جـــامـــعـــة مـحـــمـــد بـــن زايـــد
للـــذكـاء الاصـطـنـاعـي
MOHAMED BIN ZAYED UNIVERSITY
OF ARTIFICIAL INTELLIGENCE

Collecting Culturally Nuanced Image Data: Case Study with CVQA and SEA-VL

CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark

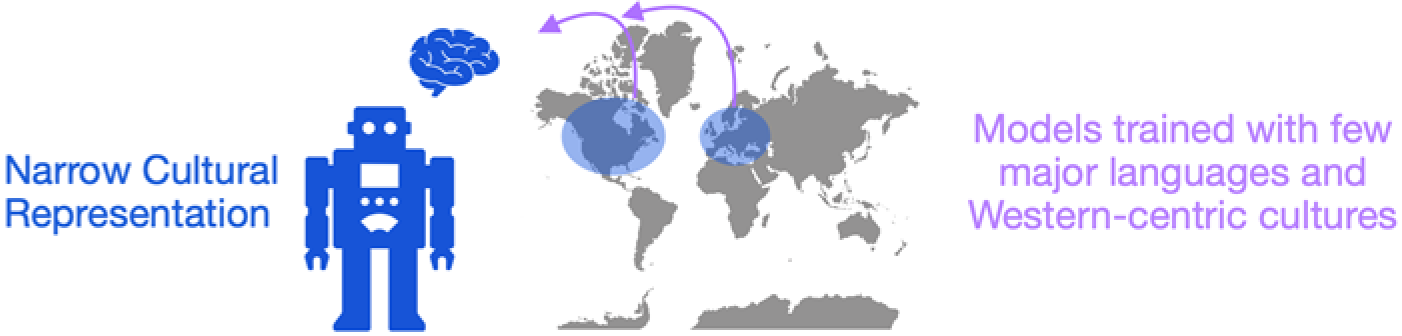
David Romero[*♣], Chenyang Lyu[*♣], Haryo Akbarianto Wibowo[♣], Teresa Lynn , Injy Hamed , Aditya Nanda Kishore , Aishik Mandal , Alina Dragonetti , Artem Abzaliev , Atnafu Lambebo Tonja , Bontu Fufa Balcha , Chenxi Whitehouse , Christian Salamea , Dan John Velasco , David Ifeoluwa Adelani , David Le Meur , Emilio Villa-Cueva , Fajri Koto , Fauzan Farooqui , Frederico Belcavello , Ganzorig Batnasan , Gisela Vallejo , Grainne Caulfield , Guido Ivetta , Haiyue Song , Henok Biadglign Ademtew , Hernán Maina , Holy Lovenia , Israel Abebe Azime , Jan Christian Blaise Cruz , Jay Gala , Jiahui Geng , Jesus-German Ortiz-Barajas , Jinheon Baek , Jocelyn Dunstan , Laura Alonso Alemany , Kumaranage Ravindu Yasas Nagasinghe , Luciana Benotti , Luis Fernando D'Haro , Marcelo Viridiano , Marcos Estecha-Garitagoitia , Maria Camila Buitrago Cabrera , Mario Rodríguez-Cantelar , Mélanie Jouitteau , Mihail Mihaylov , Naome Etori , Mohamed Fazli Mohamed Imam , Muhammad Farid Adilazuarda , Munkhjargal Gochoo , Munkh-Erdene Otgonbold , Olivier Niyomugisha , Paula Mónica Silva , Pranjal Chitale , Raj Dabre , Rendi Chevi , Ruochen Zhang , Ryandito Diandaru , Samuel Cahyawijaya , Santiago Góngora , Soyeong Jeong , Sukannya Purkayastha , Tatsuki Kuribayashi , Teresa Clifford , Thanmay Jayakumar , Tiago Timponi Torrent , Toqeer Ehsan , Vladimir Araujo , Yova Kementchedjhieva , Zara Burzo , Zheng Wei Lim , Zheng Xin Yong , Oana Ignat , Joan Nwatu , Rada Mihalcea , Thamar Solorio[♣], and Alham Fikri Aji[♣]

[♣]Core Authors (MBZUAI)

