Crowdsource, Crawl, or Generate? Creating SEA-VL, a Multicultural Vision-Language Dataset for Southeast Asia

Samuel Cahyawijaya[♛,1,2,3] **Holy Lovenia**[♛,2,3] **Joel Ruben Antony Moniz**[♛,4,5] **Tack Hwa Wong**[♛,2,6]
Mohammad Rifqi Farhansyah[✍,7] **Thant Thiri Maung**[✍,8] **Frederikus Hudi**[✍,9,10,2] **David Anugraha**[✍,11]
Muhammad Ravi Shulthan Habibi[✍,12,2,3] **Muhammad Reza Qorib**[✍,13] **Amit Agarwal**[✍,14]
Joseph Marvin Imperial[✍,15,16] **Hitesh Laxmichand Patel**[✍,14] **Vicky Feliren**[✍,17] **Bahrul Ilmi Nasution**[✍,18]
Manuel Antonio Rufino[✍,19] **Genta Indra Winata**[✍,20,2,3] **Rian Adam Rajagede**[✍,21] **Carlos Rafael Catalan**[✍,19]
Mohamed Fazli Imam[22] **Priyaranjan Pattnayak**[6] **Salsabila Zahirah Pranida**[22] **Kevin Pratama**[23]
Yeshil Bangera[24] **Adisai Na-Thalang**[25] **Patricia Nicole Monderin**[19] **Yueqi Song**[26] **Christian Simon**[27]
Lynnette Hui Xian Ng[26] **Richardy Lobo' Sapan**[12] **Taki Hasan Rafi**[28] **Bin Wang**[29] **Supryadi**[30]
Kanyakorn Veerakanjana[31] **Piyalitt Ittichaiwong**[31] **Matthew Theodore Roque**[19] **Karissa Vincentio**[3,32]
Takdanai Kreangphet[33] **Phakphum Artkaew**[34] **Kadek Hendrawan Palgunadi**[35] **Yanzhi Yu**[36]
Rochana Prih Hastuti[37] **William Nixon**[7] **Mithil Bangera**[24] **Adrian Xuan Wei Lim**[13]
Aye Hninn Khine[38] **Hanif Muhammad Zhafran**[7] **Teddy Ferdinan**[39] **Audra Aurora Izzani**[40]
Ayushman Singh[20] **Evan**[6] **Jauza Akbar Krito**[6] **Michael Anugraha**[6] **Fenal Ashokbhai Ilasariya**[6]
Haochen Li[6] **John Amadeo Daniswara**[6] **Filbert Aurelian Tjiaranata**[12] **Eryawan Presma Yulianrifat**[12]
Can Udomcharoenchaikit[41] **Fadil Risdian Ansori**[6] **Mahardika Krisna Ihsani**[22] **Giang Nguyen**[42]
Anab Maulana Barik[13] **Dan John Velasco**[19] **Rifo Ahmad Genadi**[22] **Saptarshi Saha**[43] **Chengwei Wei**[29]
Isaiah Flores[44] **Kenneth Ko Han Chen**[45] **Anjela Gail Santos**[46] **Wan Shen Lim**[26] **Kaung Si Phyo**[45]
Tim Santos[47] **Meisyarah Dwiastuti**[48] **Jiayun Luo**[6] **Jan Christian Blaise Cruz**[22,2] **Ming Shan Hee**[49]
Ikhlasul Akmal Hanif[12] **M.Alif Al Hakim**[12] **Muhammad Rizky Sya'ban**[7] **Kun Kerdthaisong**[50]
Lester James V. Miranda[51] **Fajri Koto**[22,2,3] **Tirana Noor Fatyanosa**[52] **Alham Fikri Aji**[22,2,3]
Jostin Jerico Rosal[53] **Jun Kevin**[54] **Robert Wijaya**[✍,49] **Onno P. Kampman**[✍,55,2]
Ruochen Zhang[✍,56,2] **Börje F. Karlsson**[✍,57] **Peerat Limkonchotiwat**[✍,58,59,2]

[1]Cohere [2]SEACrowd [3]IndoNLP [4]Mila - Quebec AI Institute [5]Polytechnique Montreal [6]Independent
[7]Bandung Institute of Technology [8]Ton Duc Thang University [9]Nara Institute of Science and Technology
[10]Works Applications [11]University of Toronto [12]University of Indonesia [13]National University of Singapore [14]Oracle
[15]University of Bath [16]National University Philippines [17]Monash University, Indonesia [18]The University of Manchester
[19]Samsung R&D Institute Philippines [20]Capital One [21]Universitas Islam Indonesia [22]MBZUAI [23]Meta
[24]University of New Haven [25]SCB 10X [26]Carnegie Mellon University [27]Sony Group Corporation [28]Hanyang University
[29]Institute for Infocomm Research, Singapore [30]Tianjin University
[31]Faculty of Medicine Siriraj Hospital, Mahidol University [32]Binus University [33]Srinakharinwirot University
[34]New York University [35]Institut Teknologi Sepuluh Nopember [36]Macau University of Science and Technology
[37]Universitas Gadjah Mada [38]King Mongkut's University of Technology Thonburi [39]Wrocław Tech
[40]University of Illiinois, Urbana-Champaign [41]Vidyasirimedhi Institute of Science and Technology [42]Auburn University
[43]Indian Statistical Institute, Kolkata [44]Ateneo de Manila University [45]Singapore Polytechnic
[46]University of the Philippines [47]Graphcore [48]Dataxet:Sonar [49]Singapore University of Technology and Design
[50]Thammasat University [51]Allen AI [52]Universitas Brawijaya [53]Seoul National University of Science and Technology
[54]Universitas Pelita Harapan [55]MOH Office for Healthcare Transformation [56]Brown University
[57]Beijing Academy of Artificial Intelligence (BAAI) [58]AI Singapore [59]Chulalongkorn University
[♛]**Main contributors** [✍]**Major contributors**

Abstract

Despite Southeast Asia's (SEA) extraordinary linguistic and cultural diversity, the region remains significantly underrepresented in vision-language (VL) research, resulting in AI models that inadequately capture SEA cultural nuances. To fill this gap, we present **SEA-VL**, an open-source initiative dedicated to developing culturally relevant high-quality datasets for SEA languages. 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. Our methodology employed three approaches: community-driven crowdsourcing with SEA contributors, automated image crawling, and synthetic image generation. We evaluated each method's effectiveness in capturing cultural relevance. We found that image crawling achieves approximately ∼85% cultural relevance while being

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 18685–18717
July 27 - August 1, 2025

more cost- and time-efficient than crowdsourcing, whereas synthetic image generation failed to accurately reflect SEA cultural nuances and contexts. Collectively, we gathered 1.28 million SEA culturally relevant images, more than 50 times larger than other existing datasets. This work bridges the representation gap in SEA, establishes a foundation for developing culturally aware AI systems for this region, and provides a replicable framework for addressing representation gaps in other underrepresented regions.

1 Introduction

The evolution of artificial intelligence (AI) and multimodal learning has ushered in a new era of sophisticated models that integrate textual and visual information, fundamentally transforming how machines perceive and understand the world. However, this technological revolution has created a profound asymmetry: while these advances promise universal applicability, they often disproportionately benefit certain languages and cultures (Yong et al., 2023; Pham et al., 2023; Cahyawijaya et al., 2023c; Tao et al., 2024; Cahyawijaya et al., 2024a; Myung et al., 2024; Li et al., 2024; Winata et al., 2024), leaving underrepresented cultures—particularly those of Southeast Asia (SEA)—largely overlooked (Wilie et al., 2020; Koto et al., 2020; Nguyen and Tuan Nguyen, 2020; Cahyawijaya et al., 2021; Aji et al., 2022b; Cruz and Cheng, 2022; Winata et al., 2023; Purwarianti et al., 2025; Cahyawijaya et al., 2024b; Urailertprasert et al., 2024). This digital divide not only creates significant challenges in developing culturally responsive AI technologies for underrepresented regions, but also fundamentally undermines AI's promise as a democratizing force.

Home to over 1,300 languages and rich cultural diversity, SEA stands as one of the world's most linguistically vibrant regions (Enfield, 2011; Aji et al., 2022a; Lovenia et al., 2024). However, the lack of SEA-relevant datasets, particularly in the vision-language (VL) domain (Lovenia et al., 2024), limits AI accessibility and risks cultural irrelevance or bias against SEA populations (Winata et al., 2024; Urailertprasert et al., 2024; Cahyawijaya, 2024; Forde et al., 2024). Addressing this disparity by creating datasets that authentically capture SEA's linguistic and cultural nuances requires large-scale collaborative efforts (Bell and Kampman, 2021). While previous crowdsourcing initiatives like NusaCrowd (Cahyawijaya et al.,

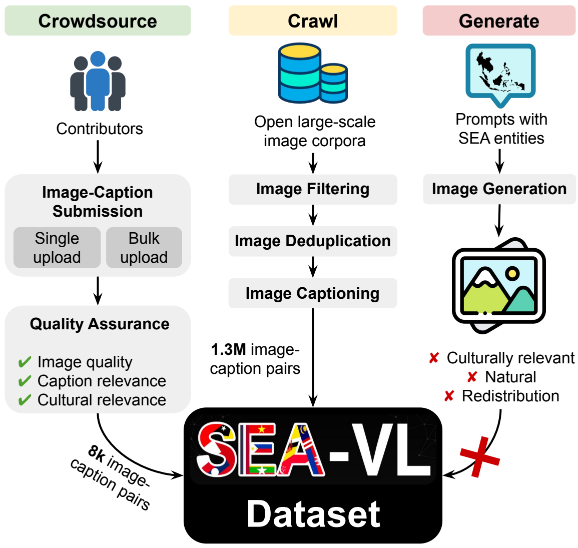


Figure 1: SEA-VL addresses the underrepresentation of SEA languages in vision-language research through a multipronged strategy for collecting culturally relevant images, via image crowdsourcing, crawling, and synthetic generation.

2023a), SEACrowd (Lovenia et al., 2024), and Aya Dataset (Singh et al., 2024), have made important strides in text-based tasks, SEA-VL takes a holistic approach to bridging the resource gap for SEA cultural representation in multimodal VL research. SEA-VL addresses this gap by developing comprehensive, high-quality VL datasets that reflect SEA's cultural heritage and linguistic diversity. SEA-VL seeks to address linguistic underrepresentation, trust, and dignity in AI, ensuring technological advancements benefit the diverse communities of SEA.

SEA-VL[1] distinguishes itself from other grassroots community-driven initiatives by going beyond relying on manual data collection only, through the participation of local contributors. With the recent popularity of various strong AI models, SEA-VL explores diverse methodologies to collect culturally relevant images in SEA. Specifically, SEA-VL collects culturally relevant image data using three different methods: (1) manual human collection, (2) a validated data filtering and deduplication pipeline on crawled images, and (3) image generation through diffusion models. 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 col-

[1]SEA-VL dataset: https://huggingface.co/collections/SEACrowd/sea-vl-multicultural-vl-dataset-for-southeast-asia-67cf223d0c341d4ba2b236e7.

lection methodologies. This evaluation not only enhances the quality and relevance of the datasets, but also provides a better understanding of the feasibility, efficiency, and quality of using AI-based data collection solutions to produce culturally relevant VL datasets, specifically for the SEA region. We also compare manual vs. automatic metadata collection to assess how well AI-based solutions generate valid, relevant metadata, which is beneficial for creating culturally relevant VL datasets.

The contributions of SEA-VL are three-fold:

- **Comprehensive, Culturally-Relevant VL Datasets:** publication of high-quality VL datasets that reflect SEA's linguistic and cultural diversity. By actively engaging local contributors, SEA-VL ensures that the data authentically represents lived experiences and regional contexts.
- **Analysis of Trade-offs in Data Collection Methods:** SEA-VL analyzes the trade-offs between effectiveness and efficiency across data collection methodologies, and demonstrates the strengths and weaknesses when employing different strategies.
- **Assessment of AI-based Solutions:** SEA-VL assesses the feasibility along with the efficiency and quality of AI-driven methods for collecting image data and its metadata, comparing these methods with manual processes for creating regionally relevant VL datasets.

2 Related Work

Crowdsource-based Data Collection Crowdsourcing is historically widely used in the machine learning community as a means to collect large amounts of high-quality human data (Crescenzi et al., 2017). Compared to alternatives such as scraping (Taesiri et al., 2024), crowdsourcing's main advantage is the ability to explicitly highlight granular variables such as demographics, opinions, and regional variations (Mostafazadeh Davani et al., 2024). The increased interest in the development of multilingual large language models (LLMs) in recent years has pushed crowdsourcing as a powerful strategy for grassroots-led data collection (Lovenia et al., 2024; Naggita et al., 2023) where representation is a key factor. As LLM research continues to grow, culture-grounded benchmarks have begun to gain traction as strong challengers for models that culturally lean towards the West (Mogrovejo et al., 2024; Winata et al., 2024; Taesiri et al., 2024). Such benchmarks are reliant on crowdsourcing to achieve the breadth and granularity needed to accurately portray multiculturality.

Underrepresented Cultures across the World Efforts for improving tools, models, and resources for low-resource languages have increased in recent years, driven by underrepresentation in widely-adopted LLMs and benchmark datasets (Cahyawijaya et al., 2023b; Koto et al., 2023, 2024; Pham et al., 2023; Song et al., 2023; Khanuja et al., 2024; Urailertprasert et al., 2024; Cahyawijaya et al., 2025; Adilazuarda et al., 2025). Beyond Southeast Asia, grassroots-led organizations have successfully spearheaded efforts to produce resources for underrepresented cultures in their region.

