exams worldwide. through a ... - **Candidate**: exams-v is a multimodal extension of the exams dataset (hardalov et al., 2020), which is collected from official state examinations crafted by the ministries of education across different countries. these assessments, taken by high school graduates, cover diverse subjects, including core disciplines...

Evidence 3 - **Rationale**: Both papers emphasize the importance of avoiding translated benchmarks and instead collecting authentic, in-language exam questions from diverse regions to capture cultural context and region-specific knowledge. - **Original**: a common but imperfect solution is translating english benchmarks into other languages. while convenient, this often falls short of capturing cultural context and nuance. translated datasets can easily reinforce western-centric knowledge and assumptions (van miltenburg et al., 2017; frank et al., 20... - **Candidate**: unlike existing benchmarks, exams-v is uniquely curated by gathering school exam questions from various countries, with a variety of education systems. this distinctive approach calls for intricate reasoning across diverse languages and relies on region-specific knowledge.

Evidence 4 - **Rationale**: Both benchmarks explicitly cover a range from high-resource to low-resource languages and emphasize linguistic diversity across multiple language families, demonstrating similar design principles. - **Original**: kaleidoscopebenchmark: we present the largest multilingual multimodal exam set, covering high resource (e.g., english, spanish) to underrepresented languages (e.g., bengali, telugu) across diverse subjects from sociology to stem. most languages (10/18) include 5+ topics, with the rest focusing on mu... - **Candidate**: the dataset includes high-resource languages like english and chinese and low-resource languages such as bulgarian, croatian, and serbian. it offers a diverse linguistic landscape, spanning germanic, slavic, and sino-tibetan language families. we also include arabic, which has a script directionalit...

---

2. Improving the cross-lingual generalisation in visual question answering

URL: View paper

Brief Assessment

Cross-Lingual VQA Generalisation[28] focuses on improving cross-lingual transfer for visual question answering using fine-tuning strategies, not on creating multilingual multimodal evaluation benchmarks with exam questions.

---

3. WorldMedQA-V: a multilingual, multimodal medical examination dataset for multimodal language models evaluation

URL: View paper

Brief Assessment

WorldMedQA-V[34] focuses specifically on medical examination datasets across four countries (Brazil, Israel, Japan, Spain) with clinical validation, while KALEIDOSCOPE covers 18 languages across 14 general subjects including STEM, humanities, and social sciences. The domain specialization and scope differ substantially.

---

4. Cross-lingual text-rich visual comprehension: An information theory perspective

URL: View paper

Brief Assessment

Text-Rich Visual Comprehension[35] focuses on cross-lingual text-rich visual question answering where image text and question languages differ, while KALEIDOSCOPE evaluates multilingual multimodal models using in-language exam questions where both are in the same language.

---

5. MSA at ImageCLEF 2025 Multimodal Reasoning: Multilingual Multimodal Reasoning With Ensemble Vision Language Models

URL: View paper

Brief Assessment

ImageCLEF Multimodal Reasoning[54] focuses on a competition system using ensemble VLMs for exam question answering, not on creating a comprehensive multilingual multimodal evaluation benchmark with in-language exam questions across diverse subjects and languages.

---

6. Seed-bench: Benchmarking multimodal large language models

URL: View paper

Brief Assessment

Seed-Bench[51] focuses on hierarchical capability evaluation of multimodal models with English-only questions, while KALEIDOSCOPE emphasizes multilingual (18 languages) and culturally authentic exam questions. The benchmarks serve different evaluation purposes.

---

7. xGQA: Cross-lingual visual question answering

URL: View paper

Brief Assessment

xGQA[22] focuses on visual question answering across 7 languages using template-based questions from GQA, while KALEIDOSCOPE covers 18 languages with 20,911 exam-based multiple-choice questions across 14 subjects. The datasets differ fundamentally in scope, question format, and evaluation methodology.

