SEA-VQA: Southeast Asian Cultural Context Dataset For Visual Question Answering

Norawit Urailertprasert, Peerat Limkonchotiwat,
Supasorn Suwajanakorn, Sarana Nutanong
School of Information Science and Technology, VISTEC, Thailand
{norawit.u_s18,peerat.l_s19,supasorn.s,snutanon}@vistec.ac.th

Abstract

Visual Question Answering (VQA) is a critical task requiring the simultaneous understanding of visual and textual information. While significant advancements have been made with multilingual datasets, these often lack cultural specificity, especially in the context of Southeast Asia (SEA). In this paper, we introduce SEA-VQA, aiming to highlight the challenges and gaps in existing VQA models when confronted with culturally specific content. Our dataset includes images from eight SEA countries, curated from the UNESCO Cultural Heritage collection. Our evaluation, comparing GPT-4 and GEMINI models, demonstrates substantial performance drops on culture-centric questions compared to the A-OKVQA dataset, a commonsense and world-knowledge VQA benchmark comprising approximately 25,000 questions. Our findings underscore the importance of cultural diversity in VQA datasets and reveal substantial gaps in the ability of current VQA models to handle culturally rich contexts. SEA-VQA serves as a crucial benchmark for identifying these gaps and guiding future improvements in VQA systems. Our code and dataset are publicly available at https://wit543.github.io/sea-vqa

1 Introduction

Visual question answering (VQA) is the task of answering questions based on an image. As exemplified in Figure 1, one may ask a question involving an object or an action in an image. The VQA system accepts the question and picture as input and answers the questions based on the image's contents. Therefore, the performance of VQA depends on the ability of the model to understand textual and visual information simultaneously. Given its applications in various domains, such as healthcare, autonomous driving, and assistive technologies, VQA is pivotal in advancing human-computer interaction by enabling machines to comprehend and respond to complex visual content and textual queries.

Past efforts in VQA evaluation datasets have generally focused on measuring the reasoning, common knowledge, and image understanding of models. Initially, these datasets (Agrawal et al., 2016) used real-world images paired with straightforward questions, requiring direct answers based on visible elements. Over time, the emphasis shifted towards complex reasoning with datasets like CLEVR (Johnson et al., 2017) and GQA (Hudson and Manning, 2019), which present questions that demand comprehension of relationships, quantities, and spatial awareness. More recent datasets focus on improving generalization across various visual data types and question formats, testing the capabilities of VQA models from multiple perspectives and reasoning tasks (Lu et al., 2022; Yue et al., 2023; Liu et al., 2023; Li et al., 2023; Yu et al., 2023; Wu et al., 2024; Fu et al., 2024).

Although these datasets have been instrumental in advancing VQA, they often lack multicultural aspects. Multilingualism is typically achieved by translating existing queries into multiple languages, which does not fully capture cultural specificities (Gao et al., 2015; Raj Khan et al., 2021; Pfeiffer et al., 2022; Tran et al., 2023). This approach overlooks the nuances and contextual knowledge unique to different cultures, limiting the robustness of VQA systems in diverse settings. For instance, xGQA (Pfeiffer et al., 2022) introduces cross-lingual VQA but focuses more on language translation rather than cultural context. ViCLEVR (Tran et al., 2023) explores visual reasoning in Vietnamese, but it is limited to a single culture and language. Table 1 provides a comprehensive comparison of existing VQA datasets by country, highlighting the diversity in answer types, image sources, languages, and question types across different datasets.

To address this gap, we propose developing a

Proceedings of the 3rd Workshop on Advances in Language and Vision Research, pages 173–185

Dataset	Answer Type	Image Source	Coverage		Question Types		
			Languages	Countries in SEA	General	Reasoning	Culture-centric
General VQA dataset							
A-OKVQA	mc	COCO	1	0	✓	×	×
Multilingual VQA dataset							
xGQA	y/n, open	GQA	8	1	✓	×	×
MaXM	y/n, open	cross3600	7	1	✓	×	×
EVJVQA	open	Self-sourced	3	1	✓	×	×
Our dataset							
SEA-VQA	mc	UNESCO	1	8	×	✓	✓

Table 1: Comparison of existing VQA dataset. Given 'y/n' represents yes/no answer types, 'Open' denotes free-form answer types, and 'mc' indicates multiple-choice questions.