Masakhane, a grassroots group based in Africa, has produced work that contributes strong benchmarks (Adelani et al., 2021, 2022b, 2023), models (Dossou et al., 2022) and evaluation metrics (Wang et al., 2024a) to alleviate resource scarcity and assess the direct applicability of widely used benchmarking methods towards non-English languages. Similarly, AI4Bharat has developed an extensive body of work, including benchmarks (Verma et al., 2025), datasets (Jain et al., 2024), tools (Khan et al., 2024; Sankar et al., 2024), and models (Gala et al., 2024) representing the Indian subcontinent. Significant efforts have also been made to promote languages indigenous to the Americas, spearheaded by the AmericasNLP community. Notable projects include textual inference, such as AmericaNLI (Ebrahimi et al., 2022), as well as advancements in machine translation (Mager et al., 2023; Rangel and Kobayashi, 2024). These groups also host workshops and shared tasks (Adelani et al., 2022a) to promote research interest.

Beyond region-wide representation, recent work has also begun to pay attention to granularity *within* countries. Resources such as MC^2 (Zhang et al., 2024) and CultureAtlas (Fung et al., 2024) produce benchmarks that highlight differing cultural variations practiced within one country, further emphasizing the issue of representing a country with only one cultural norm. There is also a strong emphasis on dialectal research, with groups like ACL SIGARAB advocating for the inclusion of diverse Arabic dialects—each with distinct morphological and stylistic variations—whereas benchmarks often rely solely on *Modern Standard* Arabic to represent the entire Middle East and North Africa region

Dataset Name	#Images	%SEA Images	Cultural Coverage[†]	ID	TH	PH	SG	MY	MM	BN	KH	LA	VN	TL
SEA-VQA (Urailertprasert et al., 2024)	488	100%	Tradition & Art, Landmark	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✗
WorldCuisines (Winata et al., 2024)	6k	15.5%	Cuisine	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
CVQA (Mogrovejo et al., 2024)	7k	17.14%	Daily Life, Local Products, Pop Culture, Landmark, Tradition & Art, Transportation, Plant & Animal, Sport & Recreation, Cuisine	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗
TotalDefMeme (Prakash et al., 2023)	7.2k	100%	Tradition & Art, Pop Culture	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
OpenViVQA (Nguyen et al., 2023)	11.2k	100%	Unknown	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
Bloom Library (Leong et al., 2022)	112k	20.54%	Unknown	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CC3M (Sharma et al., 2018)	3M	0.12% *	Unknown	Unknown										
WiT (Srinivasan et al., 2021)	11.5M	0.05% *	Unknown	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗
SEA-VL (ours)	1.3M	80% *	Daily Life, Local Products, Pop Culture, Landmark, Tradition & Art, Transportation, Plant & Animal, Sport & Recreation, Cuisine	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Summary of available datasets with culturally-relevant images, showing cultural and regional coverage. A checkmark ✓ indicates coverage in the respective country (Appendix A), while a cross ✗ indicates no coverage. SEA-VL has >50× SEA images compared to other existing datasets. [†]We follow the cultural category from Mogrovejo et al. (2024). *The number is estimated based on cultural relevance in our human evaluation.

(Abdul-Mageed et al., 2024). While interest in underrepresented languages and cultures has grown in recent years, there remains a significant gap compared to the prevalence of English in models and datasets. Efforts on Southeast Asian cultures, in particular, still need improvement in areas such as multimodality—a research gap that we strive to overcome through SEA-VL and other related open community initiatives.

3 Image Collection in SEA-VL

The goal of SEA-VL is to improve the representation of SEA cultures in VL research through various image collection strategies and to provide in-depth assessments on the trade-off of each strategy. SEA-VL employs three strategies: image crowdsourcing, image crawling, and image generation. In addition, SEA-VL explores methods to gather metadata from the collected images automatically.

3.1 Image Crowdsourcing

Despite the potentially higher noise, crowdsourcing has become a common strategy employed as a means for large-scale data collection (Cahyawijaya et al., 2023a; Lovenia et al., 2024; Singh et al., 2024). Prior works have shown that improving the data quality through data pruning brings substantial benefits to the model capability (Marion et al., 2023; Chen et al., 2024; Longpre et al., 2024; Singh et al., 2024). To further improve the quality and cultural relevance of the collected images to the SEA context, we also conduct a quality assurance phase to curate and gather feedback.

Image Collection 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 countries (Appendix A) it is relevant to. In addition, they indicate their native language and are required to include a caption in both English and their native language. Submission guidelines also specify that images must be culturally relevant, and any personally identifiable information (PII), such as faces and license plates, must be redacted before submission.

Quality Assurance After data collection, we conduct a quality check, where at least two people validate each image. If the inter-annotator agreement between two validators on a certain image is below 80%, we add another annotator for that image. Contributors must pass a screening test before participating in quality assurance. Validators determine whether an image meets quality standards, assess its cultural relevance on a 5-point scale, and verify the appropriateness of its caption. Appendix C.2 presents more details about QA.

3.2 Image Crawling

Despite the rise of crowdsourced data collection, many efforts are actively managed only in their early stages, with enthusiasm waning over time, posing sustainability and scalability challenges (Cahyawijaya et al., 2023a; Lovenia et al., 2024; Gehrmann et al., 2022; Singh et al., 2024). To address this, SEA-VL explores autonomous methods to gather culturally relevant images in SEA by crawling existing sources. A carefully designed pipeline ensures high-quality collection through curated filtering and deduplication.

Image Filtering The goal of image filtering is to select SEA culturally-relevant images from a large set of images. Based on our assessment of various image filtering strategies (see Appendix E.1), we perform image filtering through semantic similarity. Given a set of unfiltered images I_{uf}, we filter out images that have an average semantic similarity score lower than a threshold ρ compared with a set of reference culturally-relevant images I_{ref}. Specifically, given an input image $x \in I_{uf}$, we define the image filtering function $f(x)$ as:

$$f(x) = \mathbb{1}_{[\frac{1}{|I_{\text{ref}}|} \sum_{z \in I_{\text{ref}}} \Psi(\lambda(x), \lambda(z))] \geq \rho}, \quad (1)$$

where $\mathbb{1}$ denotes an indicator function, λ denotes an image encoding function, and Ψ denotes a cosine similarity function between two image representations. Given Ψ, λ, and I_{ref}, we tune the value of ρ to ensure that we end up with high-quality, culturally relevant images after filtering.

Image Deduplication Collecting data by crawling various sources tends to result in highly duplicated collections (Sharma et al., 2018; Byeona et al., 2022), causing a skewed representation towards certain image concepts. Mitigating this problem, we incorporate an effective and efficient image deduplication process after filtering the images. Image deduplication can be thought of as an unsupervised image clustering problem, where a pair of images that are closely similar is considered redundant. Specifically, given two images $x, y \in I_{uf}$ and a minimum threshold ϵ, we define an image deduplication function $g(x, y)$ as:

$$g(x, y) = \mathbb{1}_{\Psi(\lambda(x), \lambda(y)) < \epsilon}, \quad (2)$$

where $\mathbb{1}$ denotes an indicator function, λ denotes an image encoding function, and Ψ denotes a similarity function between two image representations. We explore two groups of methods for image deduplication, i.e., perceptual hashing (Hadmi et al., 2012; Hamadouche et al., 2021) and semantic similarity (Wang et al., 2014; Radford et al., 2021). Perceptual hashing encodes an image into a binary hash code, while semantic similarity encodes an image into a normalized real-valued vector.

3.3 Image Generation

With the rise of various diffusion-based image generation models (Sohl-Dickstein et al., 2015) such as Stable Diffusion (Rombach et al., 2022b; Esser et al., 2024) and DALL-E (Ramesh et al., 2021; Betker et al., 2023), we further explore the possibility of generating SEA culturally relevant synthetic images. In principle, the inference of a diffusion model reverses the diffusion process by gradually transforming random noise to obtain a sample from the desired distribution. This process is repeated several steps, gradually refining the sample until it resembles the desired distribution, resulting in samples that are diverse and realistic. On the other hand, recently proposed autoregressive image generation models (Chameleon Team, 2024; Sun et al., 2024; Wu et al., 2024) also show promising image generation quality; unlike diffusion-based models, these models generate images in an autoregressive manner using discrete image tokens.

3.4 Image Captioning

In order to make the automatically collected images more meaningful, we conducted several attempts to infer image metadata, such as captions. We originally intended to explore image captioning in both English and the target language of the respective SEA culture; however, due to the poor quality of captioning in the target language (as shown in Appendix E.2), we narrow our attempt to focus on prompting for generating culturally relevant captions in English.

4 Experiment Details

Image Filtering We incorporate image semantic similarity based on a strong yet efficient pre-trained image encoder model, i.e., CLIP-ViT (86M) (Radford et al., 2021) for our image filtering pipeline.[2] To determine the optimal threshold ρ for collecting culturally relevant SEA images, we conduct human evaluations on 3 datasets: Conceptual Captions (CC3M) (Sharma et al., 2018), COYO (Byeona et al., 2022), and WiT (Srinivasan et al., 2021). We use the SEA region images of CVQA (Mogrovejo et al., 2024) and all of SEA-VQA (Urailertprasert

[2]We explore various strategies for image filtering such as heuristics filtering based on metadata and image-text similarity (see Appendix E.1).

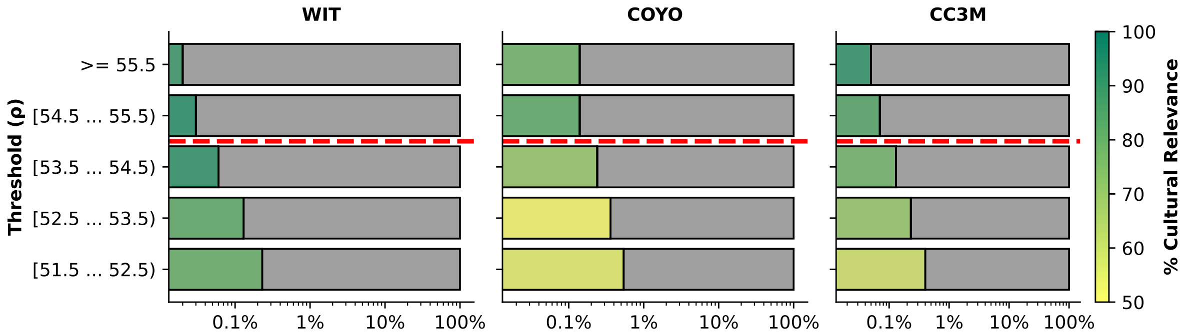


Figure 2: Human evaluation results of SEA image filtering on CC3M, COYO, and WiT datasets. Grey area indicates the proportion of images below the similarity threshold (ρ). We take the top-2 threshold groups ($\geq$54.5) as the final threshold retaining $\sim$0.15% of the total images with $\sim$85% cultural relevance.

et al., 2024) as the reference images I_{ref}. Through an exploratory data analysis, we drop all images with a similarity score (ρ) below 0.545, as only a tiny fraction of images below that threshold range are culturally relevant. This threshold is considered after the human evaluation on different similarity cluster grouping, which is presented in Figure 2 and Appendix B.2. This process removes $\sim$99% of all the images in the datasets. We cluster the remaining images into 5 groups, each with a different threshold range. We then randomly sample 50 images from each group and conduct a human evaluation to measure the cultural relevance of each group (see Appendix I.2).

Image Deduplication For image perceptual hashing, we utilize the implementation from pHash (Zauner, 2010), which encodes an image into a 64-bit binary hash code and then uses the Hamming distance as a measure of similarity. For semantic similarity, we use 3 different image embedding models, i.e., CLIP-ViT (86M) (Radford et al., 2021), SigLIP (878M) (Zhai et al., 2023), and Nomic Embed Vision v1.5 (92M) (Nussbaum et al., 2024). We perform image deduplication on the images collected from our image filtering experiment and crowdsourcing. The embedding models encode an image into a normalized embedding vector, after which cosine similarity is computed between two images. Then, we perform a human evaluation, taking 50 pairs of the top predicted samples of each method and evaluating their correctness using the criteria defined in Appendix I.3. We use 1 RTX3050 for all embedding-based models and CPU for perceptual hashing.

Image Generation We evaluate three diffusion-based image generation models: Stable Diffusion 2 (Rombach et al., 2022a), Stable Diffusion 3.5 (Esser et al., 2024) and FLUX.1-dev (Labs, 2023), and one autoregressive model: Janus-Pro (7B) (Chen et al., 2025). Images are generated for 3 cultural aspects: food, landmarks, and traditions. For food, images are generated using the prompt template "An image of people eating X" where X is the name of a Southeast Asian dish based on a list derived from the WorldCuisines dataset (Winata et al., 2024). For landmarks, the prompt is "An image of people at X", where X is a UNESCO World Heritage Site (UNESCO World Heritage Centre, n.d.) in Southeast Asia. For traditions, we use the prompt "An image of people doing X", where X is the name of a UNESCO Intangible Cultural Heritage retrieved from the metadata of SEA-VQA[3] (Urailertprasert et al., 2024).

We report the detailed hyperparameters used for each model in Appendix D. To evaluate the quality, we sample 50 generated images from each category and manually inspect them by comparing them with images from the crowdsourcing and crawling stages on two aspects: correctness and naturalness (see Appendix I.4 for details).

Image Captioning For image captioning, we explore 4 multilingual vision-language models (VLMs) within our experiments: Qwen2-VL (7B) (Bai et al., 2023; Wang et al., 2024b), Pangea (7B) (Yue et al., 2024), PaliGemma2 (10B) (Steiner et al., 2024), and Maya (8B) (Alam et al., 2024). To find the best way to collect culturally-relevant image captions, we conduct an evaluation on 2 prompting methods, i.e., location-agnostic and location-aware promptings (Mogrovejo et al.,

[3]The metadata for SEA-VQA is not publicly released and was obtained directly from the paper's authors.

Model	#Param	Precision	Throughput
Perceptual Hashing	-	2.00%	48.72
CLIP-ViT	86M	32.67%	20.34
Nomic Embed Vis.	92M	*48.67%*	21.73
SigLIP (SO)	400M	**59.33%**	3.91

Table 2: Human evaluation result of the image deduplication over 50 top predicted samples. Throughput refers to the number of images processed per second.