---

8. Cvqa: Culturally-diverse multilingual visual question answering benchmark

URL: View paper

Brief Assessment

CVQA[53] focuses on culturally-diverse visual question answering with 10k samples across 30 countries, while KALEIDOSCOPE is a larger exam-based benchmark with 20,911 questions across 18 languages emphasizing in-language, culturally authentic exam questions with 55% requiring image understanding.

---

9. Parameter-efficient cross-lingual transfer of vision and language models via translation-based alignment

URL: View paper

Brief Assessment

Translation-Based Alignment[52] focuses on parameter-efficient cross-lingual transfer of vision-language models through translation-based alignment methods, not on creating multilingual multimodal evaluation benchmarks with exam questions.

---

10. Llava-ndino: Empowering llms with multimodality for the italian language

URL: View paper

Brief Assessment

Llava-Ndino[55] focuses on developing Italian-language vision-language models through adaptation and instruction-tuning, not on creating multilingual evaluation benchmarks with exam questions across diverse languages.

Contribution 2: Open science collaboration for authentic data collection

Description: The authors conduct a large-scale open science effort involving contributors from 20 nations across four continents to manually collect exam questions in their original languages. This participatory approach ensures that the benchmark captures genuine linguistic and cultural nuances rather than relying on translations.

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 cartography of open collaboration in open source ai: Mapping practices, motivations, and governance in 14 open large language model projects

URL: View paper

Brief Assessment

Open Collaboration Cartography[72] focuses on collaborative practices in open LLM development across organizational contexts, not on multilingual benchmark creation or culturally authentic data collection for vision-language model evaluation.

---

2. The bitter lesson learned from 2,000+ multilingual benchmarks

URL: View paper

Prior Art Analysis

Bitter Lesson Multilingual[68] demonstrates that large-scale open science efforts for multilingual benchmark creation existed prior to the original paper. The candidate paper documents a comprehensive analysis of over 2,000 multilingual benchmarks from 148 countries, published between 2021 and 2024, which predates the original paper's submission. The candidate explicitly discusses participatory research and collaborative benchmark development practices that have been established in the field, including examples of community-driven data collection efforts across multiple continents and languages. This evidence shows that the original paper's claim to novelty in conducting 'a large-scale open science effort involving contributors from 20 nations across four continents' is not unique, as similar collaborative approaches have been documented and analyzed in the multilingual evaluation community.

Evidence

Evidence 1 - **Rationale**: The candidate paper's analysis of 2,000+ benchmarks from 148 countries (published 2021-2024) demonstrates that large-scale multilingual benchmark creation efforts across multiple continents were already established before the original paper's work, challenging the novelty of the 20-nation collaboration. - **Original**: in this work, we engage in a large-scale open science collection process, which brings together contributors spanning 20 nations across four continents to ensure linguistic and cultural authenticity. - **Candidate**: This position paper examinesover2,000multilingual(non-english)benchmarksfrom148countries, publishedbetween2021 and 2024, to evaluate past, present, and future practices in multilingual benchmarking.

Evidence 2 - **Rationale**: The candidate paper explicitly discusses the established practice of global collaboration for multilingual benchmark creation and notes that 'multiple teams' have been conducting 'similar problems' in this space, indicating that large-scale collaborative efforts for authentic multilingual data collection were already common practice. - **Original**: the manual curation of datasets is a costly process that requires careful attention to detail in every language to ensure high-quality, contextually relevant content for evaluation. in this work, we engage in a large-scale open science collection process - **Candidate**: global collaboration for inclusive benchmarkingthe creation of truly representative multilingual benchmarksdemandscollaborationacrosslinguistic, organizational, national, andculturalboundaries. the fragmentation of efforts we observe today leads to significant resource inefficiencies, with multiple ...

Evidence 3 - **Rationale**: The candidate paper's advocacy for international research consortia and discussion of pooling expertise across languages and cultures indicates that such collaborative approaches were already being discussed and implemented in the field, suggesting the original paper's participatory approach was not novel. - **Original**: for related participatory research see appendix c.1. - **Candidate**: we advocate for international research consortia specifically focused on multilingual benchmark development, where expertise across different languages and cultures can be pooled to create more comprehensive evaluation frameworks.

