CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark Supplementary Material

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

In this Supplementary Material, we present the following items:

1. A datasheet for the dataset documentation of CVQA (Section 1).
2. The data access and maintenance plan (Section 2).
3. The Statement of responsibility (Section 3).
4. The annotation guideline provided to CVQA annotators (Section 4).
5. Annotation Platform (Section 5).
6. Additional Statistics (Section 6)

1 Datasheet for CVQA

For documenting CVQA, we use the datasheet for datasets introduced by Gebru et al. [1], which specify the motivation, composition, collection process, preprocessing, uses and distribution of a dataset. We follow and provide this datasheet for CVQA below:

Motivation

[*] Equal Contribution

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

Q1. **For what purpose was this dataset created?** *Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

We aim to address the limitations of current Visual Question Answering (VQA) datasets, which predominantly focus on English and Western-centric images. These datasets lack diversity, especially in low-resource languages and culturally varied images. To overcome these issues, we introduce CVQA, a new benchmark designed to include culturally-driven images and questions from 28 countries covering 26 languages and 11 scripts. This benchmark aims to enhance the evaluation of multimodal AI models, encouraging the development of models with better cultural awareness and linguistic diversity.

Q2. **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

CVQA is a collaborative movement involving many people from different institutions and communities. The CVQA is led by a team of researchers from MBZUAI.

Q3. **Who funded the creation of the dataset?** *If there is an associated grant, please provide the name of the grantor and the grant name and number.*

No grant, all expenses were funded by the MBZUAI's faculty startup fund.

Q4. **Any other comments?**

No.

Composition

Q5. **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

We provide a test set that contains instances of image-question pairs. Specifically, each instance is a dictionary that contains: `image`, `ID`, `Subset`, `Question`, `Translated Question`, `Options`, `Translated Options`, `Label`, `Category`, `Image Type`, `Image Source`, `License`. We provide a more detailed description of each field and an example in the README of `https://huggingface.co/datasets/afaji/cvqa`.

Q6. **How many instances are there in total (of each type, if appropriate)?**

CVQA contains a `test-split` of 9044 instances of image-question pairs.

Q7. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

Yes, it contains all instances.

Q8. **What data does each instance consist of?** *"Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.*

We provide raw annotations, where each instance consists of an image, a question and four answer candidates - of which only one is correct.

Q9. **Is there a label or target associated with each instance?** *If so, please provide a description.*

Yes, we provide four answer candidates for each question; among each set of four, we have labelled one as correct.

Q10. **Is any information missing from individual instances?** *If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

No.

Q11. **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** *If so, please describe how these relationships are made explicit.*

Some of the images are associated with more than one question (max three).

Q12. **Are there recommended data splits (e.g., training, development/validation, testing)?** *If so, please provide a description of these splits, explaining the rationale behind them.*

All data is for test purposes.

Q13. **Are there any errors, sources of noise, or redundancies in the dataset?** *If so, please provide a description.*

The data is human-written so it is bound to errors such as typos or grammatical errors. However, we argue that these errors are naturally made and (in very small amounts) are good for benchmarking the model's robustness, as the purpose of creating the dataset.

Q14. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** *If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

It is self-contained in `https://huggingface.co/datasets/afaji/cvqa`

Q15. **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** *If so, please provide a description.*

No.

Q16. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** *If so, please describe why.*

No.

Q17. **Does the dataset relate to people?** *If not, you may skip the remaining questions in this section.*

No.

Q18. **Does the dataset identify any subpopulations (e.g., by age, gender)?**

By country.

Q19. **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** *If so, please describe how.*

No.

Q20. **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** *If so, please provide a description.*

Yes - ethnic origins. The purpose of this dataset is to capture culture in the images. However, as public faces are blurred/ unrecognisable, this feature does not pose any risk.

Q21. **Any other comments?**

No.

Collection Process

Q22. **How was the data associated with each instance acquired?** *Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the* *data validated/verified? If so, please describe how.*

The image is obtained from either web or self-made images, we selected web images with Creative Commons license. Questions and options are written by annotators.

Q23. **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** *How were these mechanisms or procedures validated?*

Manual human curation. See Section 5 and 4 for details about our annotation platform and guideline. For more information, please refer to our paper.

Q24. **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

N/A

Q25. **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

Those involved in the data collection process have been named as co-authors.

Q26. **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?** *If not, please describe the timeframe in which the data associated with the instances was created.*

The data was collected from 2023 to 2024.

Q27. **Were any ethical review processes conducted (e.g., by an institutional review board)?** *If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

Ethical Review was not required for this dataset collection.

Q28. **Does the dataset relate to people?** *If not, you may skip the remaining questions in this section.*

No, Annotators were asked to provide data about their country's culture but not specific to any individual or groups of people.