Figure 1: Examples of questions from the SEA-VQA dataset that require an understanding of cultural context. Each question is paired with an image from a specific Southeast Asian country (Thailand, Indonesia, Laos, Vietnam).

VQA dataset that challenges models in comprehending three distinct levels of concepts:

- General world knowledge, e.g., recognizing common entities such as people and animals.
- Specific cultural knowledge unique to each country.
- Understanding the contents of the image itself.

In particular, we develop a culturally specific dataset tailored to the region depicted in the image, incorporating a wider range of languages, including low-resource languages, particularly from Southeast Asia. This approach aims to improve the generalizability of VQA systems and address the current limitations in evaluating VQAs on SEA languages, which remains an open question in the field.

Our approach involves designing a data-gathering pipeline based on the utilization of large language models, such as GPT-4, to formulate questions and answers based on culturally specific images. To ensure quality in question generation, we leverage metadata for cultural questions, including cultural names, countries, and image descriptions, to assist the multi-modal large language model (MLLM) system in generating accurate questions. Additionally, human oversight in the quality-checking process ensures the integrity of the data. Our dataset comprises 515 images, 1,999 questions, and 53 cultures from 8 countries, focusing on the traditions of each culture and the reasoning behind each answer. This culturally specific approach aims to improve the generalizability of VQA systems and address the current limitations in evaluating VQAs on SEA languages, thereby ensuring robustness across diverse cultural and linguistic contexts.

2 Methodology

To formulate our dataset, the data creation pipeline consists of four steps: (i) image curation, (ii) attribute extraction, (iii) QA generation, and (iv) data quality assurance.

2.1 Image Curation

To obtain images from SEA cultures, we curate images from the UNESCO Cultural Heritage col-

lection[1]. This collection is ideal for our purpose as it encompasses a diverse range of culturally significant sites and practices, ensuring that our dataset reflects Southeast Asia's cultural heritage and diversity. Our dataset includes images from 8 countries, totaling 515 images: Cambodia (55 images), Indonesia (139 images), Laos (18 images), Malaysia 64 images), the Philippines (69 images), Singapore (8 images), Thailand (40 images), and Vietnam (122 images). For more information about cultures in each country, please refer to Appendix A.1. We base the number of cultures on those recognized and registered by UNESCO. This approach ensures that the selected cultures are officially recognized. To address the imbalance, we identify the culture each question pertains to and treat the set of questions about a particular culture as a single unit. This method helps avoid cultural imbalance in our dataset.

2.2 Attribute Extraction

The purpose of this step is to enhance the quality and relevance of the QA generation process by providing rich contextual information. Instead of using the image alone to generate questions as proposed by Agrawal et al. (2016); Schwenk et al. (2022), we found that adding more attributes extracted from images is more beneficial. To achieve this, we utilize each image's description, cultural name, and country. These attributes are generated and verified by humans in the next step, ensuring that the information provided contains insightful context for each image. This comprehensive attribute extraction process significantly improves the effectiveness of the QA generation.

2.3 QA Generation

One straightforward method to compose these pairs is by utilizing human annotators (Agrawal et al., 2016; Schwenk et al., 2022; Nguyen et al., 2023). While the human method demonstrates the best data quality, it poses challenges in terms of scalability and broader applicability. Given our goal of introducing a dataset with cultural diversity, it is crucial to develop a repeatable and economically viable approach. Our objective is to balance cost and quality in generating question-answer pairs. Therefore, in our work, we have employed a machine-human collaborative approach in which QA pairs are generated by GPT-4, while humans are employed for quality assurance (see Section 2.4).