---

3. Mmteb: Massive multilingual text embedding benchmark

URL: View paper

Brief Assessment

MMTEB[65] focuses on text embedding benchmark creation through open collaboration, not vision-language exam question collection. The candidate involves community-driven dataset aggregation for NLP tasks, while the original paper describes manual collection of culturally authentic multimodal exam questions across languages.

---

4. GIMMICK--Globally Inclusive Multimodal Multitask Cultural Knowledge Benchmarking

URL: View paper

Brief Assessment

GIMMICK[66] uses synthetic data generation with GPT-4o followed by expert annotation, rather than a large-scale open science effort with contributors from multiple nations manually collecting exam questions in original languages. The methodologies differ fundamentally in their approach to data collection.

---

5. Palm: A culturally inclusive and linguistically diverse dataset for arabic llms

URL: View paper

Prior Art Analysis

Palm[69] demonstrates a similar large-scale open science effort for culturally authentic data collection. The candidate paper describes a year-long community-driven project involving 44 researchers from 15 different Arab countries to manually create culturally and linguistically diverse instructions. This collaborative approach, like the original paper's effort across 20 nations, ensures authentic cultural representation through direct involvement of native speakers and local experts rather than relying on translations. Both projects emphasize participatory research methods where contributors from diverse geographic regions manually curate data to capture genuine cultural nuances.

Evidence

Evidence 1 - **Rationale**: Both papers describe large-scale collaborative efforts involving researchers from multiple countries to create comprehensive datasets, demonstrating that similar open science approaches existed prior to the original paper's submission. - **Original**: Through a large-scale open science effort across 18 languages, we construct kaleidoscope(see figure 1), featuring a diverse selection of knowledge domains across 14 subjects. - **Candidate**: Through a year-long, community-driven effort involving44 researchers from across 15 different arab countries, palm offers a comprehensive set of instructions that cover both msa and various regional dialects.

Evidence 2 - **Rationale**: Both papers explicitly describe engaging contributors from multiple nations to ensure cultural authenticity through participatory open science methods, showing this approach was already established. - **Original**: in this work, we engage in a large-scale open science collection process, which brings together contributors spanning 20 nations across four continents to ensure linguistic and cultural authenticity. - **Candidate**: palm is a community project involving 44 trained native speakers, all of whom are authors of this work. we aimed to incorporate local knowledge from every arab country and succeeded for 15 out of 22.

Evidence 3 - **Rationale**: Both papers emphasize manual curation through large-scale collaborative efforts leveraging local expertise, demonstrating that this participatory approach to authentic data collection was already in practice. - **Original**: the manual curation of datasets is a costly process that requires careful attention to detail in every language to ensure high-quality, contextually relevant content for evaluation. in this work, we engage in a large-scale open science collection process - **Candidate**: palm is the first dataset at the country level to cover all 22 arab countries, spanning 20 culturally relevant topics. what sets palm apart is its inclusion of instructions in both msa and local dialects, all of which are human-annotated using reliable, country-specific sources. this dataset was dev...

---

6. Crowdsource, crawl, or generate? creating sea-vl, a multicultural vision-language dataset for southeast asia

URL: View paper

Prior Art Analysis

SEA-VL[71] demonstrates that large-scale open science efforts involving contributors from multiple nations to manually collect culturally authentic data in original languages predates the ORIGINAL paper's approach. SEA-VL[71] engaged contributors from Southeast Asian countries to collect culturally relevant images with captions in native languages, ensuring linguistic and cultural authenticity through participatory methods. This directly parallels the ORIGINAL paper's claimed novelty of conducting 'a large-scale open science effort across 18 languages' with 'contributors spanning 20 nations across four continents to ensure linguistic and cultural authenticity.'

Evidence

Evidence 1 - **Rationale**: Both papers describe large-scale open science efforts involving contributors from multiple nations to ensure cultural and linguistic authenticity in data collection. - **Original**: in this work, we engage in a large-scale open science collection process, which brings together contributors spanning 20 nations across four continents to ensure linguistic and cultural authenticity. - **Candidate**: by involving contributors from sea countries, sea-vl ensures better cultural relevance and diversity, fostering greater inclusivity of underrepresented languages and cultural depictions in vl research.