www.cvqa-benchmark.org

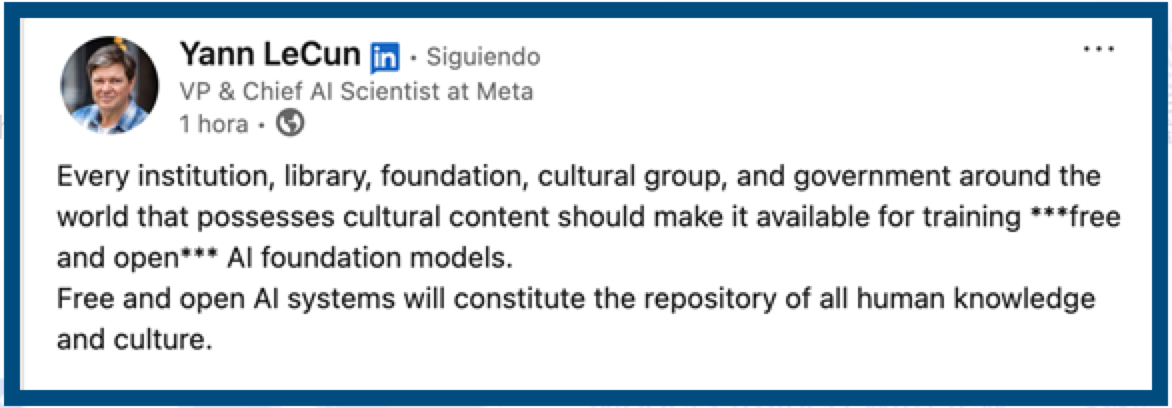
Current Models Reflect a Narrow Cultural Representation

- MLLMs are trained with data that is primarily focused on **few major languages and western-centric cultures.**
- This narrow representation makes them **biased**, have a limited world view, **poor** cultural knowledge and exhibit linguistic **imbalances.**



Current Models Reflect a Narrow Cultural Representation

- MLLMs are trained with data that is
- This narrow representation makes th
  linguistic imbalances.



Yann LeCun · Siguiendo
VP & Chief AI Scientist at Meta
1 hora ·

Every institution, library, foundation, cultural group, and government around the world that possesses cultural content should make it available for training ***free and open*** AI foundation models.
Free and open AI systems will constitute the repository of all human knowledge and culture.

MS-COCO

… weird looking vehicle…

… unusual bathroom …

… exotic fruits …

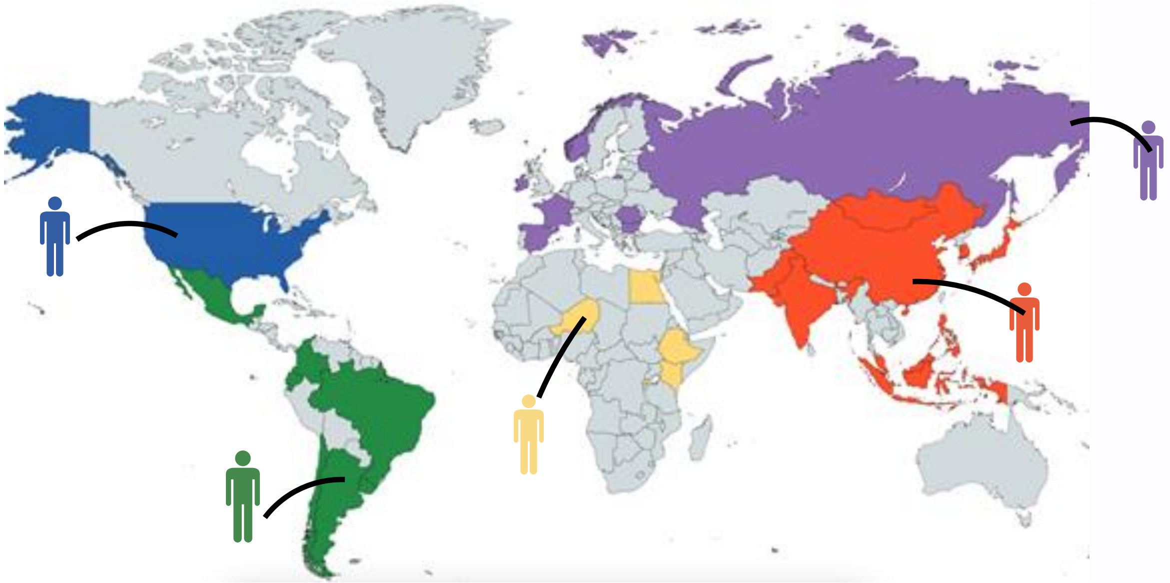
CVQA: Culturally-diverse Multilingual VQA Benchmark

- CVQA is a **multilingual, multiple-choice locally-nuanced** visual question-answering dataset.
- CVQA includes culturally-driven images and questions from across **30 countries on 4 continents.**
- Covering **31 languages with 13 scripts**, providing a benchmark with **10k questions.**