Q29. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

N/A.

Q30. **Were the individuals in question notified about the data collection?** *If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

N/A.

Q31. **Did the individuals in question consent to the collection and use of their data?** *If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

N/A.

Q32. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** *If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

N/A.

Q33. **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** *If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

N/A.

Q34. **Any other comments?**

No.

Preprocessing, Cleaning and/or Labeling

Q35. **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** *If so, please provide a description. If not, you may skip the remainder of the questions in this section.*

Yes, we removed all image metadata, automatically blurred faces and text that would reveal the answer, and finally removed images that had invalid licenses.

Q36. **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** *If so, please provide a link or other access point to the "raw" data.*

No.

Q37. **Is the software used to preprocess/clean/label the instances available?** *If so, please provide a link or other access point.*

We used a standard Python code to clean and preprocess the final instances of CVQA. During the data collection process we allowed the annotators to use "PicdeFacer", a tool that can be used for blurring faces or information: `https://picdefacer.com/en/`.

Q38. **Any other comments?**

No.

Uses

Q39. **Has the dataset been used for any tasks already?** *If so, please provide a description*

Our dataset have not been used for other tasks yet. We only use it in our paper to benchmark various models.

Q40. **Is there a repository that links to any or all papers or systems that use the dataset?** *If so, please provide a link or other access point.*

No.

Q41. **What (other) tasks could the dataset be used for?**

Primarily for benchmarking Cultural multilingual visual QA, but this dataset can potentially be used for Machine Translation and language learning game (CALL).

Q42. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** *For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

No.

Q43. **Are there tasks for which the dataset should not be used?** *If so, please provide a description*

No.

Q44. **Any other comments?**

No.

Distribution

Q45. **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** *If so, please provide a description.*

Yes, the data has been publicly released.

Q46. **How will the dataset be distributed (e.g., tarball on website, API, GitHub)?** *Does the dataset have a digital object identifier (DOI)?*

The data is available on Huggingface at: `https://huggingface.co/datasets/afaji/cvqa`.

Q47. **When will the dataset be distributed?**

CVQA is already available from June 2024 and onward.

Q48. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

Note that each instance has its own license. All data is free to use for research purposes, but not every entry is permissible for commercial use.

Q49. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

Yes, some images on Flickr, are under copyright. We advised annotators to only select those available for non-commercial use: `https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en`. We also automatically remove entries that do not conform to the copyright requirement.

Q50. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

No.

Q51. **Any other comments?**

No.

Maintenance

Q52. **Who will be supporting/hosting/maintaining the dataset?**

CVQA team at MBZUAI.

Q53. **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

You can contact the main team via email or through starting a new discussion on the CVQA Hugging Face page.

Q54. **Is there an erratum?** *If so, please provide a link or other access point*

N/A.

Q55. **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** *If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*

Yes, updates will be made on Huggingface once we have more data (e.g. new country-language pairs) or there are reported errors in the data.

Q56. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** *If so, please describe these limits and explain how they will be enforced.*

N/A.

Q57. **Will older versions of the dataset continue to be supported/hosted/maintained?** *If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*

N/A.

Q58. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** *If so, please provide a description. Will these contributions be validated/verified?*

If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Yes, it will be conducted through communications with CVQA team.

Q59. **Any other comments?**

No.

2 Data Access and Maintenance Plan

We publicly released CVQA, and it is available to download from Hugging Face: `https://huggingface.co/datasets/afaji/cvqa`. To assess model performance, we also created a leaderboard in `eval.ai` platform: `https://eval.ai/web/challenges/challenge-page/2305/overview`. Detailed information is provided in `www.cvqa-benchmark.org`. We, the authors, will be responsible for handling CVQA issues and maintaining the data accordingly.

3 Statement of Responsibility

We, the authors, bear all responsibilities in case of rights violations in CVQA. Please note that each image question in our dataset has its own distinct license. Although our data is free to use for research purposes, not all instances in CVQA are permissible for commercial use.

4 Annotation Guideline

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)**:

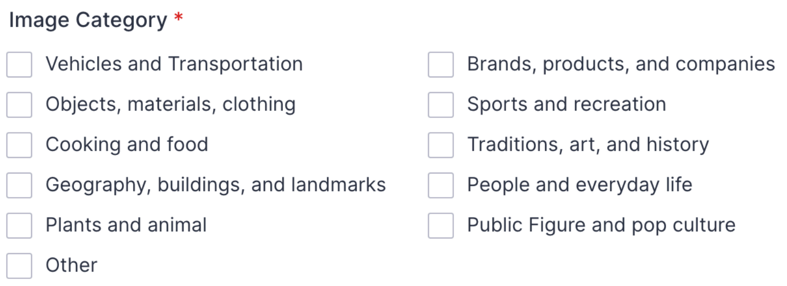


Image Category *

□ 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

- Images should be relevant to your culture/country.
- Ensure that **images are relevant to the questions being posed.** In other words, the image **is needed** to answer the question.
- If the image contains the answer’s text, you can blur/crop the image so that the image does not contain the answer.