Evidence 2 - **Rationale**: Both papers emphasize ensuring authentic representation through native speaker involvement and focus on underrepresented languages. - **Original**: our work is built around three core design principles that guide the selection, curation, processing, and addition of exams: /ctremultimodality: images are central to kaleidoscope, as we aim to evaluate how vlms integrate and reason about visual information to answer questions. we prioritize multimo... - **Candidate**: to ensure that the data collected authentically represent the lived experiences and cultural contexts of the region, an extensive human evaluation by native participants is performed using different image collection methodologies.

Evidence 3 - **Rationale**: Both papers describe manual data collection processes involving contributors providing content in their native languages with metadata to ensure cultural authenticity. - **Original**: the manual curation of datasets is a costly process that requires careful attention to detail in every language to ensure high-quality, contextually relevant content for evaluation. in this work, we engage in a large-scale open science collection process, which brings together contributors spanning ... - **Candidate**: for image collection, we ask contributors to submit only those images they personally own, avoiding images retrieved from publicly accessible platforms. contributors upload their images through a designated form, providing metadata on the location where the image was taken and to which of the 11 sea...

---

7. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming

URL: View paper

Brief Assessment

CulturalBench[67] focuses on human-AI red-teaming for cultural knowledge benchmarking, not on open science collaboration for multilingual exam collection. The candidate's approach involves iterative question generation with AI assistance rather than large-scale participatory exam curation across nations.

---

8. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages

URL: View paper

Prior Art Analysis

SEACrowd[70] demonstrates a similar large-scale open science collaborative effort for authentic multilingual data collection. The candidate paper describes a collaborative initiative involving researchers primarily from the Southeast Asian region to consolidate and standardize datasets in nearly 1,000 SEA languages. Like the original paper's approach of engaging contributors from 20 nations across four continents to manually collect exam questions in original languages, SEACrowd engaged contributors to submit datasheet forms for publicly available datasets and created standardized dataloaders. Both efforts emphasize participatory approaches with native speakers to ensure linguistic and cultural authenticity rather than relying on translations.

Evidence

Evidence 1 - **Rationale**: Both papers describe large-scale collaborative efforts involving researchers from multiple regions to ensure authentic data collection, demonstrating that similar participatory open science approaches existed prior to the original paper's work. - **Original**: in this work, we engage in a large-scale open science collection process, which brings together contributors spanning 20 nations across four continents to ensure linguistic and cultural authenticity. - **Candidate**: seacrowd represents the first comprehensive ai dataset collection initiative for sea, developed through a collaborative effort among researchers and engineers primarily based in the sea region.

Evidence 2 - **Rationale**: Both papers emphasize manual curation and careful attention to detail in data collection across multiple languages to ensure quality and cultural relevance, showing that the participatory approach for authentic multilingual data collection was already established. - **Original**: this is acutely needed in the field of machine learning, where recent studies have highlighted that dataset creators remain predominantly western-centric (longpre et al., 2025). the manual curation of datasets is a costly process that requires careful attention to detail in every language to ensure ... - **Candidate**: we invited contributors to submit datasheet forms (gebru et al., 2021) for publicly available datasets across all modalities including text, audio, and image in sea languages and/or cultures. these datasheets include detailed information about each dataset, such as data subset(s), description, task,...

Evidence 3 - **Rationale**: Both papers prioritize original, culturally authentic data sources over translations, demonstrating that the emphasis on authentic data collection through participatory methods was already present in prior work. - **Original**: we prioritized original, domain-expert-written questions (e.g., from teachers), ensuring real-world relevance and quality. the exams were gathered from various repositories, including official government websites, question banks, and other publicly available repositories with educational materials. - **Candidate**: as a proxy of the cultural relevance of sea datasets, we manually curated 259 data subsets used in seacrowd evaluation based on their data source. specifically, we categorize them whether they are 1) translated from another language, 2) crawled from local sources, or 3) hand-crafted to capture cultu...