2024). We prompt all image captioning models to highlight these cultural items, such as local food, traditions, landmarks, or other relevant elements. The prompt should be concise, consisting of 3 to 5 sentences. The specific prompts and hyperparameters are detailed in Appendix D. We manually inspect 50 random caption generations per method. See Appendix I.5 for more evaluation details.

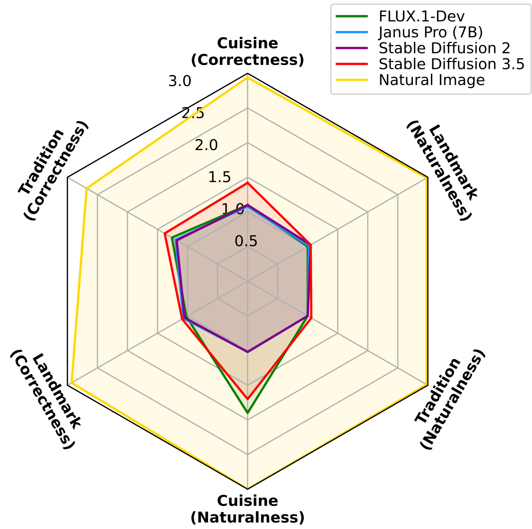


Figure 3: Human evaluation result of SEA image generation with 3-point Likert score. Natural Image refers to non-generated images taken in real life.

5 Results and Analysis

5.1 Image Filtering

The human evaluation results of image filtering with different threshold ranges are shown in Figure 2. To ensure high cultural relevance, we select the two highest threshold groups ($\geq$54.5) for our image filtering pipeline. Using this threshold, we reach $\sim$85% cultural-relevance with inter-annotator agreement (γ coefficient) of 0.6410 while retaining only $\sim$0.1% of the total images from the original dataset, e.g., from 3M images in CC3M, we gather 3,590. Using this curated threshold value, we scale the process of image filtering up to the full set of LAION (Schuhmann et al., 2021) and COYO (Byeona et al., 2022), with a total of $\sim$1.28B images[4]. From these two sources, we gather $\sim$1.72M SEA culturally-relevant images. We show the image distribution per dataset in Appendix B.2

5.2 Image Deduplication

As shown in Table 2, perceptual hashing yields a very low score compared to all semantic-similarity-based methods. This demonstrates the benefits of using pre-trained vision models and VLMs to extract semantic features from images. Among different pre-trained embedding models used, SigLIP shows the best performance in identifying duplicate images, with a 59.33% precision score, compared to CLIP-ViT and Nomic Embed Vision achieving 32.67% and 48.67%, respectively. This demonstrates that scaling models improves scene identification. Despite the higher precision, the substantially larger number of parameters of SigLIP results in much lower inference throughput. In the case of large-scale image deduplication, smaller yet performant alternative such as Nomic Embed Vision is a more suitable option as it maximizes the throughput while retaining a high deduplication precision. We then run our deduplication pipeline using Nomic Embed Vision on the $\sim$1.72M images collected from LAION and COYO resulting in $\sim$1.27M unique culturally-relevant images.

[4]We collected 2.1B image URLs, but only $\sim$60% of the images can be downloaded, on account of outdated links.

5.3 Image Generation

The results in Figure 3 demonstrate that existing image generation models struggle to produce culturally relevant SEA images. Among the evaluated models, Stable Diffusion 3.5 yields the best performance, achieving the highest correctness scores of 1.42 and 1.38 for cuisine and tradition with a moderate naturalness rating of 1.70 for cuisine. However, image generation models still fall far short of human-collected images, which achieve near-perfect correctness and naturalness scores across all categories: correctness scores remain alarmingly low, with the best model scoring <1.5 in all categories; similarly, naturalness scores are notably poor, with all models producing highly unnatural images with scores barely exceeding 1.0. This highlights a critical gap of image generation models in capturing the essence of SEA cultural elements.

Model	SEA-VQA		WorldCuisines	
	Correctness	Naturalness	Correctness	Naturalness
Human	**2.68**	**2.82**	**2.98**	**2.96**
MAYA (8B)	1.62 (+0.26)	2.34 (+0.14)	2.20 (+0.02)	<u>2.74</u> (0.00)
PALI Gemma 2 (10B)	2.04 (+0.06)	1.72 (-0.10)	2.26 (-0.16)	1.74 (+0.04)
Pangea (7B)	<u>2.36</u> (-0.24)	2.42 (-0.38)	<u>2.48</u> (-0.34)	2.52 (-0.16)
Qwen2-VL (7B)	2.10 (+0.36)	<u>2.44</u> (0.00)	2.24 (-0.14)	2.70 (-0.36)

Table 3: Human evaluation result of the image captioning phase. We use a 3-point Likert score. The values in parentheses indicate the score shift (increase or decrease) when incorporating location-aware prompting.

The primary issue we observed is the overgeneralization of cultural elements, particularly in the representation of culturally specific objects. For example, when prompting models to generate images of Pho, Khao Poon, Mee Goreng, or Mi Krop — distinct Southeast Asian noodle dishes that vary in dough ingredients, preparation, and serving presentation — the models often produced highly similar, generic-looking noodle soups, with little regard for the actual characteristics of each dish. In some cases, the results even resembled Western-style pasta, indicating a lack of grounding in the cultural specifics. Beyond object-level inaccuracies, we observed two additional recurring problems: 1) Unrealistic and different appearances in the generated images, especially when compared to actual real-world images and 2) Lack of diversity, with many generations looking nearly identical, even when the prompts referenced different countries or cultural elements. These observations align with findings from a concurrent study by Bayramli et al. (2025), which highlights the lack of cultural awareness in existing image generation models on underrepresented cultures.

5.4 Image Captioning

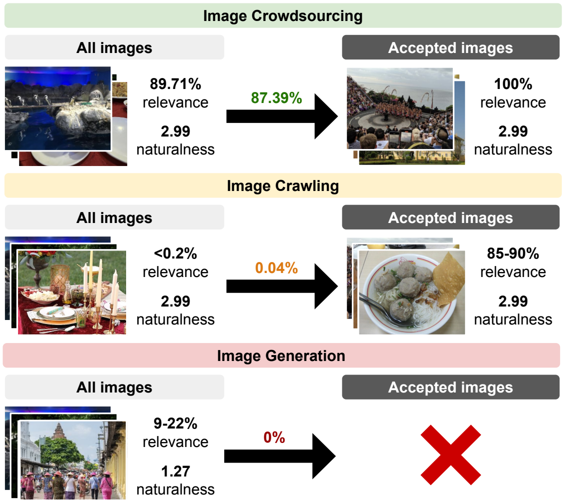
As shown in Table 3, existing VLMs can generate reasonably accurate and natural English captions for culturally relevant SEA images, though they still fall behind human-generated captions. Among the models tested, Pangea (7B) and Qwen2-VL (7B) performed best overall, with Pangea (7B) achieving the highest correctness scores in the location-agnostic setting. In comparison, Qwen2-VL (7B) excels in the location-aware setting. This suggests that existing VLMs can be a reliable option for generating synthetic captions in English. Nonetheless, there is still a huge gap for image captioning in local languages across SEA, as detailed in Appendix E.2. Our results also highlight that location-aware prompting does not consistently improve caption quality across models. For example, Qwen2-VL (7B) benefits from location-aware information in SEA-VQA but saw little improvement in WorldCuisines. Meanwhile, MAYA (8B) and PALI Gemma 2 (10B) show mixed results, with location-aware prompting slightly enhancing naturalness but not significantly improving correctness. Overall, while current models can generate high-quality English captions, there remains a gap between machine and human performance in terms of both cultural accuracy and linguistic fluency. This issue becomes more severe when the captions are in local languages, as described in Appendix E.2.

6 Discussion

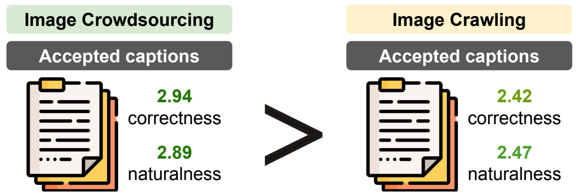
6.1 Resource Collected from SEA-VL

By virtue of all the aforementioned data collection techniques, SEA-VL, at the time of writing, is the largest gathered cultural image database for SEA, with $\sim$1.28M culturally relevant images across SEA. This is more than >50× larger than existing works, as illustrated in Table 1. Specifically, SEA-VL collects 8k manually collected images from crowdsourcing and $\sim$1.28M automatically filtered crawled images.[5] SEA-VL also brings broader outreach throughout SEA reaching underrepresented regions, e.g., Cambodia, Laos, and Timor Leste. Furthermore, SEA-VL also enables higher representation across different cultures in SEA as demonstrated by the high cultural coverage across all regions in SEA as exemplified in Table 1 and Appendix B. With this extensive cultural and

[5] We do not include the results from image generation in our published dataset due to licensing and the low cultural relevance of the generated images.



(a) Image Quality



(b) Caption Quality

Figure 4: Summary of different data collection strategies in SEA-VL. Relevance of image generation is from human evaluation on correctness. Relevance of image crowdsourcing is from annotation during quality assurance. Image naturalness is from naturalness evaluation of natural and generated images.

regional coverage of SEA-VL, we hope that AI models trained on SEA-VL can better understand and generate culturally accurate representations of SEA cultures, reducing biases and inaccuracies in representing SEA cultural contexts.

6.2 Crowdsource, Crawl, or Generate?

Figure 4 shows a clear trade-off between crowdsourcing and filtering crawled images in existing corpora. While crowdsourcing produces exceptionally high-quality images, it requires significant effort. Over 85 days, we collected 10k crowdsourced images through a labor-intensive endeavor. However, the resulting images were extremely relevant ($\geq$89%) and featured high caption quality (2.94). In contrast, filtering crawled images required only four days while still producing fairly high-quality results, with highly relevant images ($\geq$85%) and a reasonably good caption quality (2.42).

In addition, as we hint at in Section 3.2, crawling is more sustainable, with the potential to setup fully automated pipelines to continuously refresh data with minimal human intervention once initial filtering thresholds are established. In contrast to these methods, we find image generation to be completely unviable as a data collection strategy, particularly on images that require cultural nuance, as in our case. Moreover, the current image generation models come with restrictive licenses, which further limit the feasibility of using image generation as a sustainable solution for an automated culturally-relevant image collection method.

Thus, our overall recommendation might be to rely on filtering crawled images for a scalable solution, such as for the creation of large-scale training sets; crowdsourcing, on the other hand, despite being effort-intensive, is extremely useful as a means of obtaining images of high quality (e.g., creating challenging test set). At the time of writing, we would recommend avoiding the use of generated images altogether in culturally-sensitive contexts, especially within the regions with underrepresented cultures to avoid the misrepresentation of cultural identities, cultural inaccuracies, or even potentially harmful stereotypes.

7 Conclusion

We introduced SEA-VL, a corpus-building initiative covering multilingual vision data toward addressing the linguistic and cultural underrepresentation of Southeast Asian languages and cultural depictions. By leveraging diverse data collection methodologies—including crowdsourced manual collection, web crawling, and AI-generated images—along with extensive data curation procedures, SEA-VL ensures the creation of high-quality, regionally relevant vision-language datasets that authentically capture the lived experiences of SEA communities.

In summary, our findings highlight the trade-offs in data collection approaches; the potential of web-crawling for producing high-quality, culturally relevant image collections; the limited scalability and maintainability of crowdsourcing; and the limitations of existing image generation models–cultural relevance, naturalness, and licensing issues–for generating accurate, reliable, and scalable culturally relevant images. To promote open-source VL research in SEA, we release our SEA-VL dataset[6] under the CC-BY-SA 4.0 License.

[6]SEA-VL dataset: https://huggingface.co/collections/SEACrowd/sea-vl-multicultural-vl-dataset-for-southeast-asia-67cf223d0c341d4ba2b236e7.

Limitations

While SEA-VL is a step forward toward better representations of Southeast Asian culture in multilingual and multimodal AI research, we acknowledge that significant progress is still to be made. We outline several limitations of our study below.

Collection biases and limitations of outreach Similar to low-resource data collection initiatives such as SEACrowd (Lovenia et al., 2024), CVQA (Mogrovejo et al., 2024), and WorldCuisines (Winata et al., 2024), our data collection and outreach practices leveraged mainly on using social media platforms, mailing lists, and internal dissemination through the authors' networks to spread the word about the project. We acknowledge certain limitations in the nature of this practice, including the skewed or imbalanced submissions where countries that are more populous and with better infrastructure, and more connected to the original initiators of the project had higher image contributors (e.g., we saw a substantial higher number of submissions for Indonesia and Singapore compared to Myanmar, Laos, and Cambodia).

Furthermore, collection through self-taken images may only represent more popular cultures being practiced in modern times and require the contributors to be at certain places to take the photo. For future work, an ideal approach would be to have *on-the-ground* representatives for each SEA member country that can assist with data collection across culturally rich areas around SEA (e.g., traveling to rural areas beyond the cities to document non-metropolitan cultural landmarks or food). However, this type of fieldwork requires significant financial resources and manpower.

Non-holistic representation of deeper, lesser-known culture Following limitations on data collection, we acknowledge that the cultural representation of SEA is a complex cycle, making it extremely challenging to capture every cultural nuance and requires sustainable and continuous efforts from the community. Hence, our final collected image dataset might not fully represent deeper cultures in SEA at this certain timestep. To address this, we leave the submission portal open beyond the publication of this work to encourage more contributors—especially from underrepresented regions such as Cambodia, Myanmar, Brunei, and Laos—to submit more self-taken culturally relevant images. This will also serve as a good opportunity to conduct better data curation and community involvement as SEA-VL gains wider recognition across SEA.

Moreover, it is important to note that achieving a nuanced cultural curation requires multidisciplinary expertise, which our current approach only partially addresses. For example, images that had strong non-SEA influence because of historical (i.e., colonial legacies) and contemporary (i.e., globalization) reasons were curated in a relatively simplistic and ad-hoc manner. Future research should strive to integrate a systematic guideline from social science experts in order to accurately capture nuances that could have an impact on the resulting models and downstream tasks these datasets will be used for (Pouget et al., 2024).