We composed a specific instruction prompt for GPT-4 to generate questions that require understanding the depicted culture and reasoning based on detailed descriptions of the image, including cultural and geographical context. To perform an assessment of an MLLM on our dataset, we opted for a multiple-choice format comprising four options: one correct answer and three plausible but incorrect alternatives. Furthermore, to minimize the occurrence of redundant questions, we generate batches of 20 questions simultaneously. We experimented with generating between 1 and 30 questions and found that 20 questions resulted in a diverse set that remained on the topic of culture. Our goal was to determine the maximum number of questions that could still stay relevant to the topic, and we concluded that 20 questions provided the best outcome. Additionally, we instruct GPT-4 to create a question that involves reasoning, and the answers require thought rather than simple observation 1. This strategy significantly improves the diversity and complexity of the dataset. For QA generation analysis, please refer to Appendix A.2.

An example prompt: *"Create 20 challenging multiple-choice questions based on this image that require multi-step reasoning. These questions should be culturally relevant but not explicitly mention the culture in the questions themselves. Each question should have four options: one correct answer and three nearly correct alternatives. Highlight the correct answer in each set with a '<' at the end of the correct answer. Use the descriptive context provided to enhance the complexity of each question. The culture and country depicted in the image are provided below. culture:{culture} Description: {description} country: {country}"*

2.4 Data Quality Assurance

To ensure the quality and validity of our questions, we employ human reviewers to assess and filter out those that are nonsensical, unanswerable, or incorrect. This approach keeps humans in the loop, ensuring that the questions are coherent and appropriate at a reduced cost. Using GPT-4, the total cost for question generation is less than $15, which averages out to about $0.008 per question. This is significantly cheaper compared to a local labeling platform, charging more than $0.28 per question, and even Amazon Mechanical Turk, where the fee starts at $0.01, excluding the reward per question.

[1] https://whc.unesco.org/en/list/

We provide reviewers with detailed guidelines to evaluate the choices and answers in relation to the image, its cultural context, and its description. If reviewers are uncertain about an answer, they are permitted to access external knowledge sources such as a search engine to ascertain the correct response. This process requires the reviewers to consider general world knowledge, the cultural significance of the image, and its content, ensuring that the dataset maintains high standards of accuracy and cultural relevance. To ensure accuracy, we provide images, country names, culture names, and descriptions. If this information is insufficient, reviewers can use additional resources to verify each question. We also provide examples of acceptable and unacceptable questions. Reviewers are graduate students specializing in computer vision (CV) and natural language processing (NLP) from Southeast Asia to ensure familiarity with regional cultures. Table 2 provides a comprehensive overview of the dataset statistics and comparisons.

3 Experimental Results

3.1 Evaluation Setting

Test Models. We use GPT-4-TURBO and GEMINI-PRO-VISION for testing. We use the same prompt for both models. The prompt: *"Answer the following question and provide only the letter output, for example: a, b, c, d. Choose only one option, output only the choice. question:{question} choice: a) {a} b) {b} c) {c} d) {d}"*

We evaluate each question individually by inputting the prompt and image one at a time. In addition, we evaluate MLLMs on the A-OKVQA dataset (Schwenk et al., 2022) to observe the performance changes compared to our VQA dataset.

Evaluation Metrics. We use accuracy scores as the primary metric. In addition, we also demonstrate the performance of each language separately.

3.2 Main Results

Table 3 shows the performance of the two models on two datasets: A-OKVQA (Schwenk et al., 2022), a VQA dataset that requires commonsense reasoning and world knowledge to answer and our proposed dataset, SEA-VQA. We can see that with SEA-VQA, the performance of both models drastically drops compared to A-OKVQA. The results also show that GPT-4 outperforms GEMINI in both datasets, and the gap is larger for SEA-VQA. The table also provides a breakdown in terms of countries. Indonesia is the only dataset portion where GEMINI performs better than GPT-4. Another interesting point to note is that the performance of GPT-4 on the Singapore portion of the dataset (0.688) is substantially higher than the second-highest one, i.e., the Philippines (0.523). One possible explanation is due to the urbanized nature of the city-state. In the big picture, our findings demonstrate the need for improvement and adaptation in VQA systems to handle broader cultural contexts from diverse sources.