---

9. Aya dataset: An open-access collection for multilingual instruction tuning

URL: View paper

Prior Art Analysis

Aya Dataset[63] demonstrates that large-scale open science efforts involving global contributors for authentic multilingual data collection existed prior to the original paper. The candidate paper describes a year-long participatory research initiative with contributors from 119 countries collecting human-curated instruction-following data in 65 languages. This directly parallels the original paper's claim of conducting open science with contributors from 20 nations across four continents to manually collect exam questions. Both papers emphasize participatory approaches to ensure linguistic and cultural authenticity rather than relying on translations, indicating that this methodology was already established in the field.

Evidence

Evidence 1 - **Rationale**: Both papers describe large-scale participatory research initiatives involving global contributors, demonstrating that this methodology existed prior to the original paper's work. - **Original**: in this work, we engage in a large-scale open science collection process, which brings together contributors spanning 20 nations across four continents to ensure linguistic and cultural authenticity. - **Candidate**: we set about to close this gap by conducting a year-long participatory research initiative that involved working with fluent speakers of languages from around the world to collect human-curated instances of instructions and completions.

Evidence 2 - **Rationale**: The candidate paper explicitly frames participatory research with global collaborators as an established framework, indicating this approach was already recognized in the field before the original paper's work. - **Original**: our work entailed an extensive, open science process to manually collect data by working directly with native speakers of different languages (elliott et al., 2016; liu et al., 2021; thapliyal et al., 2022; li et al., 2024c; üstün et al., 2024; singh et al., 2024b). this is acutely needed in the fie... - **Candidate**: The aya initiative also serves as a valuable case study in participatory research, involving collaborators from 119 countries. We see this as an important framework for future research collaborations that aim to bridge gaps in resources.

Evidence 3 - **Rationale**: Both papers emphasize ensuring cultural authenticity through manual curation by native speakers, showing this was an established practice in multilingual dataset creation. - **Original**: the manual curation of datasets is a costly process that requires careful attention to detail in every language to ensure high-quality, contextually relevant content for evaluation. in this work, we engage in a large-scale open science collection process, which brings together contributors spanning ... - **Candidate**: The goal of the aya project is to facilitate annotations to a crowd-sourced dataset by individuals fluent in different languages. inputs from speakers of each language ensure that the dataset is more likely to be organic, articulate, and representative of the speakers' cultures.

---

10. Culturebank: An online community-driven knowledge base towards culturally aware language technologies

URL: View paper

Prior Art Analysis

CultureBank[64] demonstrates that large-scale open science efforts involving diverse contributors for authentic data collection existed prior to the original paper. The candidate paper describes a bottom-up approach utilizing online communities where contributors from different backgrounds manually collect and process cultural data. Both papers employ participatory research methods with global contributors to ensure linguistic and cultural authenticity, though they target different domains (cultural knowledge vs. multilingual exams).

Evidence

Evidence 1 - **Rationale**: This pair shows that CultureBank addresses the same limitation of formal sources by using community-driven data collection, establishing prior work in participatory cultural data gathering. - **Original**: in this work, we engage in a large-scale open science collection process, which brings together contributors spanning 20 nations across four continents to ensure linguistic and cultural authenticity. - **Candidate**: to enhance llms' culture awareness, existing studies have developed cultural knowledge databases to represent culture-related knowledge and norms, but they have several limitations. (1) they often rely on formal knowledge sources like wikipedia and online articles (nguyen et al., 2023; fung et al., ...

Evidence 2 - **Rationale**: Both papers emphasize manual curation and processing at scale to ensure quality and authenticity, demonstrating similar methodological approaches to participatory data collection. - **Original**: the manual curation of datasets is a costly process that requires careful attention to detail in every language to ensure high-quality, contextually relevant content for evaluation. - **Candidate**: we develop a bottom-up approach to process noisy self-narratives on a massive scale. using this pipeline, we develop culturebank, a cultural knowledge base with 12k cultural descriptors sourced from tiktok

---

Contribution 3: Comprehensive evaluation revealing modality-specific and cross-lingual performance disparities

Description: The authors evaluate state-of-the-art vision-language models on KALEIDOSCOPE and identify significant performance gaps: models show substantially better accuracy on text-only versus multimodal questions, struggle more with STEM subjects compared to humanities, and exhibit weaker performance on low-resource and non-Latin script languages.