Annotation Process

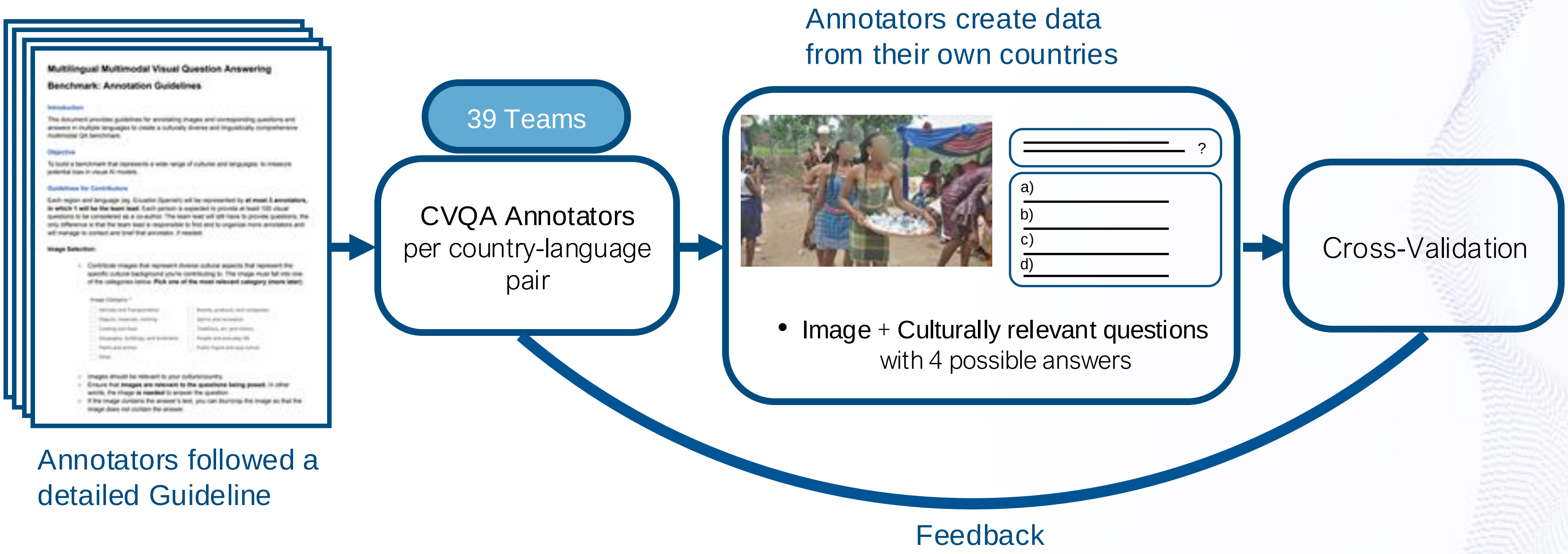
- CVQA follows a **crowd-sourcing collaboration approach**. We collaborated across communities.
- Annotators belong to various **NLP groups**, are fluent speakers and accustomed to the cultures of the locations

We collaborated with various NLP groups

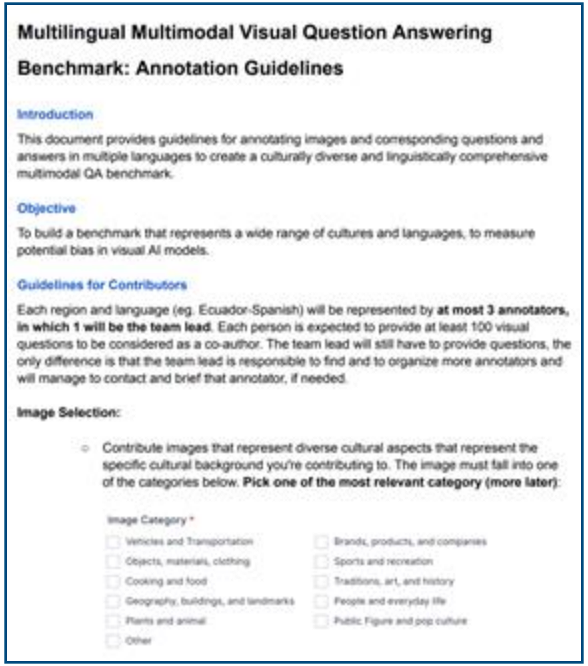


Annotation Process

- We group CVQA into **Country-Language pairs**, rather than simply on language or location.
- We developed concise **annotation guidelines** that are suitable for all Country-Language subset teams



Data Collection Design



Multilingual Multimodal Visual Question Answering Benchmark: Annotation Guidelines

Introduction

This document provides guidelines for annotating images and corresponding questions and answers in multiple languages to create a culturally diverse and linguistically comprehensive multimodal QA benchmark.

Objective

To build a benchmark that represents a wide range of cultures and languages, to measure potential bias in visual AI models.

Guidelines for Contributors

Each region and language (eg. Ecuador-Spanish) will be represented by **at most 3 annotators, in which 1 will be the team lead**. Each person is expected to provide at least 100 visual questions to be considered as a co-author. The team lead will still have to provide questions, the only difference is that the team lead is responsible to find and to organize more annotators and will manage to contact and brief that annotator, if needed.

Image Selection:

- Contribute images that represent diverse cultural aspects that represent the specific cultural background you're contributing to. The image must fall into one of the categories below. **Pick one of the most relevant category (more later)**:

Image Category *

- Vehicles and Transportation
- Objects, materials, clothing
- Cooking and food
- Geography, buildings, and landmarks
- Plants and animal
- Other
- Brands, products, and companies
- Sports and recreation
- Traditions, art, and history
- People and everyday life
- Public Figure and pop culture

Image Selection and Preparation:

- Images have to depict **diverse** cultural aspects.
- **Self-made images** are recommended but external images are allowed.
- **Anonymize** faces and text that can leak the answer.