Potentially limited generalizability using the image dataset Following certain limitations in our collected image dataset as described in the previous sections, we do not claim that any model trained or optimized using our newly collected culturally relevant SEA image dataset can effectively generalize to emerging cultural practices or underrepresented traditions. We reiterate our plan to make the submission portal open to allow the continued collection of self-taken images from the community to capture said emerging cultural practices and expand the dataset's breadth.

Ethical Considerations

We outline several practices we have conducted throughout the project to conform to ethical procedures related to data collection, privacy, and fair attribution.

Responsible credit attribution We observed justifiable and fair credit practices for our contributors for this project. We draw motivation and guidance from works documenting how low-resource language contributors emphasized lack of recognition (e.g., not being included as a co-author) in past projects related to crowdsourced data collection (Ousidhoum et al., 2024). For image contributors, we used a calibrated pointing system to encourage higher participation from SEA countries with an expected smaller number of active contributors (Lovenia et al., 2024) to reach the threshold for co-authorship. The threshold for both image contributors and validators for co-authorship qualification was 200 points. The final arrangement of authors was decided by sorting contributors with the highest garnered points in decreasing order. For more information on the contribution point system used, see Appendix H.

Safety checks for collected images We performed manual safety checks of the collected image data through consultations with the annotators to ensure that it did not contain sensitive or explicit content (e.g., images with bodily fluids like blood or costumes revealing some private human parts) which may be present in some cultural artifacts from SEA. We instructed annotators to flag and provide additional comments to images within this category for additional review. However, we found that this was not a serious issue as majority of the image submissions were centered on food, landmarks, objects, and everyday life in SEA.

Censoring personal identifiable information As part of our submission guidelines, we instructed contributors to remove and blur any personally identifiable information (PII) such as faces, car plates, and house addresses from their images before submitting to the designated form. We recommended using a free third-party PII-remover tool to do this.[7] Image validators were instructed to flag submissions with non-blurred PII to undergo re-application of the PII remover tool. For any concerns regarding PII in photos, you may contact: `seacrowd.research@gmail.com`.

Acknowledgments

We would like to thank our amazing contributors: Srishti Yadav, Raya Ramon, Anwar Choirul Mochammad, Cendekia Airlangga, Wilson Wong, Fernando Julio Cendra, Sabrina Tiun, Derry Wijaya, Randy Zakya Suchrady, Maria Bianca Therese Sta. Monica, Andy Phua, Chernenko Lada, Hendrawan Palgunadi, Dehan Al Kautsar, Elijah J. Gutierrez, Muhammad Razif Rizqullah, Le Duy Dong, Hanry Ham, Raymond Ng, Ryan Lau, Atwin Paramudya, Claire, David Samuel, Geoffrey Tyndall, Tuan Anh Vu, Asankhaya Sharma, Febriani Fitria, Pbuakhaw, and Thant Sin Tun for their hard work in submitting and validating cultural image-text pairs for SEA-VL.

This research is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. JMI is supported by the National University Philippines and the UKRI Centre for Doctoral Training in Accountable, Responsible, and Transparent AI [EP/S023437/1] of the University of Bath.

References

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The fifth nuanced Arabic dialect identification shared task. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta R. Costa-jussà, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco Guzmán, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya, Alexandre Mourachko, Safiyyah Saleem, Holger Schwenk, and Guillaume Wenzek. 2022a. Findings of the WMT'22 shared task on large-scale machine translation evaluation for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 773–800, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe

[7] https://picdefacer.com/en/

Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Adeyemi, Gilles Q. Hacheme, Idris Abdulmumim, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. 2022b. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Muhammad Farid Adilazuarda, Musa Izzanardi Wijanarko, Lucky Susanto, Khumaisa Nur'aini, Derry Wijaya, and Alham Fikri Aji. 2025. Nusaaksara: A multimodal and multilingual benchmark for preserving indonesian indigenous scripts. *Preprint*, arXiv:2502.18148.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022a. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, et al. 2022b. One country, 700+ languages: Nlp challenges for underrepresented languages and dialects in indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249.

Nahid Alam, Karthik Reddy Kanjula, Surya Guthikonda, Timothy Chung, Bala Krishna S Vegesna, Abhipsha Das, Anthony Susevski, Ryan Sze-Yin Chan, S M Iftekhar Uddin, Shayekh Bin Islam, Roshan Santhosh, Snegha A, Drishti Sharma, Chen Liu, Isha Chaturvedi, Genta Indra Winata, Ashvanth. S, Snehanshu Mukherjee, and Alham Fikri Aji. 2024. Maya: An instruction finetuned multilingual multimodal model. *Preprint*, arXiv:2412.07112.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Zahra Bayramli, Ayhan Suleymanzade, Na Min An, Huzama Ahmad, Eunsu Kim, Junyeong Park, James Thorne, and Alice Oh. 2025. Diffusion models through a global lens: Are they culturally inclusive? *arXiv preprint arXiv:2502.08914*.

Samuel J. Bell and Onno P. Kampman. 2021. Perspectives on machine learning from psychology's reproducibility crisis. *ICLR Workshop on Science and Engineering of Deep Learning*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Minwoo Byeona, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`.

Samuel Cahyawijaya. 2024. Llm for everyone: Representing the underrepresented in large language models. *Preprint*, arXiv:2409.13897.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Septiandri, James Jaya, Kaustubh Dhole, Arie Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Adilazuarda, Ryan Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Haryo Wibowo, Cuk Tho, Ichwanul Karo Karo, Tirana Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023a. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024a. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Hanung Linuwih, Bryan Wilie, Galih Muridan, Genta Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2023b. NusaWrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–945, Nusa Dua, Bali. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Putri, Wawan Cenggoro, Jhonson Lee, Salsabil Akbar, Emmanuel Dave, Nuurshadieq Nuurshadieq, Muhammad Mahendra, Rr Putri, Bryan Wilie, Genta Winata, Alham Aji, Ayu Purwarianti, and Pascale Fung. 2024b. Cendol: Open instruction-tuned generative large language models for Indonesian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14899–14914, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023c. InstructAlign: High-and-low resource language alignment via continual crosslingual instruction tuning. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 55–78, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Samuel Cahyawijaya, Ruochen Zhang, Jan Christian Blaise Cruz, Holy Lovenia, Elisa Gilbert, Hiroki Nomoto, and Alham Fikri Aji. 2025. Thank you, stingray: Multilingual large language models can not (yet) disambiguate cross-lingual word senses. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3228–3250, Albuquerque, New Mexico. Association for Computational Linguistics.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *Preprint*, arXiv:2405.09818.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024. Alpagasus: Training a better alpaca with fewer data. In *The Twelfth International Conference on Learning Representations*.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *Preprint*, arXiv:2501.17811.

Valter Crescenzi, Alvaro A. A. Fernandes, Paolo Merialdo, and Norman W. Paton. 2017. Crowdsourcing for data management. *Knowledge and Information Systems*, 53(1):1–41.

Jan Christian Blaise Cruz and Charibeth Cheng. 2022. Improving large-scale language models and resources for Filipino. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6548–6555, Marseille, France. European Language Resources Association.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John

Ortega, Ricardo Ramos, Annette Rios Gonzales, Ivan Meza-Ruiz, et al. 2022. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299.

Nicholas J Enfield. 2011. Linguistic diversity in mainland Southeast Asia. In *Dynamics of human diversity: The case of mainland Southeast Asia*, pages 63–80. Pacific Linguistics.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *Preprint*, arXiv:2403.03206.

Jessica Zosa Forde, Ruochen Zhang, Lintang Sutawika, Alham Fikri Aji, Samuel Cahyawijaya, Genta Indra Winata, Minghao Wu, Carsten Eickhoff, Stella Biderman, and Ellie Pavlick. 2024. Re-evaluating evaluation for multilingual summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19476–19493, Miami, Florida, USA. Association for Computational Linguistics.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *arXiv preprint arXiv:2402.09369*.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Aswanth Kumar M, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan. 2024. Airavata: Introducing hindi instruction-tuned llm. *Preprint*, arXiv:2401.15006.

Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina Mcmillan-major, Anna Shvets, Ashish Upadhyay, Bernd Bohnet, Bingsheng Yao, Bryan Wilie, Chandra Bhagavatula, Chaobin You, Craig Thomson, Cristina Garbacea, Dakuo Wang, Daniel Deutsch, Deyi Xiong, Di Jin, Dimitra Gkatzia, Dragomir Radev, Elizabeth Clark, Esin Durmus, Faisal Ladhak, Filip Ginter, Genta Indra Winata, Hendrik Strobelt, Hiroaki Hayashi, Jekaterina Novikova, Jenna Kanerva, Jenny Chim, Jiawei Zhou, Jordan Clive, Joshua Maynez, João Sedoc, Juraj Juraska, Kaustubh Dhole, Khyathi Raghavi Chandu, Laura Perez Beltrachini, Leonardo F . R. Ribeiro, Lewis Tunstall, Li Zhang, Mahim Pushkarna, Mathias Creutz, Michael White, Mihir Sanjay Kale, Moussa Kamal Eddine, Nico Daheim, Nishant Subramani, Ondrej Dusek, Paul Pu Liang, Pawan Sasanka Ammanamanchi, Qi Zhu, Ratish Puduppully, Reno Kriz, Rifat Shahriyar, Ronald Cardenas, Saad Mahamood, Salomey Osei, Samuel Cahyawijaya, Sanja Štajner, Sebastien Montella, Shailza Jolly, Simon Mille, Tahmid Hasan, Tianhao Shen, Tosin Adewumi, Vikas Raunak, Vipul Raheja, Vitaly Nikolaev, Vivian Tsai, Yacine Jernite, Ying Xu, Yisi Sang, Yixin Liu, and Yufang Hou. 2022. GEMv2: Multilingual NLG benchmarking in a single line of code. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 266–281, Abu Dhabi, UAE. Association for Computational Linguistics.

Azhar Hadmi, Abdellah Ait Ouahman, Brahim Ait Es Said, and William Puech. 2012. Perceptual image hashing.

Maamar Hamadouche, Khalil Zebbiche, Mohamed Guerroumi, Hanane Tebbi, and Youcef Zafoune. 2021. A comparative study of perceptual hashing algorithms: Application on fingerprint images. In *The 2nd International Conference on Computer Science's Complex Systems and their Applications*.

Sparsh Jain, Ashwin Sankar, Devilal Choudhary, Dhairya Suman, Nikhil Narasimhan, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M Khapra, and Raj Dabre. 2024. Bhasaanuvaad: A speech translation dataset for 13 indian languages. *Preprint*, arXiv:2411.04699.

Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad B, Varun G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh M. Khapra. 2024. IndicLLMSuite: A blueprint for creating pre-training and fine-tuning datasets for Indian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15831–15879, Bangkok, Thailand. Association for Computational Linguistics.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10258–10279, Miami, Florida, USA. Association for Computational Linguistics.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Black Forest Labs. 2023. Flux. https://github.com/black-forest-labs/flux.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8608–8621, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer‘s guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James Validad Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Jann Railey Montalan, Ryan Ignatius Hadiwijaya, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus Irawan, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, Reynard Adha Ryanda, Sonny Lazuardi Hermawan, Dan John Velasco, Muhammad Dehan Al Kautsar, Willy Fitra Hendria, Yasmin Moslem, Noah Flynn, Muhammad Farid Adilazuarda, Haochen Li, Johanes Lee, R. Damanhuri, Shuo Sun, Muhammad Reza Qorib, Amirbek Djanibekov, Wei Qi Leong, Quyet V. Do, Niklas Muennighoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Tai Ngee Chia, Ayu Purwarianti, Sebastian Ruder, William Chandra Tjhi, Peerat Limkonchotiwat, Alham Fikri Aji, Sedrick Keh, Genta Indra Winata, Ruochen Zhang, Fajri Koto, Zheng Xin Yong, and Samuel Cahyawijaya. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Manuel Mager, Rajat Bhatnagar, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2023. Neural machine translation for the indigenous languages of the americas: An introduction. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 109–133.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*.

Yann Mathet. 2017. The agreement measure γcat a complement to γ focused on categorization of a continuum. *Computational Linguistics*, 43(3):661–681.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

David Orlando Romero Mogrovejo, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesus-German Ortiz-Barajas, Emilio Villa Cueva, Jinheon Baek, Soyeong Jeong, Injy Hamed, Zheng Xin Yong, Zheng Wei Lim, Paula Mónica Silva, Jocelyn Dunstan, Mélanie Jouitteau, David LE MEUR, Joan Nwatu, Ganzorig Batnasan, Munkh-Erdene Otgonbold, Munkhjargal Gochoo, Guido Ivetta, Luciana Benotti, Laura Alonso Alemany, Hernán Maina, Jiahui Geng, Tiago Timponi Torrent, Frederico Belcavello, Marcelo Viridiano, Jan Christian Blaise Cruz, Dan John Velasco, Oana Ignat, Zara Burzo, Chenxi Whitehouse, Artem Abzaliev, Teresa Clifford, Gráinne Caulfield, Teresa Lynn, Christian Salamea-Palacios, Vladimir Araujo, Yova Kementchedjhieva, Mihail Minkov Mihaylov, Israel Abebe Azime, Henok Biadglign Ademtew, Bontu Fufa Balcha, Naome A Etori, David Ifeoluwa Adelani, Rada Mihalcea, Atnafu Lambebo Tonja, Maria Camila Buitrago Cabrera, Gisela Vallejo, Holy Lovenia, Ruochen Zhang, Marcos Estecha-Garitagoitia, Mario Rodríguez-Cantelar, Toqeer Ehsan, Rendi Chevi, Muhammad Farid Adilazuarda, Ryandito Diandaru, Samuel Cahyawijaya, Fajri Koto, Tatsuki Kuribayashi, Haiyue Song, Aditya Nanda Kishore Khandavally, Thanmay Jayakumar, Raj Dabre, Mohamed Fazli Mohamed Imam, Kumaranage Ravindu Yasas Nagasinghe, Alina Dragonetti, Luis Fernando D'Haro, Olivier NIYOMUGISHA, Jay Gala, Pranjal A Chitale, Fauzan Farooqui, Thamar Solorio, and Alham Fikri Aji. 2024. CVQA: Culturally-diverse multilingual visual question answering benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Aida Mostafazadeh Davani, Mark Diaz, Dylan K Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526, Miami, Florida, USA. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *arXiv preprint arXiv:2406.09948*.