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. CLAIM: Mitigating Multilingual Object Hallucination in Large Vision-Language Models with Cross-Lingual Attention Intervention

URL: View paper

Brief Assessment

CLAIM[58] focuses on mitigating multilingual object hallucination in vision-language models through attention intervention, not on comprehensive evaluation of performance disparities across modalities and languages as in the original paper.

---

2. Zero-shot cross-lingual knowledge transfer in vqa via multimodal distillation

URL: View paper

Brief Assessment

Zero-Shot Multimodal Distillation[59] focuses on knowledge transfer methods for extending English VQA models to non-English languages, not on comprehensive evaluation of performance disparities across modalities and languages.

---

3. Evaluating general vision-language models for clinical medicine

URL: View paper

Brief Assessment

Clinical Medicine VLMs[61] focuses on evaluating vision-language models specifically for clinical medicine applications (gastroenterology, radiology, dermatology), not on multilingual or cross-lingual performance disparities that are central to the original paper's contribution.

---

4. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training

URL: View paper

Brief Assessment

UC2[5] focuses on cross-lingual cross-modal pre-training methods and model architecture, not on comprehensive evaluation of existing models across modalities and languages. The candidate does not demonstrate prior work on systematic performance disparity analysis.

---

5. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models

URL: View paper

Prior Art Analysis

M3exam[16] demonstrates prior work that identified similar performance disparities across modalities and languages in vision-language model evaluation. The candidate paper reports that models perform substantially better on text-only questions compared to multimodal

questions, with specific performance gaps documented (e.g., text-only models achieving 60.36% vs multimodal models achieving lower scores). The candidate also documents cross-lingual performance disparities, showing models perform worse on low-resource and non-latin script languages. These findings predate the original paper and establish that such performance gap analyses were already conducted in multilingual multimodal evaluation contexts.

Evidence

Evidence 1 - **Rationale**: Both papers evaluate and compare text-only versus multimodal model performance, documenting performance disparities between these modalities. - **Original**: all models perform substantially better on text-only questions, revealing a clear disparity across modalities. The gap widens in larger modelsl; for instance, gpt-4o shows a 21.6% difference between text-only and multimodal performance - **Candidate**: in table 4, we present the performance of various models on english questions, as there are no existing llms handling both multilingual and multimodal settings. in addition to multimodal models, we provide random guess baselines, the performance of the flan-t5 model (xxl version) [13], and the perfo...

Evidence 2 - **Rationale**: Both papers identify and document cross-lingual performance disparities, specifically noting worse performance on low-resource and non-latin script languages. - **Original**: model performance varies across languages, with better results in high-resource languages and weaker performance in midand low-resource ones (section 4.3). crosslingual transfer appears to play a role, as models perform better on average in languages using latin scripts compared to those with non-la... - **Candidate**: when comparing performance across different languages, we observe that existing models generally perform worse for non-latin languages, such as chinese (despite being relatively high-resource), as well as low-resource languages like javanese (even though it mostly uses the latin script).

Evidence 3 - **Rationale**: Both papers document domain-specific performance disparities, with models showing different performance levels across subject categories like STEM versus humanities. - **Original**: we observe a significant performance gap between questions requiring knowledge of humanities & social sciences and those focused on stem subjects (section 4.4). on average, models present accuracy of 83.7% for humanities versus 59.2% for stem - **Candidate**: notably, across all languages, the model tends to underperform in the math category. this suggests that the reasoning skills required in these questions present a great challenge for the model. conversely, the model exhibits relatively stronger performance in the natural science and social science s...