- Image source:
    - 1. Self/personal picture **(highly preferable).** You may ask your family/friend to donate their photos, if possible.
    - 2. We also accept external images from:
        - Flickr: https://www.flickr.com/explore (**please make sure the associated license to the image is Creative Commons**), this can be selected at the the top left of Flickr ("Any License").
        - WikimediaCommons: https://commons.wikimedia.org/wiki/Main_Page (**here you do not need to select any license for the images)**,
        - Unsplash: https://unsplash.com/ (**please make sure to search the image first and they select the license: Free**). More details (Tutorial) at the end of this document.
        - Dollar Street: https://www.gapminder.org/dollar-street (**here you do not need to select any license for the images),** this webpage has images only from some countries, please make sure to select your country to find images if applicable.

          More detailed instructions for each web page are shown at the end of this document.
    - If you use an external image, you'll need to put the url of the original image.
- The image must be reasonable quality (not pixelated or blurry, can be understandable). You can upload images of any ratio as long as it is not too tall or wide (e.g.: don't submit panorama pictures).
- Do not show personally identifiable information (PII) such as faces, car plates, house addresses. Faces of public figures or fictional characters are ok. Also, **please be sure to blur text in the image that will leak the answer**. "PicdeFacer" can be used for blurring: https://picdefacer.com/en/. Tutorial on using PicdeFacer is shown at the end of this document.

Question and Answer Creation:

After finding the image, you must now formulate 1-3 questions + answers from that image. Specifically:

- The question must be answerable **only by looking at the image.**
- Ensure that the questions are culturally relevant and specific to the image content.
- Provide answers that are concise, accurate, and directly related to the question.
- You will also need to provide 1 correct option and 3 other incorrect options (distractors). For the distractors, choose options that are relevant, not obvious wrong answers.

- **The question must be answerable even without the multiple-choice.** Example of the invalid question: ("What song is not performed by this musician" – not answerable if you don't know the choices)
- Make sure the questions are **written fluently in both the local language and English.** Use a grammar checker if needed i.e. if you are not fluent in English.
- Be mindful of cultural sensitivities and avoid stereotyping or misrepresenting cultural aspects.
- Ensure there are **variations on your question**. Identity questions are fine, eg "What is this", or "where is this". But additionally adding more complex/difficult questions would be great. For example, multi-hop reasoning, counting, referencing, or questions that require local commonsense knowledge to be answered.

Category Definition

When selecting a category, pick one of the most relevant. Please follow the guideline:

- Vehicles and Transportation: Local public transport, local vehicles.

- **Objects, Materials, Clothing:** Questions about local/traditional clothes. Unique/local tools or items.

- **Cooking and Food:** Local dishes and food/drink. This category includes native fruits in the context of the image if that fruit is served as a food/drink.

- **Geography, Buildings, Landmarks:** Popular/common landmarks, local architecture/buildings. Local monuments.

- **Plants and Animals:** Plants and animals commonly found in the region.

- **Brands, Products, and Companies:** Questions about understanding local yet popular brands or companies. Even if the brand is about food/transportation, if the main focus of the question is the brand recognition itself, then it should be under this category.

- **Sports & Recreation:** Local sports and fun activities. Focuses on the activity itself rather than the location (in that case, it goes to the 'landmark' category).

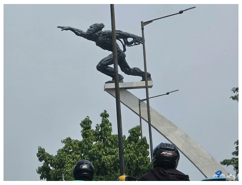
- **Tradition, Art, History:** Local ceremonies/festivals/events, local dance/music, folklores. Historical artifacts.

- **People & Everyday Life:** Focuses on the people themselves: i.e., common habits/customs, common occupations and jobs, routine religious activities, everyday activities/routines.

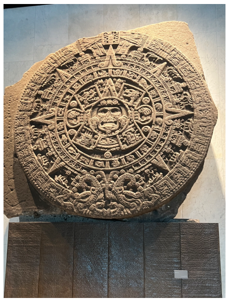
- **Public Figures & Pop Culture:** Questions on the understanding of common public figures (e.g., politicians, artists, musicians, etc.). Common pop culture such as movies and games. If the category is still ambiguous to you, pick the one you think is the most appropriate.