3.3 Error Analysis

We organize the error analysis into two parts: errors made by both models and errors made by only one of the models. Both models perform poorly on questions requiring the ability to differentiate subtle variations of cultures originating from the same region or cultures that exist across multiple SEA nations with local variations. Such questions require the knowledge and understanding of differences in attires, musical instruments, and cultural performances that look similar even for humans who are not from this region. For example, the Thai cultural performances of Nora and Khon may look similar to those unfamiliar with the SEA cultural context. Additionally, cultural diffusion across the Southeast Asian region historically means that similar cultures can exist in different countries. This is evident in the Royal Ballet of Cambodia and Thai Khon, which may seem similar to outsiders, but locals can distinguish them by their costumes and dance patterns.

Furthermore, the models struggle particularly with questions that require recognizing specific cultural elements in an image to determine the reasoning behind the action or role of the subjects depicted. Neither model performs well on questions involving musical performance, requiring the ability to recognize musical instruments. For example, consider this question: "This instrument is a part of which traditional performance art?" When shown a canang, which is typically used in Mak Yong theatre, both models incorrectly answer "Wayang Kulit."

In addition to commonly occurring errors found in GEMINI and GPT-4, there are also error patterns specific to either model.

- GEMINI often fails to adhere to instructions to select one answer, frequently outputting multiple choices, e.g., (a, b), or (a, b, c).
- GPT-4 struggles with the determination of a per-

Country	Total			Images/Culture		Questions/Culture		Questions/Image	
	Cultures	Images	Questions	Avg.	Std.	Avg.	Std.	Avg.	Std.
Cambodia	6	55	304	9.17	1.17	50.67	3.56	5.53	2.85
Indonesia	12	139	752	11.58	7.12	62.67	40.22	5.41	3.92
Laos	2	18	72	9.00	1.41	36.00	15.56	4.00	4.24
Malaysia	7	64	189	9.14	1.57	27.00	5.86	2.95	1.46
Philippines	6	69	153	11.50	4.14	25.50	6.72	2.22	1.33
Singapore	1	8	32	8.00	0.00	32.00	0.00	4.00	1.69
Thailand	4	40	184	10.00	0.00	46.00	13.04	4.60	3.12
Vietnam	15	122	313	8.13	3.23	20.87	9.71	2.57	0.73
Over All	53	515	1999	9.57	2.33	37.59	11.83	3.91	2.42

Table 2: The dataset statistics on the number of cultures, images, and associated questions. The table provides metrics on the average number of images per culture and questions per culture and image, complete with standard deviations.

Language	GEMINI	GPT-4
Proposed Dataset, SEA-VQA		
Cambodia	0.257	0.467
Indonesia	0.453	0.336
Laos	0.278	0.375
Malaysia	0.360	0.492
Philippines	0.307	0.523
Singapore	0.219	0.688
Thailand	0.348	0.478
Vietnam	0.176	0.495
Average (Macro)	0.300	0.482
Average (Micro)	0.275	0.365
Existing VQA Benchmark		
A-OKVQA (Micro)	0.760	0.822

Table 3: Accuracy of GEMINI and GPT-4 on culture-specific questions from the SEA-VQA dataset and the general knowledge-based A-OKVQA dataset. The table presents model performance across various Southeast Asian countries.

son's age, the length of objects, and actions within a cultural context. For instance, when asked about the typical range of diameters for instruments shown in an image (an image of a gong in Vietnam), the model incorrectly suggested 15 to 35 centimeters, whereas the correct answer is 25 to 80 centimeters. In response to a question about an image showing a traditional ensemble, the correct label should have been "A khene orchestra concert." However, due to a focus only on visible actions and objects, GPT-4 answer was "A bamboo dance."

These examples highlight areas where the model's accuracy can be improved.

4 Conclusion and Future Work

In conclusion, we propose a VQA dataset for the Southeast Asian cultural context called *SEA-VQA*. Our dataset is generated from MLLM while using humans in the data quality assurance process. Using this approach, we are able to generate 1,999 questions from 8 countries and 53 cultures with limited human efforts. Results from assessments using our SEA-VQA dataset reveal that, although MLLMs demonstrate reasonable performances in standard VQA benchmarks, there is a gap in understanding local cultural knowledge.