Keziah Naggita, Julienne LaChance, and Alice Xiang. 2023. Flickr africa: Examining geo-diversity in large-scale, human-centric visual data. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 520–530, New York, NY, USA. Association for Computing Machinery.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Nghia Hieu Nguyen, Duong TD Vo, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2023. Openvivqa: Task, dataset, and multimodal fusion models for visual question answering in vietnamese. *Information Fusion*, 100:101868.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

Nedjma Ousidhoum, Meriem Beloucif, and Saif M Mohammad. 2024. Building Better: Avoiding Pitfalls in Developing Language Resources when Data is Scarce. *arXiv preprint arXiv:2410.12691*.

Viet H Pham, Thang M Pham, Giang Nguyen, Long Nguyen, and Dien Dinh. 2023. Semi-supervised neural machine translation with consistency regularization for low-resource languages. *arXiv preprint arXiv:2304.00557*.

Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Steiner, Xiaohua Zhai, and Ibrahim M Alabdulmohsin. 2024. No filter: Cultural and socioeconomic diversity in contrastive vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 106474–106496. Curran Associates, Inc.

Nirmalendu Prakash, Ming Shan Hee, and Roy Ka-Wei Lee. 2023. Totaldefmeme: A multi-attribute meme dataset on total defence in singapore. In *Proceedings of the 14th Conference on ACM Multimedia Systems*, pages 369–375.

Ayu Purwarianti, Dea Adhista, Agung Baptiso, Miftahul Mahfuzh, Yusrina Sabila, Aulia Adila, Samuel Cahyawijaya, and Alham Fikri Aji. 2025. NusaDialogue: Dialogue summarization and generation for underrepresented and extremely low-resource languages. In *Proceedings of the Second Workshop in South East Asian Language Processing*, pages 82–100, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Julio Rangel and Norio Kobayashi. 2024. Advancing nmt for indigenous languages: A case study on yucatec mayan and chol. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 138–142.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022a. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022b. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

Ashwin Sankar, Srija Anand, Praveen Srinivasa Varadhan, Sherry Thomas, Mehak Singal, Shridhar Kumar, Deovrat Mehendale, Aditi Krishana, Giri Raju, and Mitesh Khapra. 2024. Indicvoices-r: Unlocking a massive multilingual multi-speaker speech corpus for scaling indian tts. *NeurIPS 2024 Datasets and Benchmarks*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. *Preprint*, arXiv:1503.03585.

Yueqi Song, Simran Khanuja, Pengfei Liu, Fahim Faisal, Alissa Ostapenko, Genta Winata, Alham Fikri Aji, Samuel Cahyawijaya, Yulia Tsvetkov, Antonios Anastasopoulos, and Graham Neubig. 2023. GlobalBench: A benchmark for global progress in natural language processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14157–14171, Singapore. Association for Computational Linguistics.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Mike Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. 2024. Paligemma 2: A family of versatile vlms for transfer. *Preprint*, arXiv:2412.03555.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. 2024. Autoregressive model beats diffusion: Llama for scalable image generation. *Preprint*, arXiv:2406.06525.

Mohammad Reza Taesiri, Giang Nguyen, Sarra Habchi, Cor-Paul Bezemer, and Anh Nguyen. 2024. Imagenet-hard: The hardest images remaining from a study of the power of zoom and spatial biases in image classification. *Advances in Neural Information Processing Systems*, 36.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

UNESCO World Heritage Centre. n.d. Unesco world heritage list. `https://whc.unesco.org/en/list/`. Accessed: 2025-01-10.

Norawit Urailertprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. 2024. SEA-VQA: Southeast Asian cultural context dataset for visual question answering. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 173–185, Bangkok, Thailand. Association for Computational Linguistics.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. Milu: A multi-task indic language understanding benchmark. *Preprint*, arXiv:2411.02538.

Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1386–1393.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Hassan Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Abdi Mohamed, Hassan Ayinde, Oluwabusayo Olufunke Awoyomi, Lama Alkhaled, Sana Al-azzawi, Naome A. Etori, Millicent Ochieng, Clemencia Siro, Njoroge Kiragu, Eric Muchiri, Wangari Kimotho, Lyse Naomi Wamba Momo, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Nasir Iro, Saheed S. Abdullahi, Stephen E. Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Raphael Ogbu, Sam Ochieng', Verrah Akinyi Otiende, Chinedu Emmanuel Mbonu, Sakayo Toadoum Sari, Yao Lu, and Pontus Stenetorp. 2024a. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Ching Lam Cheng, Daud Abolade, Emmanuele Chersoni, Enrico Santus, Fariz Ikhwantri, Garry Kuwanto, Hanyang Zhao, Haryo Akbarianto Wibowo, Holy Lovenia, Jan Christian Blaise Cruz, Jan Wira Gotama Putra, Junho Myung, Lucky Susanto, Maria Angelica Riera Machin, Marina Zhukova, Michael Anugraha, Muhammad Farid Adilazuarda, Natasha Santosa, Peerat Limkonchotiwat, Raj Dabre, Rio Alexander Audino, Samuel Cahyawijaya, Shi-Xiong Zhang, Stephanie Yulia Salim, Yi Zhou, Yinxuan Gui, David Ifeoluwa Adelani, En-Shiun Annie Lee, Shogo Okada, Ayu Purwarianti, Alham Fikri Aji, Taro Watanabe, Derry Tanti Wijaya, Alice Oh, and Chong-Wah Ngo. 2024. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. *Preprint*, arXiv:2410.12705.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. 2024. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *Preprint*, arXiv:2410.13848.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Fikri Aji. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63, Singapore. Association for Computational Linguistics.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2024. Pangea: A fully open multilingual multimodal llm for 39 languages. *Preprint*, arXiv:2410.16153.

Christoph Zauner. 2010. Implementation and benchmarking of perceptual image hash functions.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

Chen Zhang, Mingxu Tao, Quzhe Huang, Jiuheng Lin, Zhibin Chen, and Yansong Feng. 2024. MC^2: Towards transparent and culturally-aware NLP for minority languages in China. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8832–8850, Bangkok, Thailand. Association for Computational Linguistics.

A Southeast Asian Countries

We present an overview of the countries in Southeast Asia (SEA), including their population data in Figure 6, which provides key demographic information for a better understanding of the region's population. We also show a visual representation of SEA region in Figure 5.

Figure 5: Map of Southeast Asia.

No.	Abbr.	Country	Flag	Population
1	BN	Brunei		0.5M
2	KH	Cambodia		17.6M
3	ID	Indonesia		280.7M
4	LA	Laos		8.0M
5	MY	Malaysia		34.6M
6	MM	Myanmar		55.8M
7	PH	Philippines		114.2M
8	SG	Singapore		6.0M
9	TH	Thailand		66.0M
10	TL	Timor-Leste		1.4M
11	VN	Vietnam		100.3M
Southeast Asia				**685.1M**
Middle East & North Africa				576.7M
North America				592.2M
Europe				774.6M

Figure 6: Southeast Asian countries and their populations as of 2025. Efforts and resources in the region still lag behind compared to other, more-represented regions even if SEA's total population is as-large or larger.

B Distribution of SEA-VL Dataset

B.1 From Image Crowdsourcing

Table 4 provides an overview of the SEA-VL dataset distribution, detailing the number of accepted images from image crowdsourcing and their cultural relevance scores. A total of 8,018 images were accepted, with an average of 2.6 validators per image. The median relevance score is 4.67, while the average score is 4.38, with a standard deviation of 0.65. Regionally, Indonesia contributes the largest number of images (3,242) with an average relevance score of 4.54. Relevance-wise, it's followed by Cambodia (208), Myanmar (586 images, 4.41), and Malaysia (453 images, 4.38). Other countries such as Thailand, Vietnam, Singapore, and the Philippines are also represented. These statistics highlight the distribution and cultural relevance of images across Southeast Asia.

Criteria for Data Quality Flags in SEA-VL Dataset To ensure the quality and cultural relevance of the images in the SEA-VL dataset, we define three key evaluation metrics (Figure 7).

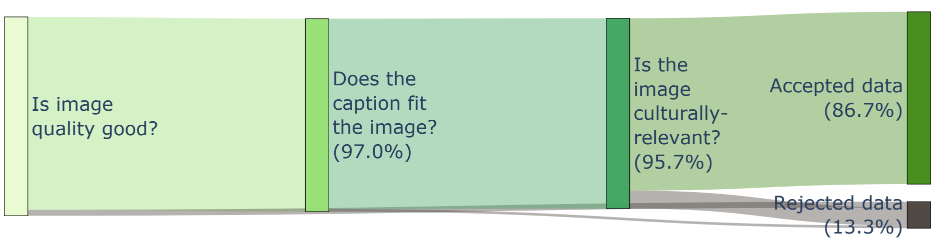


Figure 7: Curating the images obtained from SEA-VL crowdsourcing.

Overall			
# Data	8018	**Relevance**	
# Validator per data	2.6	Median	4.67
		Avg.	4.38
		Std.	0.65

Per region An image can be relevant for more than one region.		
Country	**# Data**	**Avg. relevance**
Brunei	72	4.26
Cambodia	208	4.54
Timor-Leste	12	4.22
Indonesia	3242	4.54
Laos	157	4.32
Malaysia	453	4.38
Myanmar	586	4.41
Phillippines	543	4.21
Singapore	1542	4.07
Thailand	1006	4.41
Vietnam	541	4.32

Table 4: Statistics of accepted data from image crowdsourcing. Relevance refers to image cultural relevance (using 5-point Likert score).

- **Is the Image Quality Good?**
  - **True**: If the average photo quality score is greater than 0.5 and at least **two annotators** have reviewed the image.
  - **None**: If the average photo quality score is greater than 0.5, but fewer than two annotators provided a review.
  - **False**: If the average photo quality score is **0.5 or lower**.
- **Does the Caption Fit the Image?**
  - **True**: If the average caption fit score is greater than 0.5 and at least **two annotators** have reviewed the image.
  - **None**: If the average caption fit score is greater than 0.5, but fewer than two annotators provided a review.
  - **False**: If the average caption fit score is **0.5 or lower**.
- **Is the Image Culturally Relevant?**
  - **True**: If the average cultural relevance score (found in SEA) is **3 or higher**, and at least **two annotators** have reviewed the image.
  - **None**: If the average cultural relevance score is **3 or higher**, but fewer than two annotators provided a review.
  - **False**: If the average cultural relevance score is **below 3**.
- **Overall Data Quality**
  - **True**: If the image meets **all three** conditions:
    * `average_photo_quality` > 0.5
    * `average_found_in_SEA_score` ≥ 3
    * `average_caption_fit` > 0.5
    * **At least two annotators** provided reviews

– **None**: If all three conditions are met, but fewer than two annotators reviewed the image.
– **False**: If any of the three criteria do not meet the required threshold.

These flags help ensure that the dataset maintains high-quality images, culturally relevant content, and appropriate captions, allowing for more robust research applications.

B.2 Detailed Statistics of Image Filtering

For the accepted data from image filtering, we begin by filtering three existing datasets. The detailed statistics for all three datasets—Conceptual Captions 3M (Sharma et al., 2018), COYO (2M) (Byeona et al., 2022), and WiT (Srinivasan et al., 2021)[8]—are presented in Table 5, which outlines their alignment with the threshold ρ in our image filtering experiment.

To scale up the experiment, we use two large-scale datasets, i.e., COYO (700M) (Byeona et al., 2022) and LAION (Schuhmann et al., 2021). From 747M image URLS in COYO, we successfully crawled 467.5M images ($\sim$62.5%), while for LAION we collected 1.3B URLs and gathered 826.5M images ($\sim$66.74%). We show the histogram for each threshold range for COYO and LAION in Figure 8.

The accepted data from image filtering is obtained from the Platinum, i.e., $[54.5...55.5)$, and Diamond, i.e., ≥ 55.5 threshold groups in Figure 8. We then run image deduplication on the combined filtered COYO and LAION data, and end up with a total of 1.28M images.

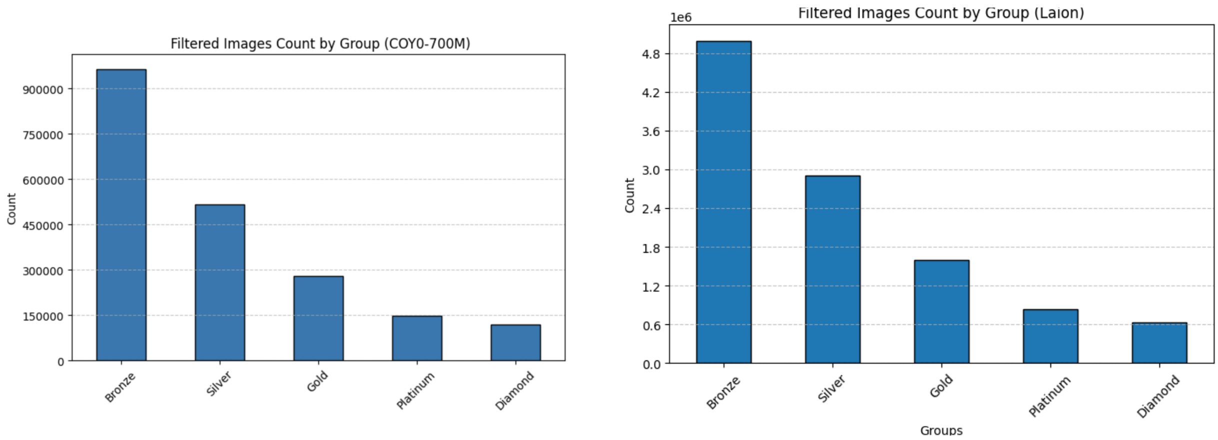


Figure 8: Histogram of images per filtering threshold in **(left)** COYO and **(right)** LAION. The X-axis labels denotes the different threshold groups with Bronze=$[51.5 \ldots 52.5)$, Silver=$[52.5 \ldots 53.5)$, Gold=$[53.5 \ldots 54.5)$, Platinum=$[54.5 \ldots 55.5)$, and Diamond=≥ 55.5

C Additional Detail on Image Crowdsourcing

The SEA-VL project page can be accessed at `https://seacrowd.github.io/seavl-launch/`.