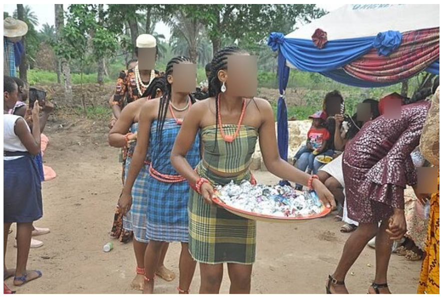
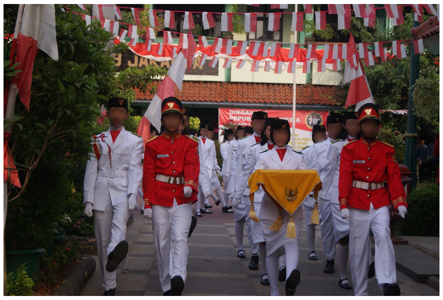
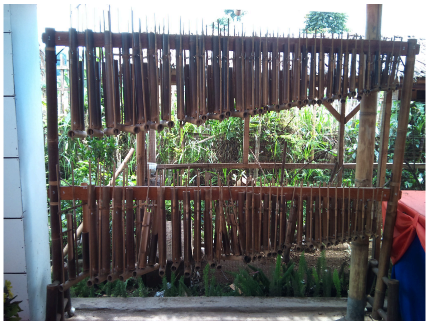
Examples

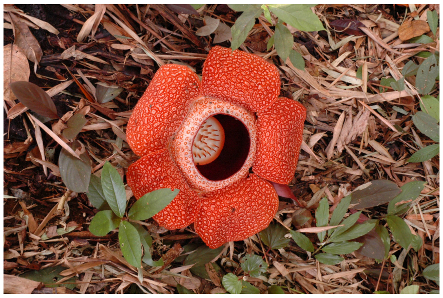
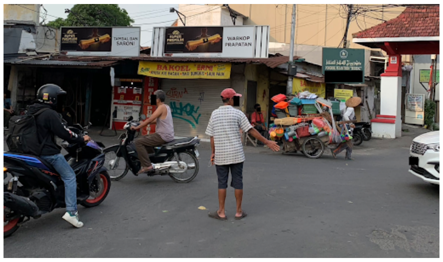
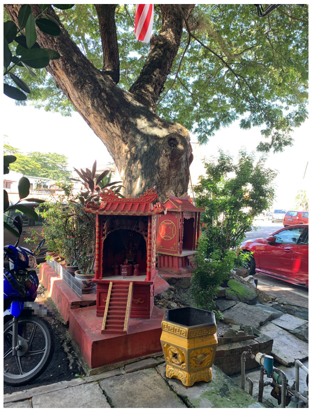
Examples that can be improved

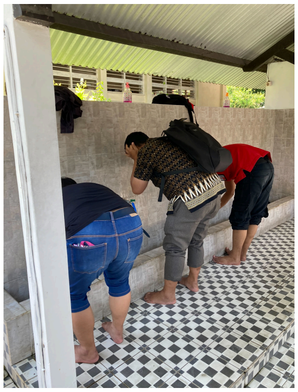
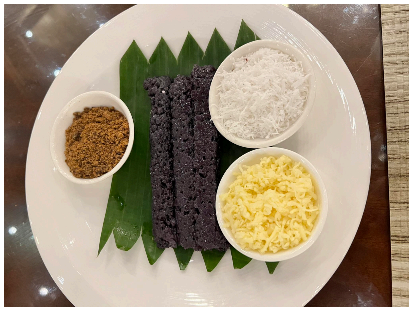
	Make sure the image is needed to respond the question, example: 1) ¿En qué mes se celebra esta fiesta? (In which month is this celebration held?) Correct 2) ¿En qué mes se celebra la fiesta de la "Mama Negra"? (In which month is the celebration of the "Mama Negra" held?) Wrong–As this question can be answered without looking at the image.
	**Make sure the question is not ambiguous:** 1) Where is this monument located? Wrong–Not specific, the answer could be a city, country, province,etc. 2) In which **city** is this monument located? Correct–specifically asking about the city
	Make sure the question is not too vague: 1) What is this? Question wording can be more specific 2) What is the name of this vehicle? Correct–specifically asking about the vehicle name.

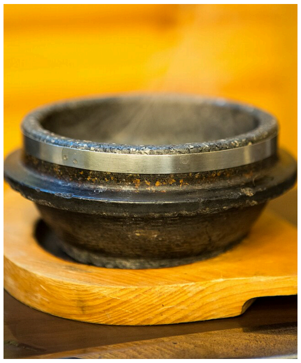
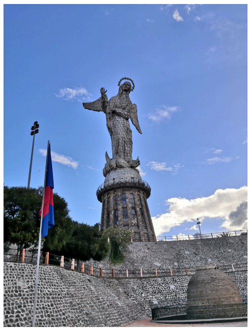
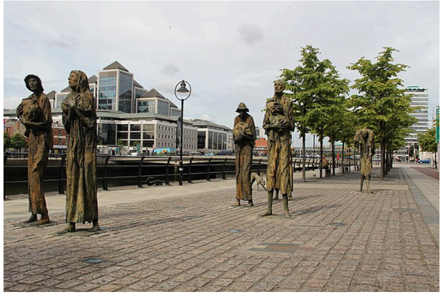
Acceptable examples