Question Creation:

- Questions have to be **culturally relevant**.
- To answer the question, **the image must be required**.
- The questions need to be answerable **without** the multiple choices.

Data Collection Design

- We gathered images and created question-answer pairs based on the cultures of **various locations.**
- The question-answer pairs were created in their respective local languages, along with **parallel** English translations
- CVQA uses **common knowledge** as a proxy of culture, we define the following categories:

Categories

1. Vehicles and Transportation
2. Cooking and Food
3. People and Everyday Life
4. Sports and Recreation
5. Plants and Animals
6. Objects, Materials and Clothing
7. Brands and Products
8. Geography, Buildings, and Landmarks
9. Tradition, Art and History
10. Public Figure and Pop-Culture

CVQA Samples

Igbo - Nigeria

Category: Tradition/ Art / History – Igbo/Nigeria

Kedụ mmemme ndị a na-eme?
(Which ceremony are they doing?)

A. Ị gba nkwụ (Traditional marriage)
B. Ncheta Ọmụmụ (Birthday)
C. Emume cheiftaincy (Chieftaincy ceremony)
D. Emume iri ji ọhụrụ (New yam festival)

Malay - Malaysian

Category: People and everyday life – Malay/Malaysian

Roh manakah yang disembah dengan altar ini?
(Which deity is worshiped on this altar?)

A. Datuk Gong (Na Tuk Kong)
B. Buddha (Buddha)
C. Brahma (Brahma)
D. Vishnu (Vishnu)

Spanish - Mexico

Category:Tradition / Art / History – Spanish/Mexico

¿Qué se muestra en la imagen? (What is shown in the image?)
A. el calendario azteca/ piedra del sol (the aztec calendar/ aztec sun stone)
B. una serpiente azteca (an aztec serpent)
C. coatlicue (coatlicue)
D. tláloc (tlaloc)

Korean - South Korea

Category: Object, Clothing, and Material – Korean/South Korea

이런 종류의 요리에 사용되는 그릇을 무엇이라고 부르나요?
(What is this type of bowl called in cooking?)

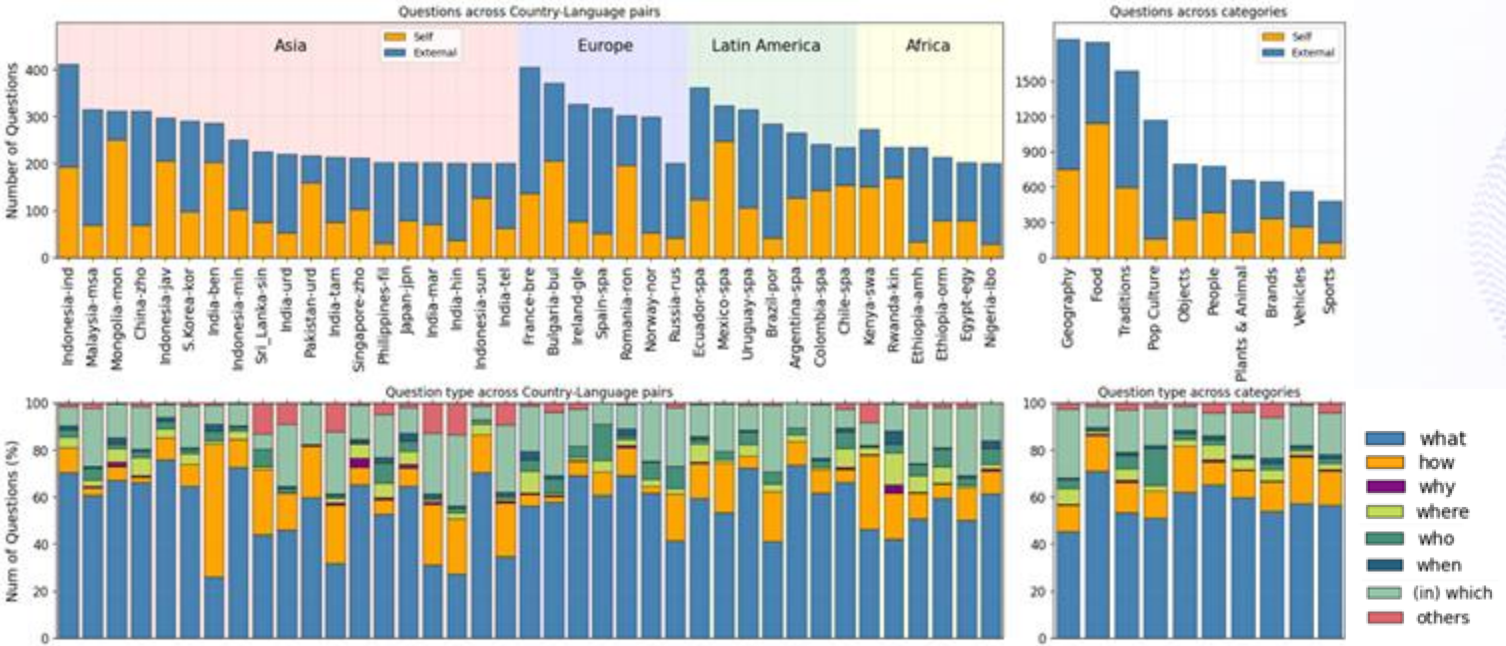
A. 돌솥 (Dolsot)
B. 복주머니 (Bokjumeoni)
C. 냄비 (Pot)
D. 팬 (Pan)