C.1 Image Submission

The image submission form used during the image collection phase is provided in Figure 9, where contributors input necessary information, including captions and cultural relevance details for each image. Additionally, Figure 10 showcases the bulk upload UI tool, designed to streamline the submission process, allowing contributors to upload multiple images at once while inputting essential metadata. The contribution progress of image submissions is shown in Figure 12.

C.2 Quality Assurance

Contributors validate each submitted image using the form in Figure 11. Validators assess image quality, ensuring clarity, no offensive content, and that the image is not overly cropped or AI-generated.[9] Images should also be culturally relevant to Southeast Asia, either by being unique to the region (e.g., local food,

[8]We use the SEA languages subset of WiT from SEACrowd (Lovenia et al., 2024)
[9]Contributors have to pass a short screening test before becoming validators.

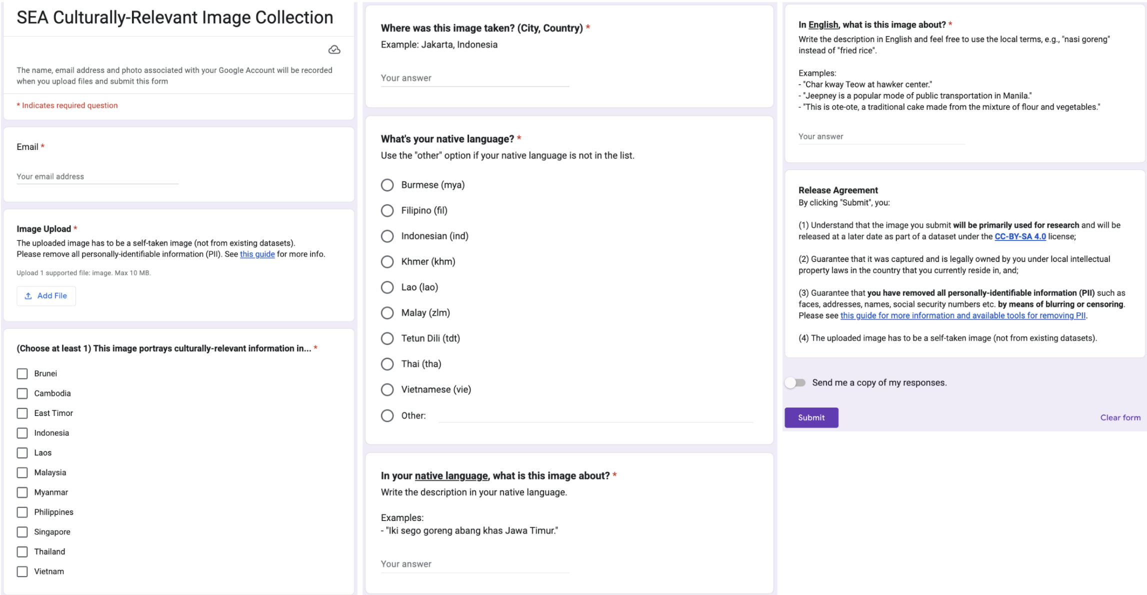


Figure 9: SEA-VL image submission form for single upload.

landmarks) or strongly reflective of SEA culture (e.g., SEA-specific celebrations). Additionally, validators ensure images do not contain personally identifiable information (PII).

The annotation guidelines (Figure 14) include 5 options for cultural relevance: Option 1 is for images uniquely associated with SEA, such as local foods or landmarks. Option 2 covers images that strongly reflect SEA culture or lifestyle and have a low degree of similarity to other cultures. Option 3 includes images that may not originally be from SEA but are very common in SEA culture. Option 4 pertains to images with some SEA affiliation but stronger ties to other cultures. Option 5 is for images unrelated to SEA. Regarding the caption, validators must also determine whether it aligns with the image, selecting from three options: yes, no, or unsure (Figure 11). The contribution progress for image validation is tracked and shown in Figure 13.

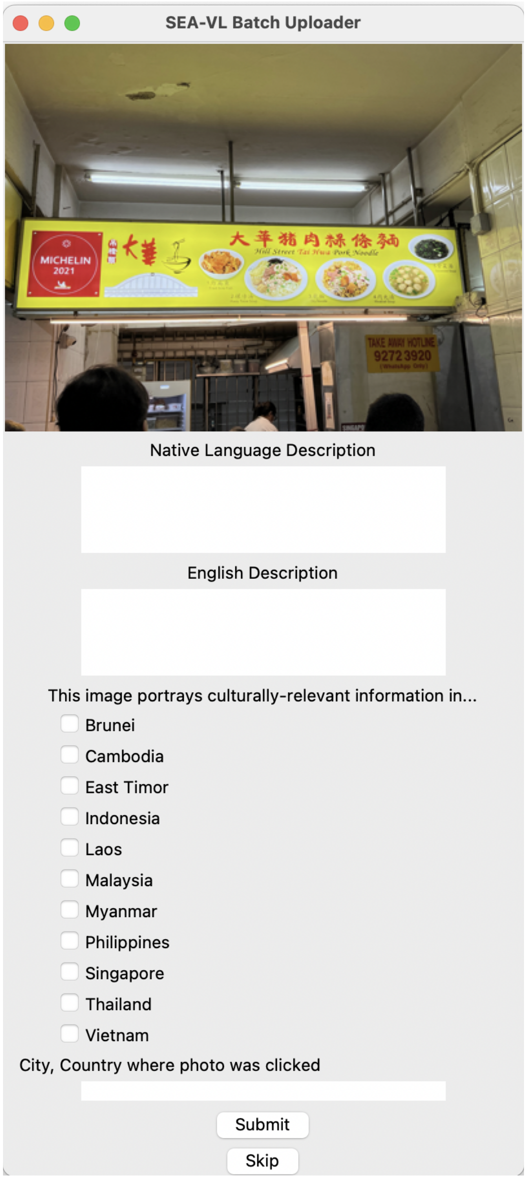


Figure 10: SEA-VL image submission UI tool for bulk upload.

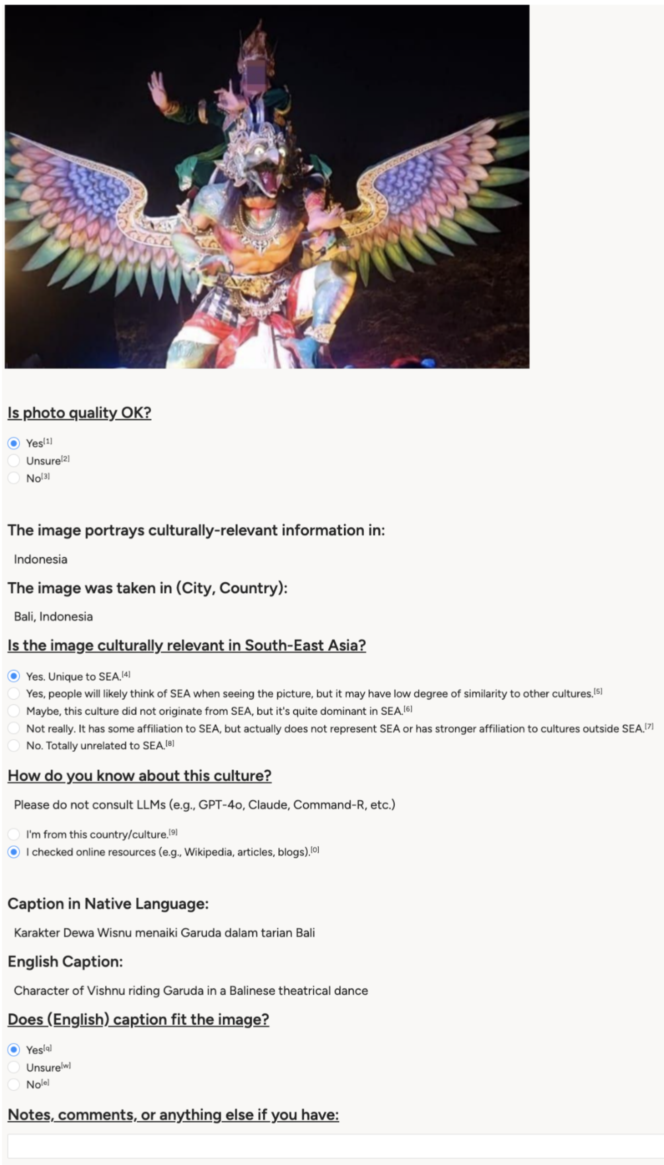


Figure 11: The SEA-VL image validation form used in the quality assurance phase.

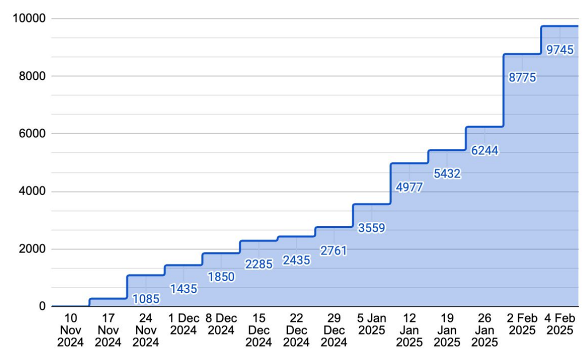


Figure 12: SEA-VL image submission contribution progress.

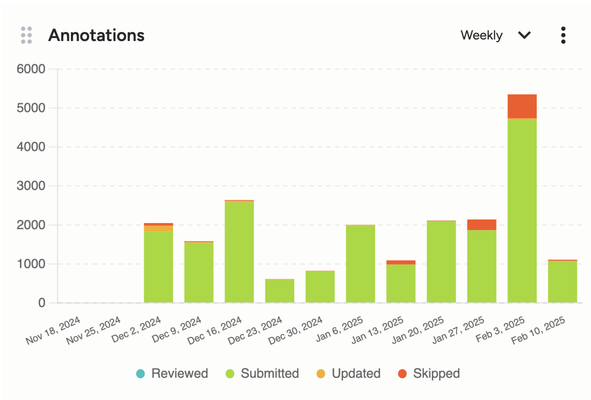


Figure 13: SEA-VL image validation contribution progress.

SEA-VL Annotation Guideline

This document explains the annotation guidelines for SEA-VL data validation. Please note that to become a SEA-VL annotator, you must first pass this short screening test. Once you pass, we will contact you and provide an official validator account to validate our data.

Validation Mechanism

You will be given an image and its short caption provided by the contributors. We want to ensure that the data is suitable for SEA-VL, that is, relevant to South East Asia and in an acceptable quality. You don't have to be native in the corresponding SEA countries to validate the data, however, please utilize Google, Wikipedia, or any other trustworthy sources when validating the data that you are not familiar with.

Validation Questions

Please consult the following guideline for answering the validation questions.

1. Is the photo quality OK and appropriate?

- Answer OK if the image quality is acceptable. Note that we don't require high-quality, professional photography. Any smartphone or amateur photography is acceptable. Importantly, ensure that the captured object is clear, not overly cropped, and not blurry. Ensure that the photo is also appropriate, e.g., it does not contain harmful stereotypes (eg. High-school gang fight in Indonesia, Tawuran), offensive or suggestive images, or any other illegal content.
- Rotated images are also considered OK, as humans can understand images that are rotated just fine. We want to ensure that our system is robust.
- Note that we expect the images to be originated from the contributors. So if you suspect that the image is taken from the internet or is AI-generated, also select NO.

2. Is the image culturally relevant in South-East Asia?

- **Option 1: Yes. Unique to SEA.**
  For images or cultures that originate from SEA. Examples include:
  - Local food such as Pad Thai.
  - Local, unique and very popular buildings like Petronas Towers or Monumen Nasional.
  - Local clothes such as Batik.
  - Local activities such as the Ati-Atihan Festival.
  - Culturally/historically significant paintings.
  - Local brands that are very well-known locally and even internationally (e.g., IKEA is well-known to be Swedish, we want our model to be familiar with our brands too). Eg. Indomie.
- **Option 3: Maybe. Not originally from SEA but very common in SEA culture.**
  For images/concepts not originally from SEA but very ubiquitous in SEA culture and everyday life and/or have SEA-specific local nuances. Examples include:
  - Local buildings/places that are not necessarily unique, but have strong SEA-vibe and local influence, e.g., in their architecture. For example, random mosques with local architectures, or random local markets/housing complexes with distinguishable SEA elements. Note that generic homes/hotels that are very typical globally (commonly seen elsewhere outside SEA) should not be considered.
  - Non-SEA celebrations with SEA-specific nuances: For instance, Chinese New Year celebrations in Singapore or Eid al-Fitr celebrations in Malaysia.
  - Typical concepts or mundane objects with SEA-specific twists: An example could be candy with Rambutan flavor, or objects such as signs, written with SEA languages.
  - Foods that are not originated from the region, but very prominent in SEA, e.g. Chinese dishes that are popular in Singapore.
  - Non-local brands that are very strong/prominent in SEA (but not generally worldwide), e.g., MSG like Aji no Moto, or Mixue.
- **Option 5: No. Totally unrelated to SEA.**
  Images/concepts that have nothing to do with SEA. Examples include:
  - Events like the Super Bowl.
  - International landmarks (e.g., Statue of Liberty).
  - Generic objects with no cultural relevance (e.g., random chair).
  - International brand that has no significance in SEA specifically, eg a photo of a random HSBC office.