	**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) ¿En dónde se exhibe esta pieza? (Where is this piece exhibited?) **A. En el museo nacional de antropología (In the National Museum of Atroplogy)** B. en el castillo de Chapultepec (In the Chapultepec Castle) C. En el zócalo de la ciudad de Mexico (In the Mexico City zocalo) D. En Teotihuacan (In Teotihuacan)

	**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) Kedụ ebe a na-eme mmemme a? (Where is this ceremony held?) **A. Ụlọ nna nwunye (The home of the bride's father)** B. Ahia (The market) C. Ụlọ nna di (The home of the groom's father) D. Ụlọ nsọ (The church)
	**Category: Tradition/ Art / History – Indonesian/Indonesia** Pada tahun berapakah foto ini diambil? (In what year is this photo taken?) **A. 2015 (2015)** B. 2020 (2020) C. 2023 (2023) D. 2010 (2010) Apa nama pasukan yang ada di foto ini? (What is the name of the squad in this photo?) **A. Paskibraka (Paskibraka)** B. Brimob (Brimob) C. TNI (TNI) D. ABRI (ABRI) Apa tugas utama pasukan ini? (What is the main purpose of this squad?) **A. Mengibarkan bendera (Hoisting the flag)** B. Mengawal presiden (Escorting president) C. Menjaga keamanan (Maintaining security) D. Mengiringi pengantin (Accompanying the bride and groom)
	**Category: Tradition, Art, History – Sundanese/Indonesia** Naon kagunaan ieu hiji alat? (What is the use of this tool?) **A. Alat musik (Musical instrument)** B. Alat pertahanan diri (Self defence tool) C. Jemuran (Clothes drying equipment) D. Alat masak (Cooking tool) – Ieu hiji alat teh asalna ti propinsi mana di Indonesia? (This tool comes from which province in Indonesia?) **A. Jawa Barat (West Java)** B. Bali (Bali)

	C. Bengkulu (Bengkulu) D. Sumatra Barat (West Sumatra)
	**Category: Plants and animals – Malay/Malaysia** Apakah nama bunga dalam gambar ini? (What is the name of the flower in this picture?) **A. Pakma (Rafflesia)** B. Bunga raya (Hibiscus) C. Anggerik (Orchid) D. Bunga kertas (Bougainvillea) Di rantau Asia manakah bunga itu boleh ditemui? (In which region of Asia can the flower be found?) **A. Asia Tenggara (Southeast Asia)** B. Asia Timur (East Asia) C. Asia Selatan (South Asia) D. Asia Tengah (Central Asia)
	**Category: People and everyday life - Javanese/Indonesia** Opo arane wong seng nang tengah embong iki? (What is the term for the man in the middle of the road?) **A. Polisi cepek (Polisi cepek)** B. Tukang parkir (Parking assistance man) C. Mlijo (Grocery man) D. Tukang becak (Pedicap man) Opo seng dilakukno wong seng nang tengah dalan iku? (What does the man in the middle of the road do?) **A. Ngatur prapatan (Managing the intersection)** B. Njaluk donasi (Asking for donations) C. Ngawasi pelanggaaran lalu lintas (Looking out for traffic violations) D. Nunjukno arah (Showing directions)
	**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) Apakah agama yang diamalkan oleh pengguna altar ini? (What religion do the users of these altars practice?) **A. Taoism (Taoisme)** B. Buddha (Buddhisme) C. Islam (Islam) D. Hindu (Hinduisme)

	**Category: People and Everyday Life – Indonesian/Indonesia** Apa yang orang-orang ini lakukan? (What are these people doing?) **A. Berwudhu (Performing ablution)** B. Mandi (Taking a bath) C. Yoga (Yoga) D. Beribadah (Praying) Dimana biasanya orang-orang melakukan aktivitas di foto ini? (Where do people usually do the activity in this photo?) **A. Masjid (Mosque)** B. Gereja (Church) C. Pemandian umum (Public bath) D. Gym (Gym)
	**Category: Cooking and Food – Tagalog/Philippines** Anong tawag sa kakanin na ito? (What is the name of this rice cake?) **A. Puto Bumbong (Puto Bumbong)** B. Suman (Suman) C. Kutsinta (Kutsinta) D. Sapin-Sapin (Sapin-Sapin) Tuwing kailan ito madalas tinitinda sa Pilipinas? (When is this food usually sold in the Philippines?) **A. Christmas Season (Christmas Season)** B. Independence Day (Independence Day) C. Labor Day (Labor Day) D. National Heroes Day (National Heroes Day) Ano tawag dun sa brown? (What do you call the brown object?) **A. Muscovado (Muscovado)** B. Latik (Toasted coconut) C. Chocolate (Chocolate) D. Caramel (Caramel)