MBZUAI

CVQA: Data Statistics

- No. of countries: 30
- No. of languages: 31
- No. of images: 5,239
- No. of questions: 10,374

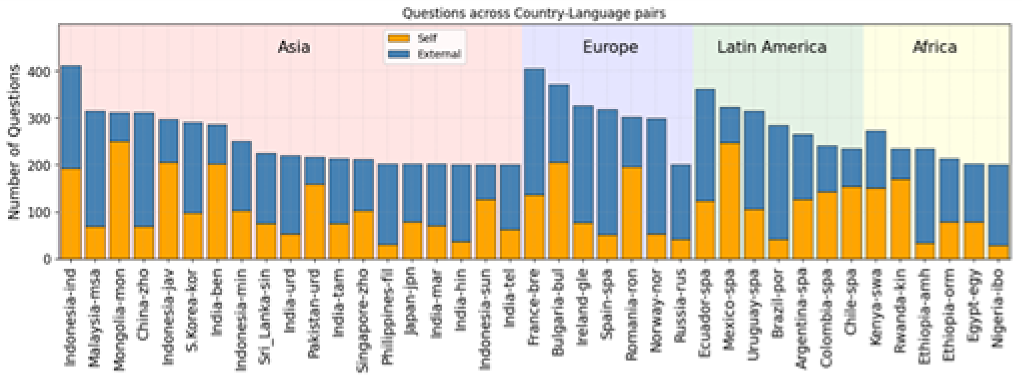
- No. of scripts: 13
- No. of country-language pair: 39
- Avg. questions per image: 1.98
- Avg. words per question: 7.6



CVQA: Data Statistics

Country	Language	Script
Africa		
Egypt	Egyptian Arabic	Arabic
Ethiopia	Amharic	Amharic
Ethiopia	Oromo	Latin
Kenya	Swahili	Latin
Nigeria	Igbo	Latin
Rwanda	Kinyarwanda	Latin
Asia		
China	Chinese	Chinese
India	Bengali	Bengali
India	Hindi	Devanagari
India	Marathi	Devanagari
India	Tamil	Tamil
India	Telugu	Telugu
India	Urdu	Perso-Arabic
Indonesia	Indonesian	Latin
Indonesia	Javanese	Latin
Indonesia	Minangkabau	Latin
Indonesia	Sundanese	Latin
Japan	Japanese	Japanese
South Korea	Korean	Hangul
Malaysia	Malay	Latin
Mongolia	Mongolian	Cyrillic
Pakistan	Urdu	Perso-Arabic
Philippines	Filipino	Latin
Singapore	Chinese	Chinese
Sri Lanka	Sinhala	Sinhalese
Europe		
Bulgaria	Bulgarian	Cyrillic
France	Breton	Latin
Ireland	Irish	Latin
Norway	Norwegian	Latin
Romania	Romanian	Latin
Russia	Russian	Cyrillic
Spain	Spanish	Latin
Latin America		
Argentina	Spanish	Latin
Brazil	Portuguese	Latin
Chile	Spanish	Latin
Colombia	Spanish	Latin
Ecuador	Spanish	Latin
Mexico	Spanish	Latin
Uruguay	Spanish	Latin

- No. of countries: 30
- No. of languages: 31
- No. of scripts: 13



CVQA covers several less commonly studied languages and regions →

- Ireland-Irish
- Indonesia-Minangkabau
- India-Tamil
- France-Breton
- Nigeria-Igbo
- Mongolia-Mongolian

- Kenya-Swahili
- Egypt-Egyptian Arabic
- Ecuador-Spanish
- Argentina-Spanish
- Brazil-Portuguese
- South Korea - Korean

Evaluation of Open and Closed-source Models

- Among **open models**, LLaVa achieves the best performances but still **significantly behind** closed models.
- All models obtain **worse performances when the question is asked in local languages**, emphasizing the models lower capability in handling non-English prompts.

Table 3: Average performance of MLLMs on our CVQA dataset with English prompts (EN) and local language prompts (LOC).