---

6. Uncovering bias in large vision-language models at scale with counterfactuals

URL: View paper

Brief Assessment

Bias in VLMs[57] focuses on social bias evaluation (toxicity, stereotypes) in vision-language models using counterfactual images, not on performance disparities across modalities, languages, or subject domains in multilingual exam benchmarks.

---

7. Cultural bias mitigation in vision-language models for digital heritage documentation: A comparative analysis of debiasing techniques

URL: View paper

Brief Assessment

Cultural Bias Mitigation[60] focuses on bias mitigation techniques in vision-language models for heritage documentation, not on comprehensive multilingual evaluation revealing performance disparities across modalities and languages as in the original paper.

---

8. Multilingual Vision-Language Models, A Survey

URL: View paper

Brief Assessment

Multilingual VLM Survey[19] is a survey paper that reviews existing models and benchmarks rather than conducting original empirical evaluations. It does not present new experimental findings that would refute the novelty of KALEIDOSCOPE's comprehensive evaluation work.

---

9. Benchmarking vision language models for cultural understanding

URL: View paper

Brief Assessment

Cultural Understanding Benchmark[56] focuses on cultural understanding across countries and cultural facets (clothing, food, rituals, traditions), not on general modality-specific or cross-lingual performance disparities in vision-language models. The evaluation dimensions differ fundamentally from KALEIDOSCOPE's focus on text-only vs. multimodal performance gaps, STEM vs. humanities disparities, and Latin vs. non-Latin script performance.

---

10. Naturalbench: Evaluating vision-language models on natural adversarial samples

URL: View paper

Brief Assessment

NaturalBench[62] focuses on natural adversarial samples from English image-text corpora (Flickr30k, DOCCI) with a vision-centric design to prevent blind solutions, not on systematic cross-lingual or modality-specific evaluation across diverse languages and scripts as in KALEIDOSCOPE.

Appendix: Text Similarity Detection

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

References

- [0] Kaleidoscope: In-language Exams for Massively Multilingual Vision Evaluation View paper
- [1] Pali-x: On scaling up a multilingual vision and language model View paper
- [2] Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features View paper
- [3] Multilingual diversity improves vision-language representations View paper
- [4] Towards cross-lingual explanation of artwork in large-scale vision language models View paper
- [5] Uc2: Universal cross-lingual cross-modal vision-and-language pre-training View paper
- [6] Cross-Lingual Adaptation for Vision-Language Model via Multimodal Semantic Distillation View paper
- [7] On scaling up a multilingual vision and language model View paper
- [8] Centurio: On drivers of multilingual ability of large vision-language model View paper
- [9] Mitigating multilingual hallucination in large vision-language models View paper
- [10] A culturally-diverse multilingual multimodal video benchmark & model View paper
- [11] Pali: A jointly-scaled multilingual language-image model View paper