	**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) 그릇의 재질은 무엇인가요? (What is the material of the bowl?) **A. 돌 (Stone)** B. 도자기 (Ceramic) C. 유리 (Glass) D. 스테인리스 스틸 (Stainless Steel)
	**Category: Landmark and building - Spanish/Ecuador** ¿Cómo se llama este monumento ubicado en Quito? (What is the name of this monument located in Quito?) **A. Virgen de El Panecillo (The Virgin of El Panecillo)** B. Manto de María (Manto de María) C. Mitad del mundo (Middle of the world) D. Cristo de la concordia (Christ of peace)
	**Category: Landmark and building - Irish/Ireland** Cén cathair ina bhfuil na dealbha seo? (In which city are these statues?) **A. Cathair Bhaile Átha Cliath (Dublin City)** B. Páras (Paris) C. Cathair Corcaigh (Cork City) D. Beirlín (Berlin) Cén eachtra stairiúil atá léirithe sna dealbha seo? (What historical event is depicted in these statues?) **A. An Ghorta Mhór (The Great Famine)** B. Éirí Amach 1916 (The 1916 Rising) C. Teitheadh na n-Iarlaí (The flight of the Earls) D. Cogadh 1835 (The 1835 war) Cén abhainn atá le taobh na ndealbh seo? (What river is beside these statues?) **A. An Life (The Liffey)** B. An tSionann (The Shannon) C. Abhainn an Rí (King's River) D. An Thames (The Thames)

5 Annotation Platform

We use JotForm as our annotation platform. For question entry, contributors can upload and write questions in both languages in the form. The interface can be seen in Figure 1. During validation, contributors can see all the data submitted by other contributors (Figure 2) and select the entry to see a detailed preview of the submission (Figure 3), here they can edit the data directly, provide comments, or confirm the data by starring the entry.

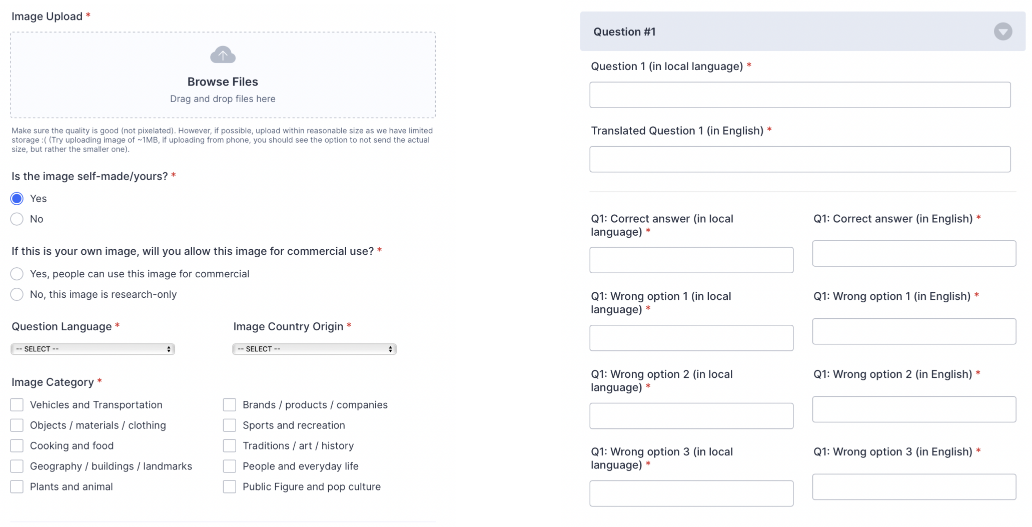


Image Upload *

Browse Files
Drag and drop files here

Make sure the quality is good (not pixelated). However, if possible, upload within reasonable size as we have limited storage :((Try uploading image of ~1MB, if uploading from phone, you should see the option to not send the actual size, but rather the smaller one).