Select Option 2 ("Yes, people will likely think of SEA when seeing the picture, but it may have a low degree of similarity to other cultures.") or Option 4 ("Not really. It has some affiliation to SEA, but actually does not represent SEA or has stronger affiliation to cultures outside SEA.") if you are unsure between corresponding categories.

3. Does the image contain a person's face, phone number, ID, or car plate numbers?

Ensure that the image properly obfuscates personally identifiable information (PII) such as faces, IDs, car plates, etc. Note that the face of a public figure is not considered PII.

Figure 14: SEA-VL annotation guideline for data validation.

D Hyperparameters

Our SEA-VL experiment repository can be accessed on GitHub.[10]

D.1 Image Deduplication

We divided the images into 50 randomly sampled subsets for comparison. Similarity was then assessed using four methods: (1) perceptual hashing via the imagehash[11] library (Hamming distance, distance $= 16$, CPU computation); (2) Nomic Embed Vision v1.5 (feature embeddings, threshold $= 0.95$, GPU computation); (3) SigLIP (feature embeddings, threshold $= 0.85$, GPU computation); and (4) CLIP-ViT (feature embeddings, threshold $= 0.95$, GPU computation). The best threshold value for each method is determined based on the ratio between precision and number of duplicated images captured (as a proxy to recall). These methods differ in their hardware requirements and similarity measurement strategies.

D.2 Image Generation

For diffusion models, high-quality images were generated using the hyperparameters specified in the guidance-distilled[12] settings of Flux.1-Dev. The inference process was configured with 50 sampling steps and CFG at a scale of 3.5. The scheduler settings remained at their default values, with the Flow Matching Euler Discrete[13] scheduler employed for both Flux.1-Dev and Stable Diffusion 3.5, while the DDIM[14] scheduler was utilized for Stable Diffusion 2. The generated images have a resolution of 1024×1024 pixels. For autoregressive models, we followed the default hyperparameters specified in the official implementation[15]. This configuration included a CFG scale of 5.0, a temperature value of 1.0, and the generation of 576 visual tokens per image. Due to architectural limitations in Janus-Pro, the output images were constrained to a resolution of 384×384 pixels.

D.3 Image Captioning

We utilize two types of prompting for image captioning: Location-Agnostic Prompt and Location-Aware Prompt, as detailed in Figure 15 and 16. The prompt is designed for images that may include culturally significant elements from Southeast Asia. The caption should highlight these cultural items, such as local food, traditions, landmarks, or other relevant elements, and should be concise, consisting of 3 to 5 sentences. In terms of Location-Aware Prompts, the specific country's location is retrieved from the dataset metadata if the image is associated with a specific location within Southeast Asia, and this information is incorporated into the prompt as a hint. In this case, the caption must mention the cultural elements from the specified location, providing more context about the region's culture and traditions.

For both prompts, we generate captions in six different languages: English, Thai, Malay, Tagalog, Indonesian, and Vietnamese. This multilingual approach allows for a diverse representation of Southeast Asian culture in different linguistic contexts. Each language provides its unique cultural perspective on the image, ensuring that the captions are both culturally and linguistically appropriate. The captions were generated deterministically using greedy decoding, prioritizing coherence and reproducibility, while disabling the repetition penalty to maintain a fair comparison across languages.

E Additional Detail on Image Crawling

E.1 Other Methods Explored for Image Filtering

We explore various approaches for image filtering, such as rule-based approaches through URL whitelisting/blacklisting, EXIF geolocation filtering, and other heuristics. Nonetheless, these methods are not

[10] https://github.com/SEACrowd/sea-vl-experiments
[11] https://github.com/JohannesBuchner/imagehash
[12] https://huggingface.co/docs/diffusers/main/api/pipelines/flux
[13] https://huggingface.co/docs/diffusers/api/schedulers/flow_match_euler_discrete
[14] https://huggingface.co/docs/diffusers/api/schedulers/ddim
[15] https://github.com/deepseek-ai/Janus

Location-Agnostic Prompt

English: Write a caption in English for an image that may include culturally significant objects or elements from Southeast Asia. The caption should specifically name Southeast Asian cultural items, such as cuisine, traditions, landmarks, or other related elements if they appear in the image. The caption should be concise, consisting of 3 to 5 sentences.

Thai: เขียนแคปชั่นเป็นภาษาไทยสำหรับรูปที่อาจมีวัตถุทางวัฒนธรรมหรือองค์ประกอบจาก Southeast Asia แคปชั่นควรจะมีชื่อวัตถุวัฒนธรรมของ Southeast Asian ได้แก่ ชื่ออาหาร วัฒนธรรม สถานที่ หรืออะไรก็ตามที่ปรากฏบนรูปภาพ แคปชั่นควรมีความยาว 3 ถึง 5 ประโยค

Malay: Tulis kapsyen dalam bahasa Melayu untuk imej yang mungkin mengandungi objek atau unsur penting budaya dari Asia Tenggara. Kapsyen harus menyebut secara khusus item budaya Asia Tenggara, seperti makanan, tradisi, mercu tanda, atau elemen lain yang berkaitan jika ia muncul dalam imej. Kapsyen hendaklah ringkas, terdiri daripada 3 hingga 5 ayat.

Tagalog: Magbigay ng caption sa Tagalog para sa isang litrato na maaaring may elemento o aspetong makahulugan para sa Timog-Silangang Asya. Dapat pinapangalanan ng caption na ito ang mga cultural at tradisyunal na bagay sa Timog-Silangang Asya tulad ng pagkain, tradisyon, lugar, o anumang kaugnay na elemento kung ito ay kasama sa litrato. Ang caption ay dapat na maigsi, mga 3 hanggang 5 pangugusap lamang.

Indonesian: Tulis deskripsi dalam bahasa Indonesia untuk gambar yang mungkin mengandung objek atau elemen budaya penting di Asia Tenggara. Deskripsi harus menyebutkan secara spesifik barang-barang dalam budaya Asia Tenggara, seperti makanan, tradisi, tempat bersejarah, atau elemen terkit lainnya jika mereka muncul dalam gambar. Deskripsi harus singkat, terdiri dari 3 sampai 5 kalimat.

Vietnamese: Viết chú thích bằng tiếng Anh cho một hình ảnh mà nó có thể chứa các vật thể hoặc yếu tố văn hóa quan trọng của Đông Nam Á. Chú thích phải nêu tên cụ thể của các yếu tố văn hóa Đông Nam Á - chẳng hạn như ẩm thực, phong tục, địa danh hoặc các yếu tố liên quan khác - nếu chúng xuất hiện trong hình ảnh. Chú thích phải ngắn gọn, chỉ bao gồm từ 3 đến 5 câu.

Figure 15: Location-Agnostic Prompts in English and multiple SEA languages.

Location-Aware Prompt

English: This is an image from {Location}. Write a caption in English for an image that may include culturally significant objects or elements from {Location}. The caption should specifically name Southeast Asian cultural items, such as cuisine, traditions, landmarks, or other related elements if they appear in the image. The caption should be concise, consisting of 3 to 5 sentences.

Thai: นี้คือรูปจาก Thailand เขียนแคปชั่นในภาษาไทยสำหรับรูปที่อาจมีวัตถุทางวัฒนธรรมจาก Thailand แคปชั่นควรกล่าวถึงวัตถุวัฒนธรรมทาง Southeast Asian ได้แก่ ชื่ออาหาร วัฒนธรรม สถานที่ หรืออื่นๆที่อาจปรากฏบนรูป แคปชั่วควรมีความยาว 3 ถึง 5 ประโยค

Malay: Ini adalah imej dari Malaysia. Tulis kapsyen dalam bahasa Melayu untuk imej yang mungkin mengandungi objek atau unsur penting budaya dari Malaysia. Kapsyen harus menyebut secara khusus item budaya Asia Tenggara, seperti makanan, tradisi, mercu tanda, atau elemen lain yang berkaitan jika ia muncul dalam imej. Kapsyen hendaklah ringkas, terdiri daripada 3 hingga 5 ayat.

Tagalog: Ito ay isang litrato mula sa Philippines. Magsulat ng caption sa Tagalog para sa isang litrato na may elemento o bagay na makabuluhan sa kultura ng Philippines. Dapat pinapangalanan ng caption na ito ang mga cultural at tradisyunal na bagay sa Timog-Silangang Asya tulad ng pagkain, tradisyon, lugar, o anumang kaugnay na elemento kung ito ay kasama sa litrato. Ang caption ay dapat na maigsi, mga 3 hanggang 5 pangugusap lamang.

Indonesian: Ini adalah gambar dari Indonesia. Tulis deskripsi dalam bahasa Indonesia untuk gambar yang mungkin mengandung objek atau elemen budaya penting dari Indonesia. Deskripsi harus menyebutkan secara spesifik barang-barang dalam budaya Asia Tenggara, seperti makanan, tradisi, tempat bersejarah, atau elemen terkit lainnya jika mereka muncul dalam gambar. Deskripsi harus singkat, terdiri dari 3 sampai 5 kalimat.

Vietnamese: Đây là hình ảnh từ Vietnam. Hãy viết chú thích bằng tiếng Anh cho một hình ảnh mà nó có thể chứa các vật thể hoặc yếu tố văn hóa quan trọng của Vietnam. Chú thích phải nêu tên cụ thể của các yếu tố văn hóa Đông Nam Á - chẳng hạn như ẩm thực, phong tục, địa danh hoặc các yếu tố liên quan khác - nếu chúng xuất hiện trong hình ảnh. Chú thích phải ngắn gọn, chỉ bao gồm từ 3 đến 5 câu.

Figure 16: Location-Aware Prompts in English and multiple SEA languages.

Threshold	CC3M (3M Images)			COYO (1.66M Images)			WiT (1.46M Images)		
	Relevance	#Images	%Images	Relevance	#Images	%Images	Relevance	#Images	%Images
< 51.5	-	2.99M	99.12%	-	1.65M	98.58%	-	1.44M	98.25%
$[51.5 \dots 52.5)$	58%	11885	0.40%	54%	8925	0.54%	80%	9627	0.66%
$[52.5 \dots 53.5)$	70%	6824	0.23%	40%	5919	0.36%	82%	6715	0.46%
$[53.5 \dots 54.5)$	78%	3841	0.13%	70%	3996	0.24%	92%	4377	0.30%
$[54.5 \dots 55.5)$	84%	2091	0.07%	82%	2323	0.14%	94%	2590	0.18%
≥ 55.5	92%	1499	0.05%	78%	2294	0.14%	90%	2162	0.15%

Table 5: The detailed filtering statistics of the image filtering on CC3M, COYO, and WiT datasets.

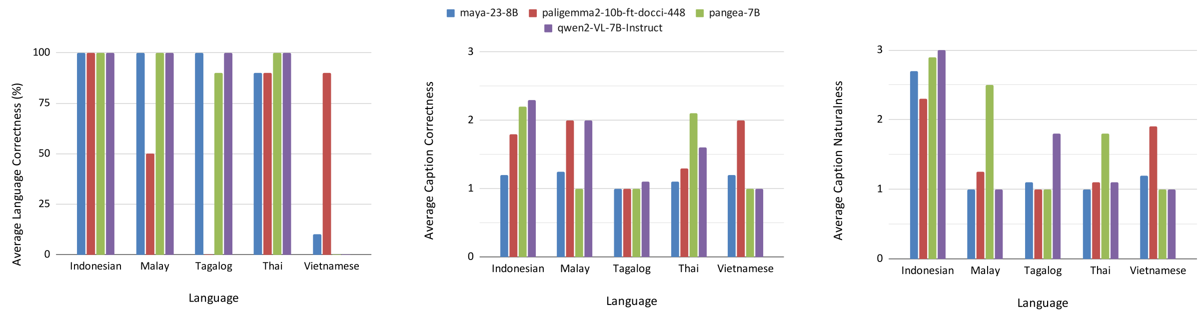


Figure 17: Bar charts that show performance of various multilingual VLM models on captioning images with a cultural context in the corresponding language, as measured by 3 metrics: **(left)** the average correctness of the language of the generated caption, **(center)** the average correctness of the caption, and **(right)** the average naturalness of the caption.

very noisy and tend to be source-dependent rendering them unreliable and not scalable. Beyond rule-based approaches, we explore two semantic similarity based approaches, i.e., text-image similarity and image-image similarity. In text-image similarity, we first collect terms that are culturally relevant to SEA, e.g., name of local dishes, name of places, etc. While for image-image similarity, we first collect culturally relevant images from existing source datasets, i.e., CVQA (Mogrovejo et al., 2024) and SEA-VQA (Urailertprasert et al., 2024).

E.2 Image Captioning in Local Languages

As discussed in Section 3.4, we initially intended to have image captioning in both English as well as the language corresponding to the respective SEA culture's target language. However, we decided to focus purely on English captions based on insight obtained from an initial pilot study. In this sub-section, we outline the pilot study and describe the findings that justified this choice.

For each among 5 languages (and their corresponding cultures), we randomly select 10 images and have each image captioned by each of our four chosen multilingual VLMs (Maya (8B) (Alam et al., 2024), PaliGemma2 (10B) (Steiner et al., 2024), Pangea (7B) (Yue et al., 2024), and Qwen2-VL (7B) (Bai et al., 2023; Wang et al., 2024b)). These models are prompted with a language-aware prompt, and are instructed to generate captions for the corresponding images in the target language. We thus obtain 200 image-caption pairs. We then manually evaluate the so generated captions on 3 parameters: the correctness of the language the caption was generated in, the correctness of the caption, and the naturalness of the caption. The language correctness is posed as a simple binary yes-no question. The correctness and naturalness of the caption are both measured using a 3-point scale, which we describe in Appendix Section I.5.

We present our findings in Figure 17. Overall, we find that most models struggle to output captions that are correct and natural in the SEA language corresponding to the context of the image shown (the one exception here, perhaps, is Indonesian, where, to our pleasant surprise, we see both Pangea (7B) and Qwen2-VL (7B) being fairly correct and remarkably natural). Interesting, we find that these models are often unable to respect even the requested language, particularly in the case of Vietnamese; we often find the models defaulting to English captions in these cases.