- [12] Evaluating multimodal large language models for visual question-answering in italian View paper
- [13] A Survey of State of the Art Large Vision Language Models: Benchmark Evaluations and Challenges View paper
- [14] Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models View paper
- [15] Chitrarth: Bridging vision and language for a billion people View paper
- [16] M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models View paper
- [17] DC-CLIP: Multilingual CLIP Compression via vision-language distillation and vision-language alignment View paper
- [18] TowerVision: Understanding and Improving Multilinguality in Vision-Language Models View paper
- [19] Multilingual Vision-Language Models, A Survey View paper
- [20] Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models View paper
- [21] Large multilingual models pivot zero-shot multimodal learning across languages View paper
- [22] xGQA: Cross-lingual visual question answering View paper
- [23] M2-VLP: Enhancing Multilingual Vision-Language Pre-Training via Multi-Grained Alignment View paper
- [24] Alignmmbench: Evaluating chinese multimodal alignment in large vision-language models View paper
- [25] MVL-SIB: A Massively Multilingual Vision-Language Benchmark for Cross-Modal Topical Matching View paper
- [26] Cross-lingual vision-language navigation View paper
- [27] M4u: Evaluating multilingual understanding and reasoning for large multimodal models View paper
- [28] Improving the cross-lingual generalisation in visual question answering View paper
- [29] Jina-VLM: Small Multilingual Vision Language Model View paper
- [30] Cross-Lingual Semantic Alignment in Large Language Models via Context-Aware Training View paper
- [31] Lgvlm-miot: A lightweight generative visual-language model for multilingual iot applications View paper
- [32] M5--A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks View paper
- [33] Unifying cross-lingual and cross-modal modeling towards weakly supervised multilingual vision-language pre-training View paper
- [34] WorldMedQA-V: a multilingual, multimodal medical examination dataset for multimodal language models evaluation View paper
- [35] Cross-lingual text-rich visual comprehension: An information theory perspective View paper
- [36] Exploring vision language models for multimodal and multilingual stance detection View paper
- [37] Aya Vision: Advancing the Frontier of Multilingual Multimodality View paper
- [38] Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training View paper
- [39] Multimodal machine translation through visuals and speech View paper
- [40] Multilingual Evaluation of Image-Text Retrieval in Visionâ□□Language Models: A Metric-Based Perspective View paper
- [41] mblip: Efficient bootstrapping of multilingual vision-llms View paper
- [42] Distilling Multilingual Vision-Language Models: When Smaller Models Stay Multilingual View paper
- [43] Evaluating Multimodal Language Models as Visual Assistants for Visually Impaired Users View paper
- [44] uCLIP: Parameter-Efficient Multilingual Extension of Vision-Language Models with Unpaired Data View paper
- [45] Rethinking Multilingual Vision-Language Translation: Dataset, Evaluation, and Adaptation View paper
- [46] MMAFFBen: A Multilingual and Multimodal Affective Analysis Benchmark for Evaluating LLMs and VLMs View paper
- [47] CVLUE: A New Benchmark Dataset for Chinese Vision-Language Understanding Evaluation View paper
- [48] Towards Multilingual Vision-Language Models View paper
- [49] EverydayMMQA: A Multilingual and Multimodal Framework for Culturally Grounded Spoken Visual QA View paper
- [50] LinguaMark: Do Multimodal Models Speak Fairly? A Benchmark-Based Evaluation View paper
- [51] Seed-bench: Benchmarking multimodal large language models View paper
- [52] Parameter-efficient cross-lingual transfer of vision and language models via translation-based alignment View paper
- [53] Cvqa: Culturally-diverse multilingual visual question answering benchmark View paper
- [54] MSA at ImageCLEF 2025 Multimodal Reasoning: Multilingual Multimodal Reasoning With Ensemble Vision Language Models View paper
- [55] Llava-ndino: Empowering llms with multimodality for the italian language View paper
- [56] Benchmarking vision language models for cultural understanding View paper
- [57] Uncovering bias in large vision-language models at scale with counterfactuals View paper
- [58] CLAIM: Mitigating Multilingual Object Hallucination in Large Vision-Language Models with Cross-Lingual Attention Intervention View paper
- [59] Zero-shot cross-lingual knowledge transfer in vqa via multimodal distillation View paper
- [60] Cultural bias mitigation in vision-language models for digital heritage documentation: A comparative analysis of debiasing techniques View paper
- [61] Evaluating general vision-language models for clinical medicine View paper
- [62] Naturalbench: Evaluating vision-language models on natural adversarial samples View paper
- [63] Aya dataset: An open-access collection for multilingual instruction tuning View paper
- [64] Culturebank: An online community-driven knowledge base towards culturally aware language technologies View paper
- [65] Mmteb: Massive multilingual text embedding benchmark View paper
- [66] GIMMICK--Globally Inclusive Multimodal Multitask Cultural Knowledge Benchmarking View paper
- [67] CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming View paper
- [68] The bitter lesson learned from 2,000+ multilingual benchmarks View paper
- [69] Palm: A culturally inclusive and linguistically diverse dataset for arabic llms View paper
- [70] SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages View paper
- [71] Crowdsource, crawl, or generate? creating sea-vl, a multicultural vision-language dataset for southeast asia View paper
- [72] A cartography of open collaboration in open source ai: Mapping practices, motivations, and governance in 14 open large language model projects View paper