Is the image self-made/yours? *

● Yes
○ No

If this is your own image, will you allow this image for commercial use? *

○ Yes, people can use this image for commercial
○ No, this image is research-only

Question Language *
-- SELECT --

Image Country Origin *
-- SELECT --

Image Category *

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

Question #1

Question 1 (in local language) *

Translated Question 1 (in English) *

Q1: Correct answer (in local language) *

Q1: Correct answer (in English) *

Q1: Wrong option 1 (in local language) *

Q1: Wrong option 1 (in English) *

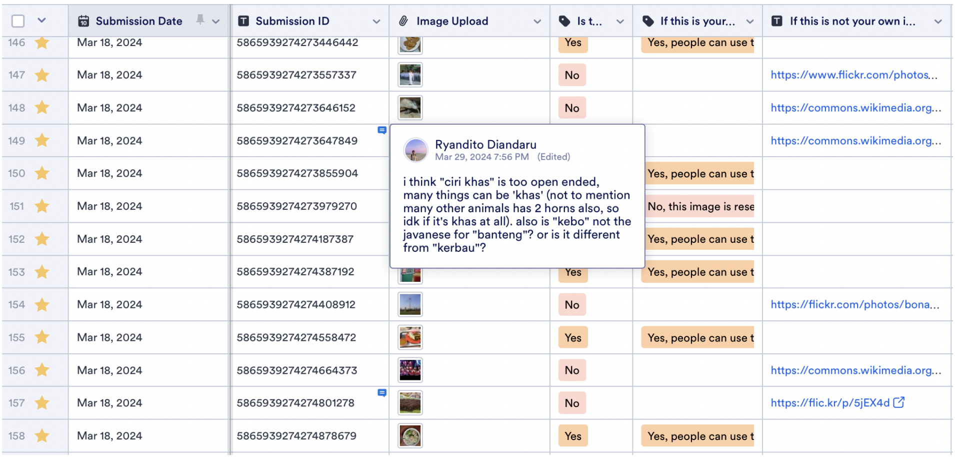
Q1: Wrong option 2 (in local language) *

Q1: Wrong option 2 (in English) *

Q1: Wrong option 3 (in local language) *

Q1: Wrong option 3 (in English) *

Figure 1: Annotation interface for entering image and questions



□	Submission Date	Submission ID	Image Upload	Is t...	If this is your...	If this is not your own i...
146 ★	Mar 18, 2024	5865939274273446442		Yes	Yes, people can use t	
147 ★	Mar 18, 2024	5865939274273557337		No		https://www.flickr.com/photos...
148 ★	Mar 18, 2024	5865939274273646152		No		https://commons.wikimedia.org...
149 ★	Mar 18, 2024	5865939274273647849				https://commons.wikimedia.org...
150 ★	Mar 18, 2024	5865939274273855904			Yes, people can use t	
151 ★	Mar 18, 2024	5865939274273979270			No, this image is rese	
152 ★	Mar 18, 2024	5865939274274187387			Yes, people can use t	
153 ★	Mar 18, 2024	5865939274274387192		Yes	Yes, people can use t	
154 ★	Mar 18, 2024	5865939274274408912		No		https://flickr.com/photos/bona...
155 ★	Mar 18, 2024	5865939274274558472		Yes	Yes, people can use t	
156 ★	Mar 18, 2024	5865939274274664373		No		https://commons.wikimedia.org...
157 ★	Mar 18, 2024	5865939274274801278		No		https://flic.kr/p/5jEX4d
158 ★	Mar 18, 2024	5865939274274878679		Yes	Yes, people can use t	

Ryandito Diandaru
Mar 29, 2024 7:56 PM (Edited)

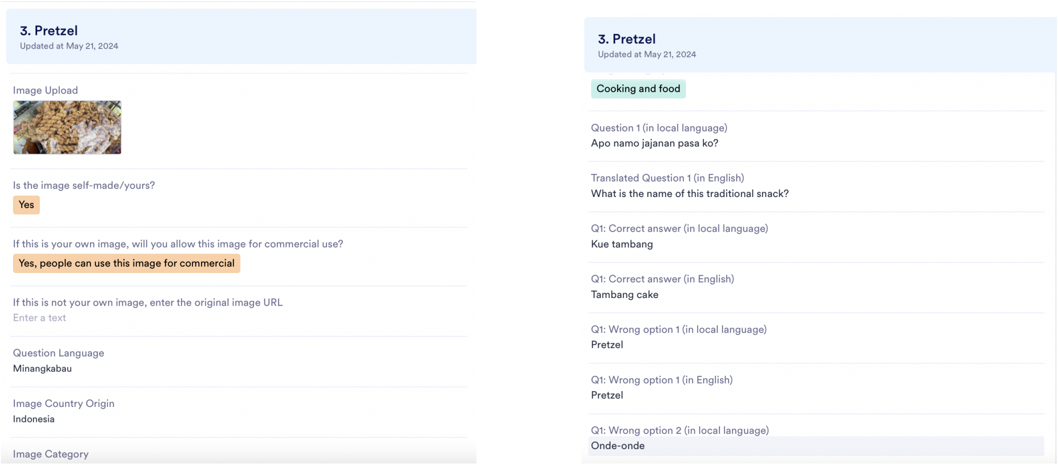
i think "ciri khas" is too open ended, many things can be 'khas' (not to mention many other animals has 2 horns also, so idk if it's khas at all). also is "kebo" not the javanese for "banteng"? or is it different from "kerbau"?