LLaVA-1.5-7B		M-CLIP		CLIP		mBLIP-mT0		mBLIP-BLOOMZ		InstructBLIP		Gemini-1.5-Flash		GPT-4o	
EN	LOC	EN	LOC	EN	LOC	EN	LOC	EN	LOC	EN	LOC	EN	LOC	EN	LOC
49.6	35.5	38.0	33.7	42.7	30.6	31.3	30.9	39.3	32.7	49.0	31.9	66.9	68.5	75.4	74.3

Evaluation of Open and Closed-source Models

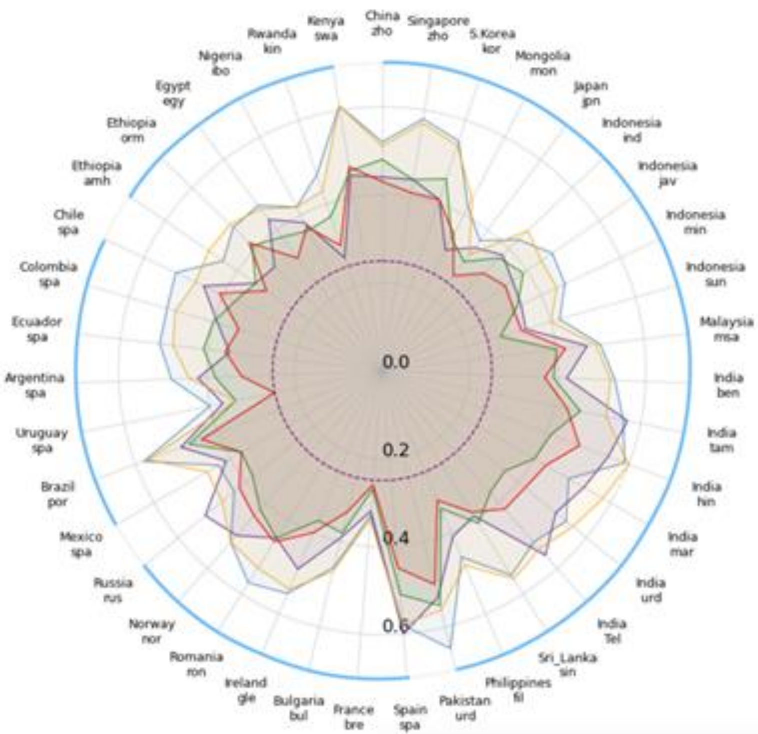
- Among **open models**, LLaVa achieves the best performances but still **significantly behind** closed models.
- All models obtain **worse performances when the question is asked in local languages**, emphasizing the models lower capability in handling non-English prompts.

Table 3: Average performance of MLLMs on our CVQA dataset with English prompts (EN) and local language prompts (LOC).

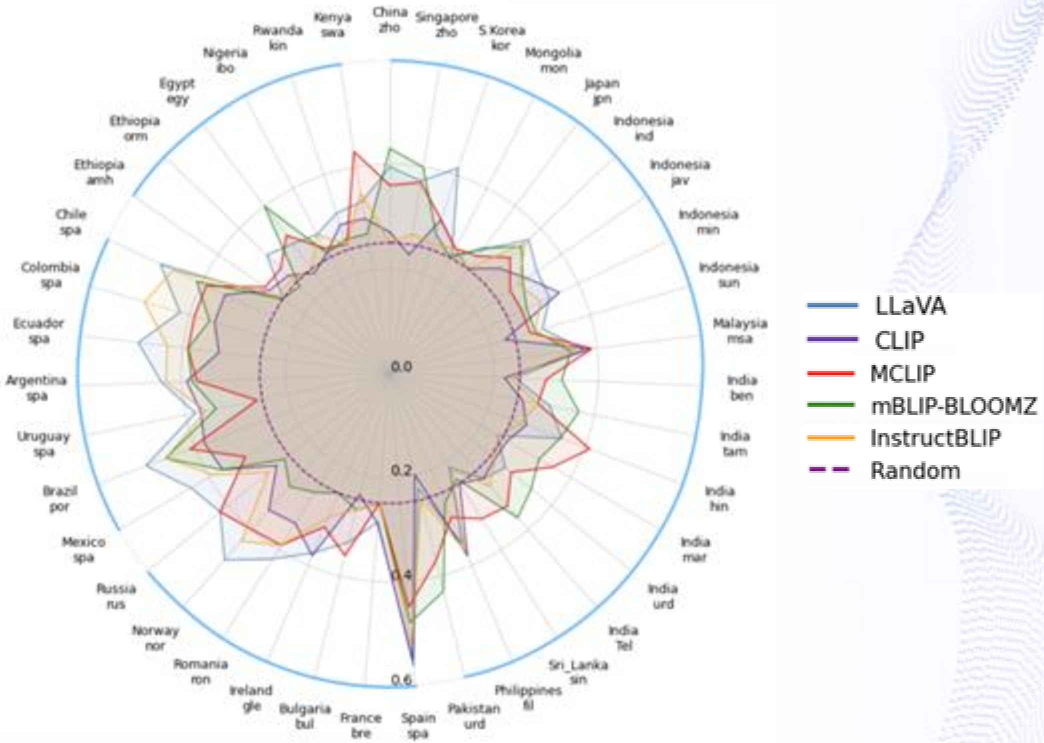
LLaVA-1.5-7B		M-CLIP		CLIP		mBLIP-mT0		mBLIP-BLOOMZ		InstructBLIP		Gemini-1.5-Flash		GPT-4o	
EN	LOC	EN	LOC	EN	LOC	EN	LOC	EN	LOC	EN	LOC	EN	LOC	EN	LOC
49.6	35.5	38.0	33.7	42.7	30.6	31.3	30.9	39.3	32.7	49.0	31.9	66.9	68.5	75.4	74.3