In future work, we aim to apply this process to other underrepresented languages and dialects from the region. We plan to explore more languages and images from open-source projects in SEA, i.e., SEACrowd (Lovenia et al., 2024), to extend from monolingual to multilingual VQAs. Additionally, we will explore generating VQAs using multiple models to improve accuracy and robustness. We also plan to add more attribute extraction methods to create more variation in VQAs. In addition, we also plan to explore the integration of virtual reality technology to enhance the richness of the dataset.

5 Limitations

One limitation is the quality of image data and description. Expanding the dataset size requires a greater source of image data; however, ensuring that these images accurately represent the relevant cultures is challenging, thus limiting the number of usable images. For the selection of cultures, we rely on those officially recognized and registered with UNESCO. This approach may restrict

the scope of represented cultures, as many local cultures that are not registered or are in the process of registration are excluded. Despite these limitations, using UNESCO as a source allows us to extend our research beyond Southeast Asia, incorporating cultures from around the globe.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. VQA: Visual Question Answering.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702, Long Beach, CA, USA. IEEE.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, Honolulu, HI. IEEE.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2023. SEED-Bench-2: Benchmarking Multimodal Large Language Models.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023. MMBench: Is Your Multi-modal Model an All-around Player?

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. 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, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus, 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, Ngee Chia Tai, Ayu Purwarianti, Sebastian Ruder, William 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.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering.

Ngan Luu-Thuy Nguyen, Nghia Hieu Nguyen, Duong T. D. Vo, Khanh Quoc Tran, and Kiet Van Nguyen. 2023. VLSP2022-EVJVQA Challenge: Multilingual Visual Question Answering. *Journal of Computer Science and Cybernetics*, pages 237–258.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-Lingual Visual Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland. Association for Computational Linguistics.

Humair Raj Khan, Deepak Gupta, and Asif Ekbal. 2021. Towards Developing a Multilingual and Code-Mixed Visual Question Answering System by Knowledge Distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1753–1767, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge.

Khiem Vinh Tran, Hao Phu Phan, Kiet Van Nguyen, and Ngan Luu Thuy Nguyen. 2023. ViCLEVR: A Visual Reasoning Dataset and Hybrid Multimodal Fusion Model for Visual Question Answering in Vietnamese.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. 2024. Q-Bench: A Benchmark for General-Purpose Foundation Models on Low-level Vision.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu

Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI.

A Appendix

A.1 Cultures List

Table 4 catalogs diverse cultural practices across Southeast Asia and adjacent regions. Organized by country, the table highlights traditional cultural heritage, such as Cambodia's Kun Lbokator and Indonesia's Wayang puppet theatre, demonstrating each nation's commitment to preserving its cultural identity. Noteworthy, entries like Tugging rituals and games are shared between countries, indicating cultural ties that transcend national borders.

A.2 QA Generation Analysis

We can use a language model like GPT-4 to automate the generation of questions, choices, and answers. However, the generated contents may contain inaccuracy, irrelevant information, and formatting inconsistencies. To combat these issues, human involvement is still necessary to ensure the dataset quality.

We observe that nearly 15% of all questions generated from GPT-4 cannot be used for VQA purposes. These are questions lacking definitive answers, e.g., the determination of the time of day the image was captured (e.g., morning, evening, noon, night), the emotion of a person depicted in the image (e.g., happy, excited, stressed, sad). This observation indicates that a significant portion of these questions require further refinement and human oversight to ensure they are appropriate and useful for VQA tasks.