F Contributor Details

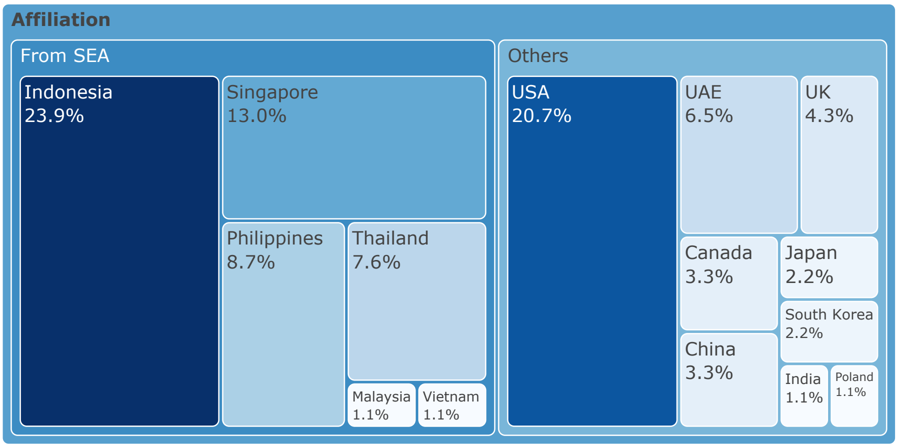
The details of our authors and their contribution points are provided here. The full contribution point tracking monitor can be accessed here.

G Contributor Demographic

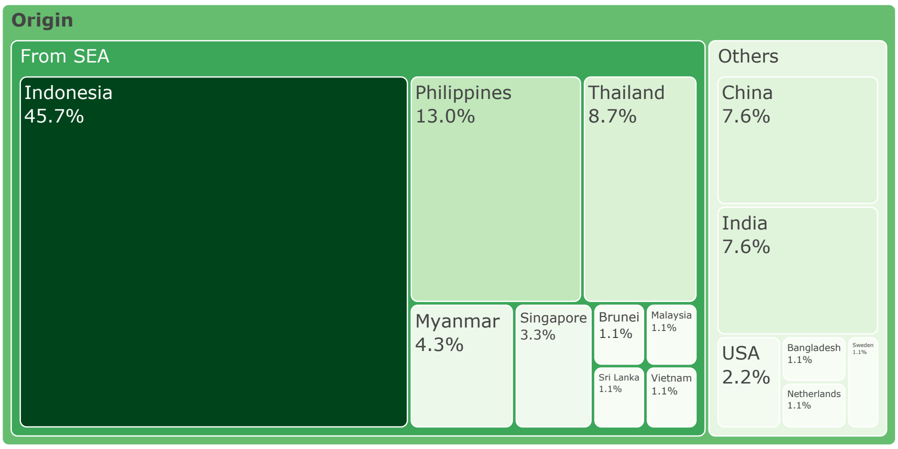
We describe our author demographic and image validator demographic in Figure 18 and Figure 19, respectively. Figure 18 shows the demographic distribution of SEA-VL authors based on their affiliation and origin countries.

Affiliation (a) The largest group of authors are affiliated with Indonesia, followed by the USA, Singapore, and the Philippines. Other countries with notable author affiliations include Thailand, the UAE, the UK, and Canada.

Origin (b) When considering the origin countries, Indonesia again has the largest representation (45.7%). The Philippines accounts for 13.0%, followed by Thailand, China, India, and Myanmar. Other countries like the USA, Brunei, and Malaysia are also represented.



(a) Based on affiliation country



(b) Based on origin country

Figure 18: SEA-VL author demographic based on their affiliation or origin countries.

In terms of the demographic breakdown of image validators, the majority of image validators are from Indonesia, followed by the Philippines, Thailand, and non-SEA countries. Other countries with smaller representation include Singapore, Myanmar, and Brunei, Malaysia, and Vietnam, as shown in Figure 19.

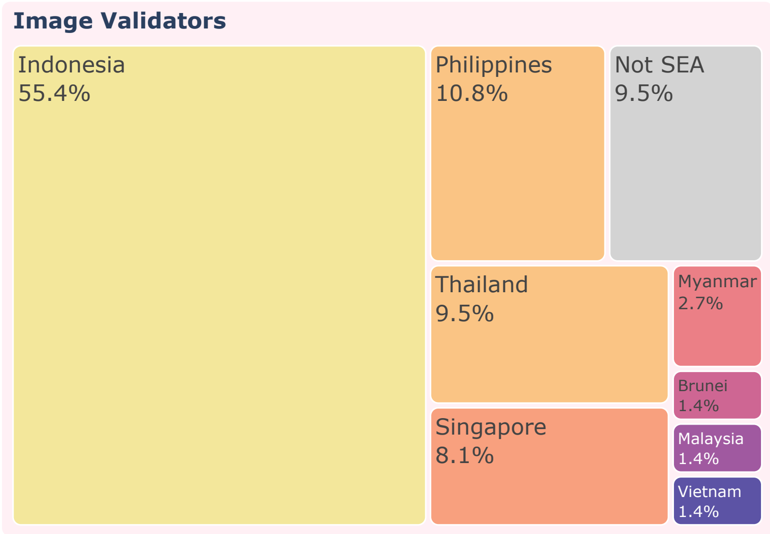


Figure 19: SEA-VL image validator demographic based on their origin countries.

H Contribution Point System

We discuss additional information regarding the contribution point system formulated at the start of the SEA-VL initiative as a form of credit attribution to collaborators from the community. We provide a breakdown of the types of contribution activities and their corresponding points to be awarded in Table 6. For transparency, this point system is discussed at every town hall meeting to inform new collaborators and volunteers to the project. We categorize the types of contributions into two: **open contribution** which includes image collection and image validation and are open to any potential collaborators and volunteers and; **closed contribution** which includes performing model experiments, evaluations, paper writing, and management, coordination, and communication of project progress. Tasks under closed contribution are assigned to selected collaborators and original initiators of the project who have the resource and compute capacity such as GPU equipment (see Appendix D for more details) and experience in paper writing.

Similar to previous corpus-building initiatives such as SEACrowd (Lovenia et al., 2024), CVQA (Mogrovejo et al., 2024), and WorldCuisines (Winata et al., 2024), we set a threshold of **200 points** for co-authorship denoting significant contribution in this project (Figure 20). We resolve the authorship order based on the decreasing order of points (collaborators with the highest number of points will either come first or come last, depending on their preference). On the other hand, collaborators who did not reach the given threshold will be acknowledged instead.

Activity	Awarded Points
Image Collection	2 pts per image (Indonesia, Singapore, and the Philippines), 3 pts per image (Thailand, Malaysia, and Vietnam), 4 pts per image (Brunei, Laos, Cambodia, and Myanmar, East Timor)
Image Validation	1 pt per image
Model Experiments	100 pts - no limit (based on difficulty, compute resources, and time)
Evaluation Procedures	100 pts - no limit (based on difficulty, compute resources, and time)
Paper Writing	100 pts - no limit (based on designated sections)
Management, Coordination, and Communication	100 pts - no limit (based on difficulty, compute resources, and time)

Table 6: The point system used for crediting various forms of contributions from the collaborators of the community. We set **200 points** as the threshold for co-authorship. We resolve the authorship order based on the decreasing order of points (collaborators with the highest number of points will come first).

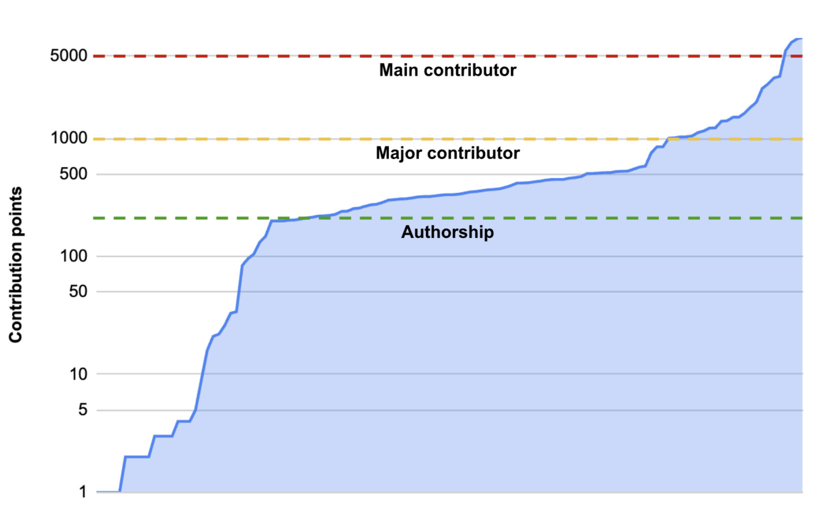


Figure 20: SEA-VL contribution points and thresholds.

I Human Evaluation

I.1 Annotator Distribution

All human evaluators involved in the evaluations described in Sections 4 and 5 are either originally from or affiliated with **Southeast Asia**. The breakdown of the annotators in each experiment is as follows:

- **Image Captioning**: 1 from Singapore, 1 from the Philippines, and 1 from Indonesia.
- **Image Deduplication**: 3 from Indonesia.
- **Image Filtering**:
  - **CC3M**: 2 from Indonesia and 1 from Thailand.
  - **COYO**: 2 from Indonesia and 1 from Singapore.
  - **WIT**: 2 from Indonesia and 1 from the Philippines.
- **Image Generation**: 2 from Indonesia and 1 from the Philippines.

I.2 Image Filtering Evaluation

The Image Filtering Evaluation process involves human annotators assessing images from three datasets: WiT, CC3M, and COYO. All annotators are given the same set of samples for each dataset, where 50 images are randomly selected from each of the following dataset tiers: bronze, silver, gold, platinum, and diamond, resulting in a total of 250 images per dataset. For each image, annotators are asked to classify it as "Yes," "No," or "Not Sure" based on whether the image is relevant to SEA. Three annotators are assigned to WiT and COYO, while five annotators work on CC3M. The evaluation results are then aggregated and averaged across the annotators for each dataset and category. We measure the inter-annotator agreement of the human evaluation using γ coefficient (Mathet et al., 2015; Mathet, 2017).

I.3 Image Duplication Evaluation

Given two images, three annotators were asked to assess whether the images are duplicates (binary decision). The metric "duplicated" was defined loosely to the annotators, with annotators encouraged to consider whether having the image pair would be redundant for training purposes. Each annotator was provided with the same set of 75 image pairs, each related to either cuisine or tradition. Finally, for each cultural domain, the duplication score was averaged across the three annotators.

I.4 Image Generation Evaluation

Three annotators were assigned to assess the quality of the generated images. Each annotator was tasked with evaluating a distinct set of 250 samples, each focusing on a specific cultural domain: one annotator

Score	Correctness Description
3	The image correctly describes the given query.
2	The image somewhat correctly describes the given query.
1	The image is irrelevant to the query.

Table 7: Scoring rubric for **correctness** in **Image Generation Evaluation**.

Score	Naturalness Description
3	The image is natural and culturally relevant.
2	The image feels somewhat natural.
1	The image is unnatural and looks machine generated.

Table 8: Scoring rubric for **naturalness** in **Image Generation Evaluation**.

assessed cuisine, another assessed landmarks, and the third assessed traditions. Each sample consisted of a query (caption) describing an aspect of culture from a specific South East Asian country, along with the corresponding image. The annotators were instructed to evaluate the given caption based on correctness and naturalness according to the rubric in Table 7 and 8 respectively. Finally, for each cultural domain, the correctness and naturalness scores are averaged independently based on the source of the image (whether generated by a model or taken by a human).

I.5 Image Captioning Evaluation

Score	Correctness Description
3	The caption correctly describes the given image.
2	The caption somewhat correctly describes the given image.
1	The caption is irrelevant to the image.

Table 9: Scoring rubric for **correctness** in **Image Captioning Evaluation**.

To assess the quality of generated captions, three annotators were presented with the same set of 450 samples, each consisting of an image along with its corresponding caption. The annotators were instructed to evaluate the given caption based on correctness and naturalness according to the rubric in Table 9 and 10 respectively. Finally, the correctness and naturalness scores are averaged independently based on the source of the generation (whether by a model or a human) across three annotators.

I.6 Failure Mode in SEA Image Captioning

We further conducted human evaluation on 40 examples per language for Indonesian, Malay, Thai, Tagalog, and Vietnamese to break down the failure modes of image captioning for SEA languages. The highlights of the observed failure modes are as follows:

- Answering in English or other non-target languages: Most frequent in Vietnamese (74.4%), Tagalog (28.2%), and negligible in the other languages (<5%).
- Captions that appear unnatural or machine-generated: Common in Tagalog, Thai, and Vietnamese (82–85%), Malay (64.1%), and much less so in Indonesian (7.5%).
- Captions irrelevant to the image: Particularly high in Tagalog (97.4%), Vietnamese (82.1%), Thai (55.0%), Malay (48.7%), and lower in Indonesian (30.0%).

From our observations, Indonesian performs best among the tested languages, likely due to its relatively higher representation in the captioning model's training data. Thai and Vietnamese show the most severe degradation, which we attribute to the usage of Thai (non-Latin) script or Vietnamese alphabets. Despite sharing the Latin script, Tagalog performs poorly, with frequent phrase repetitions leading to irrelevant outputs. Malay captions, while more accurate than Tagalog, often sound unnatural, possibly due to

Score	Naturalness Description
3	The caption seems to be naturally written by native speakers.
2	The caption feels somewhat natural.
1	The caption is unnatural and looks machine-generated.

Table 10: Scoring rubric for **naturalness** in **Image Captioning Evaluation**.

vocabulary overlap with Indonesian, which may cause generation to lean stylistically toward Indonesian rather than native Malay phrasing

J Samples of Images and Captions Collected

We provide qualitative samples of the collected image-text pairs here.[16]

[16] https://github.com/SEACrowd/seacrowd.github.io/blob/master/docs/SEA_VL_Appendix_J.pdf