Performance per Country-Language Pair

English-only question-option pairs



Local-language question-option pairs



Performance across Categories

- **People and Everyday life** achieves the best accuracies across most of the models.
- **Cooking & Food and Pop culture** exhibit low accuracies, demonstrating that the high diversity of these categories across different cultures poses a great challenge for MLLMs.

Table 5: Accuracy of models across categories. Per category, the best performing models on English (EN) and local language (LOC) question-option pairs are bolded and underlined, respectively.

Categories	LLaVA-1.5-7B		M-CLIP		CLIP		mBLIP-mT0		mBLIP-BLOOMZ		InstructBLIP	
	EN	LOC	EN	LOC	EN	LOC	EN	LOC	EN	LOC	EN	LOC
Brands	**49.9**	<u>36.5</u>	37.2	35.7	36.6	29.7	33.7	30.8	40.5	35.1	48.4	32.6
Food	**45.4**	<u>31.9</u>	34.5	29.1	39.2	30.4	28.1	27.6	37.7	29.8	44.4	30.6
Geography	**47.1**	<u>38.2</u>	37.1	34.2	41.8	31.9	30.6	31.6	35.0	32.3	45.3	33.2
Objects	51.8	33.0	39.4	34.5	39.7	25.4	34.3	33.0	43.1	34.0	**52.3**	29.1
People	58.9	38.1	45.0	37.8	46.8	30.9	35.3	34.7	46.3	36.7	**59.8**	34.0
Plants & Animals	**55.7**	<u>37.5</u>	43.7	32.0	48.0	27.2	35.2	35.5	46.0	36.0	55.4	35.1
Pop Culture	44.5	<u>36.3</u>	33.7	31.5	**46.1**	36.3	28.8	29.9	35.7	30.7	45.1	34.6
Sports	**50.7**	<u>39.1</u>	39.3	33.3	43.5	32.4	32.6	31.4	40.1	34.9	50.5	34.7
Tradition	**50.4**	<u>35.8</u>	37.0	35.2	41.9	32.2	31.6	31.5	39.0	32.2	47.9	30.8
Vehicles	50.6	<u>41.4</u>	39.5	41.1	44.6	30.5	35.6	33.9	42.0	34.0	**55.0**	33.0

Performance across Categories

- **People and Everyday life** achieves the best accuracies across most of the models.
- **Cooking & Food and Pop culture** exhibit low accuracies, demonstrating that the high diversity of these categories across different cultures poses a great challenge for MLLMs.

Table 5: Accuracy of models across categories. Per category, the best performing models on English (EN) and local language (LOC) question-option pairs are bolded and underlined, respectively.

Categories	LLaVA-1.5-7B		M-CLIP		CLIP		mBLIP-mT0		mBLIP-BLOOMZ		InstructBLIP	
	EN	LOC	EN	LOC	EN	LOC	EN	LOC	EN	LOC	EN	LOC
Brands	**49.9**	36.5	37.2	35.7	36.6	29.7	33.7	30.8	40.5	35.1	48.4	32.6
Food	**45.4**	31.9	34.5	29.1	39.2	30.4	28.1	27.6	37.7	29.8	44.4	30.6
Geography	**47.1**	<u>38.2</u>	37.1	34.2	41.8	31.9	30.6	31.6	35.0	32.3	45.3	33.2
Objects	51.8	33.0	39.4	<u>34.5</u>	39.7	25.4	34.3	33.0	43.1	34.0	**52.3**	29.1
People	58.9	<u>38.1</u>	45.0	37.8	46.8	30.9	35.3	34.7	46.3	36.7	**59.8**	34.0
Plants & Animals	**55.7**	37.5	43.7	32.0	48.0	27.2	35.2	35.5	46.0	36.0	55.4	35.1
Pop Culture	44.5	36.3	33.7	31.5	**46.1**	36.3	28.8	29.9	35.7	30.7	45.1	34.6
Sports	**50.7**	<u>39.1</u>	39.3	33.3	43.5	32.4	32.6	31.4	40.1	34.9	50.5	34.7
Tradition	**50.4**	<u>35.8</u>	37.0	35.2	41.9	32.2	31.6	31.5	39.0	32.2	47.9	30.8
Vehicles	50.6	<u>41.4</u>	39.5	41.1	44.6	30.5	35.6	33.9	42.0	34.0	**55.0**	33.0

Performance across Image Source

- For self-made images, the performance of LLaVa and CLIP tends to be **lower compared to web images.**
- While this is **not consistent** across all models, this indicate that **web-images might be more representative** of the data used to train these models.