Country	Culture
Cambodia	Kun Lbokator, traditional martial arts in Cambodia
Cambodia	Lkhon Khol Wat Svay Andet
Cambodia	Chapei Dang Veng
Cambodia	Royal ballet of Cambodia
Cambodia	Sbek Thom, Khmer shadow theatre
Cambodia, Philippines, Republic of Korea, Viet Nam	Tugging rituals and games
Indonesia	Jamu wellness culture
Indonesia	Gamelan
Indonesia	Traditions of Pencak Silat
Indonesia	Pinisi, art of boatbuilding in South Sulawesi
Indonesia	Three genres of traditional dance in Bali
Indonesia	Noken multifunctional knotted or woven bag, handcraft of the people of Papua
Indonesia	Saman dance
Indonesia	Indonesian Angklung
Indonesia	Indonesian Batik
Indonesia	Education and training in Indonesian Batik intangible cultural heritage for elementary, junior, senior, vocational school and polytechnic students, in collaboration with the Batik Museum in Pekalongan
Indonesia	Indonesian Kris
Indonesia	Wayang puppet theatre
Indonesia, Malaysia	Pantun
Laos	Traditional craft of Naga motif weaving in Lao communities
Laos	Khaen music of the Lao people
Malaysia	Mek Mulung
Malaysia	Songket
Malaysia	Silat
Malaysia	Dondang Sayang
Malaysia	Mak Yong theatre
Malaysia, China	Ong Chun/Wangchuan/Wangkang ceremony, rituals and related practices for maintaining the sustainable connection between man and the ocean
Philippines	Aklan piña handloom weaving
Philippines	The School of Living Traditions (SLT)
Philippines	Buklog, thanksgiving ritual system of the Subanen
Philippines	Darangen epic of the Maranao people of Lake Lanao
Philippines	Hudhud chants of the Ifugao
Singapore	Hawker culture in Singapore, community dining and culinary practices in a multicultural urban context
Thailand	Songkran in Thailand, traditional Thai New Year festival
Thailand	Nora, dance drama in southern Thailand
Thailand	Nuad Thai, traditional Thai massage
Thailand	Khon, masked dance drama in Thailand
Viet Nam	Art of pottery-making of Chăm people
Viet Nam	Art of Xòe dance of the Tai people in Viet Nam
Viet Nam	Practices of Then by Tày, Nùng and Thái ethnic groups in Viet Nam
Viet Nam	The art of Bài Chòi in Central Viet Nam
Viet Nam	Xoan singing of Phú Thọ province, Viet Nam

Continued on next page

Country	Culture
Viet Nam	Practices related to the Viet beliefs in the Mother Goddesses of Three Realms
Viet Nam	Ví and Gim folk songs of Ngh Tĩnh
Viet Nam	Art of Đn ca tài t music and song in southern Viet Nam
Viet Nam	Worship of Hùng kings in Phú Th
Viet Nam	Gióng festival of Phù Đông and Sóc temples
Viet Nam	Ca trù singing
Viet Nam	Quan H Bc Ninh folk songs
Viet Nam	Nha Nhac, Vietnamese court music
Viet Nam	Space of gong culture

Table 4: **Table of Cultures**: The table organizes the cultures used in the dataset by country, providing a comprehensive overview of diverse cultural elements across different nations.

A.3 More Culture Examples

We provided more examples of challenging cultural elements to elevate the visual question-answering (VQA) capabilities of our dataset. The text highlighted in green represents the correct answer, while the responses from GPT-4 and GEMENI are displayed in the box below.

These examples feature Khaen music of the Lao people, a traditional form of music recognized by UNESCO for its unique use of bamboo pipes; Songket weaving from Malaysia, a luxurious fabric interwoven with gold and silver threads; Aklan piña handloom weaving from the Philippines, known for its intricate process of weaving pineapple leaf fibers; and children playing the Suling, a key instrument in the Gamelan ensemble of Indonesia. Each example has been carefully selected to challenge the understanding and appreciation of these unique cultural expressions.

Example: Laos

Figure 2: **Culture:** Khaen music of the Lao people

What type of traditional ensemble performance is shown in the image?
A. A choir concert
B. A bamboo dance
C. A khene orchestra concert
D. A traditional puppet show

GPT4	GEMINI
B	C

Example: Malaysia

Figure 3: **Culture:** Songket

What is the name of the fabric pattern used in this headgear?
A. Batik
B. Pua Kumbu
C. Songket
D. Tenun

GPT4	GEMINI
B	A

Example: the Philippines

Figure 4: **Culture:** Aklan piña handloom weaving

What characteristic makes the tool in the image appropriate for fiber extraction?
A. Flexibility
B. Sharpness
C. Weight
D. Porosity

GPT4	GEMINI
B	A

Example: Indonesia

Figure 5: **Culture:** Gamelan

Which musical instrument is predominantly played by the children in the image?
A. Angklung
B. Kendang
C. Suling
D. Bonang

GPT4	GEMINI
C	A