Figure 2: Annotation interface for validation. Contributors can comment, edit, and star the entries

6 Additional Statistics

6.1 Most-Frequent Words in the Questions

Figure 4 shows word clouds for the most frequent words in CVQA per category. We exclude stopwords as well as 'picture', 'photo', and 'image' from the list, since most questions contain these words. In this VQA context, we can treat them as stopwords.



3. Pretzel
Updated at May 21, 2024

Image Upload

Is the image self-made/yours?
Yes

If this is your own image, will you allow this image for commercial use?
Yes, people can use this image for commercial

If this is not your own image, enter the original image URL
Enter a text

Question Language
Minangkabau

Image Country Origin
Indonesia

Image Category

3. Pretzel
Updated at May 21, 2024

Cooking and food

Question 1 (in local language)
Apo namo jajanan pasa ko?

Translated Question 1 (in English)
What is the name of this traditional snack?

Q1: Correct answer (in local language)
Kue tambang

Q1: Correct answer (in English)
Tambang cake

Q1: Wrong option 1 (in local language)
Pretzel

Q1: Wrong option 1 (in English)
Pretzel

Q1: Wrong option 2 (in local language)
Onde-onde

Figure 3: During validation, contributors can preview the submission from other contributors

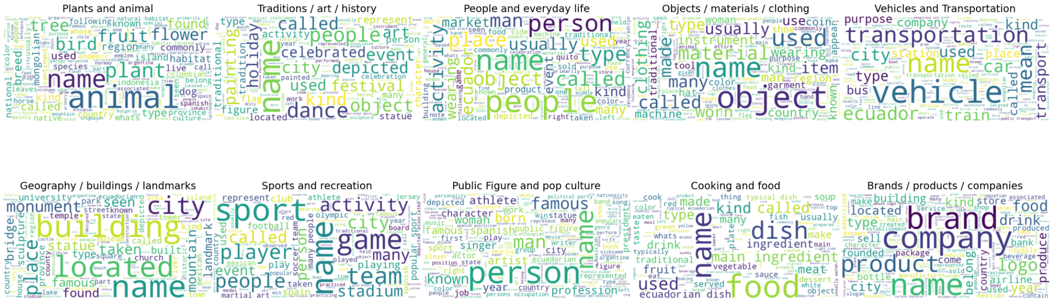


Figure 4: Word Cloud in CVQA per category

6.2 CVQA Annotator Demographic

Figure 5 illustrates the demographic statistics of the annotators, based on an anonymous questionnaire we provided. At the time of writing, we have information for 36 out of 76 annotators. As such, this breakdown is a rough representation of the annotation group.

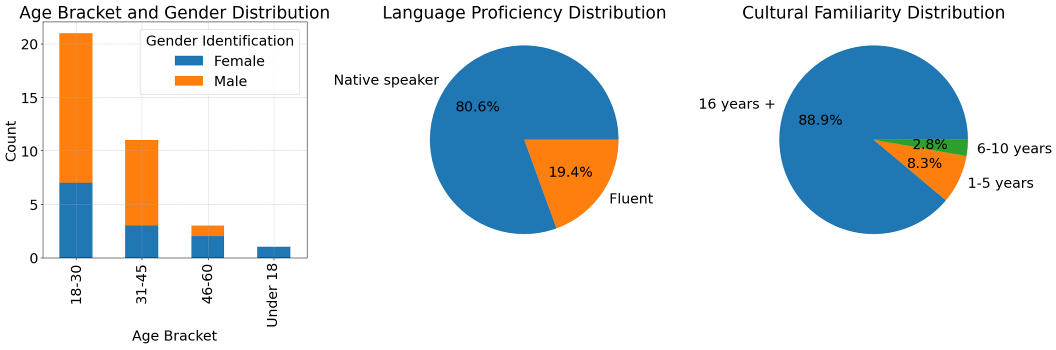


Figure 5: Annotator demographic statistics

6.3 Country-Language Pairs and Scripts

In Table 1, we provide information on the script used in each Country-Language pair.

Country	Language	Script
Africa		
Egypt	Egyptian Arabic	Arabic
Ethiopia	Amharic	Amharic
Ethiopia	Oromo	Latin
Nigeria	Igbo	Latin
Asia		
China	Chinese	Chinese
India	Bengali	Bengali
India	Tamil	Tamil
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

Table 1: The list of Country-Language pairs covered in CVQA and their corresponding scripts.

References

[1] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. I. au2, and K. Crawford. Datasheets for datasets, 2021.