Table 6: Accuracy of different models divided by image source

Image Source	LLaVA-1.5-7B		M-CLIP		CLIP		mBLIP-mT0		mBLIP-BLOOMZ		InstructBLIP	
	EN	LOC	EN	LOC	EN	LOC	EN	LOC	EN	LOC	EN	LOC
Self-made Image	48.8	34.2	38.1	34.3	41.2	30.1	31.2	31.5	40.1	33.4	48.3	31.5
Web Image	49.7	37.4	37.4	33.3	43.1	31.8	31.9	31.2	38.7	32.3	49.1	33.1

Paper, Data and Leaderboard

Paper

Data

Leaderboard

Labels are Open !!

SEA-VL

Welcome to SEACrowd! 👋

We are a community dedicated to bridging the gap between multilingual AI and Southeast Asian AI and enhancing the quality of AI research and researchers in the region.

See what indigenous and non-indigenous languages are under our study.

🚨 *Call for contributors!* 🚨

*Following the success of our SEACrowd project, we're excited to announce **SEA-VL**, a new open-source initiative to create high-quality vision-language datasets for Southeast Asian (SEA) languages! Check it out here.*

Data Collection Scheme

Task 1: Submit a SEA Culturally-Relevant Image (1-2 points per photo)

Submission is simple! Just go to this form and provide your self-taken, culturally relevant photo with a brief description.

Points:

- 2 points ~~1 point~~ for images from Indonesia, Singapore, and Phillippines
- 3 points ~~1.5 points~~ for images from Thailand, Malaysia, and Vietnam
- 4 points ~~2 points~~ for images from Brunei, East Timor, Cambodia, Laos, Myanmar

Task 2: Review Image-Description Pairs (1 point per review)

To participate in reviewing, contributors must first pass this short screening test. Check our annotation guideline to learn more!

Authorship

Why Contribute?

As with SEACrowd, every contribution to SEA-VL will earn points. Reaching 200 points in Phase 1 will guarantee co-authorship in our publication for ACL 2025. You'll also be eligible for our exclusive merch once you surpass 300 points in Phase 1!

Example: Image Tracking

[SEA-VL] SEA Culturally-Relevant Image Collection (Responses) : Monitor

Total submissions	5134
Unique contributors	75

What kind of images have we collected?

Based on image location:
Brunei: 61, Cambodia: 63, East Timor: 9, Indonesia: 1917, Laos: 76, Malaysia: 309, Myanmar: 5, Phillippines: 85, Singapore: 1168, Thailand: 832, Vietnam: 381, Others: 228

No	Submission time	Image location	Image caption (English)
5134	23 Jan 2025	Phoenix, United States	Bengbeng, a chocolate snack from Indonesia
5133	23 Jan 2025	Phoenix, United States	Soto ayam with boiled egg
5132	23 Jan 2025	Victoria, Canada	Peanut sauce chicken satay with lontong and acar
5131	23 Jan 2025	Urbana, United States	Rupiah bank notes when Indonesia was under Japanese occupation
5130	23 Jan 2025	Urbana, United States	A picture of Garuda Pancasila, Indonesian national emblem, a picture of Soekarno, and old Rupiah bank notes
5129	23 Jan 2025	Urbana, United States	A keris, its sheath, and the belt used to hold it
5128	23 Jan 2025	Urbana, United States	A keris decorated by agates along with its sheath
5127	23 Jan 2025	Urbana, United States	Tenun machine from Indonesia used to weave songket
5126	23 Jan 2025	St Louis, United States	Orangutan, an animal native to Kalimantan Island
5125	23 Jan 2025	Urbana, United States	Malam and canting used to make batik

Example: Image Validation

Is photo quality OK?

- Yes[1]
- Unsure[2]
- No[3]

The image portrays culturally-relevant information in:

Vietnam

The image was taken in (City, Country):

Hanoi, Vietnam

Is the image culturally relevant in South-East Asia?

- Yes. Unique to SEA.[4]
- Yes, people will likely think of SEA when seeing the picture, but it may have low degree of similarity to other cultures.[5]
- Maybe, this culture did not originate from SEA, but it's quite dominant in SEA.[6]
- Not really. It has some affiliation to SEA, but actually does not represent SEA or has stronger affiliation to cultures outside SEA.[7]
- No. Totally unrelated to SEA.[8]

How do you know about this culture?

Please do not consult LLMs (e.g., GPT-4o, Claude, Command-R, etc.)

- I'm from this country/culture.[9]
- I checked online resources (e.g., Wikipedia, articles, blogs).[0]

Challenges and Conclusion

- Community-based image collection enables us to collect data across diverse culture
- Scale up is difficult
- Ensuring quality needs effort; proper data validation, proper annotation guideline; ensure the annotators are motivated