Seeing Culture:
A Benchmark for Visual Reasoning and Grounding

Anonymous ACL submission

Abstract

Multimodal vision-language models (VLMs) have achieved substantial progress in various tasks that demand combined understanding of visual and textual content, particularly in cultural understanding tasks, with the emergence of new cultural datasets. However, these datasets frequently fall short of providing cultural reasoning while underrepresenting many cultures. In this paper, we introduce the Seeing Culture Benchmark (SCB), focusing on cultural reasoning with a novel approach that requires VLMs to reason on culturally rich images in two stages: i) selecting the correct visual option in a multiple-choice visual question answering (VQA) approach, and ii) segmenting the relevant cultural artifact as evidence of reasoning. Visual options in the first stage are systematically organized into three types: those originating from the same country, those from different countries, or a mixed group. Notably, all options are derived from a singular category for each type. Progression to the second stage occurs only after a correct visual option is chosen. Our benchmark encompasses 1,065 images capturing 138 cultural artifacts across five categories from seven Southeast Asia (SEA) countries, whose diverse cultures are often overlooked. Additionally, the benchmark provides 3,178 questions, of which 1,093 are unique and meticulously curated by human annotators. Our evaluation of various VLMs reveals the complexities involved in cross-modal cultural reasoning and underscores the disparity between visual reasoning and spatial grounding in scenarios that are culturally nuanced. The SCB serves as a crucial benchmark for identifying these shortcomings so as to guide future developments in the field of cultural reasoning.

1 Introduction

Recent multimodal VLMs are impressive on various tasks, such as VQA and visual grounding, which require assessing the understanding of visual and textual information. For instance, VQA

Figure 1: Comparison between our benchmark (SCB) and the recent studies on cultural understanding (Mogrovejo et al., 2024; Bhatia et al., 2024) and reasoning (Urailertprasert et al., 2024). SCB requires reasoning on cultural artifacts via diverse and rich visuals.

Figure 2: The presented collection of images from our SCB encompasses visual representations of cultural concepts from seven countries, categorized across five dimensions: music, game, dance, celebration, and wedding. These images exhibit either a variety of cultural artifacts situated in diverse contexts (e.g., the depiction of the *Balinese legong dance* showcases multiple characters, two *princesses rangkesari*, and one *condong*, with corresponding questions) or integrated distractors in addition to the primary concept (e.g., the image featuring the *banduria*, which displays Spanish guitars on the right side while the *bandurias* are positioned on the left). The segmentation masks of concepts are best viewed in color.

tasks with open-ended or multiple-choice questions have been used on various generic topics such as healthcare and entertainment. At the same time, visual grounding, which entails the segmentation of an object based on textual input, has predominantly expanded on general scene understanding via recent VLMs. However, their performance may vary significantly across different cultural contexts, highlighting the need for new benchmarks to assess and improve their performance in diverse cultural contexts. While recent studies (Nayak et al., 2024; Wang et al., 2025; Mogrovejo et al., 2024; Bhatia et al., 2024) attempt to address this gap with the focus of cultural understanding, there remains a pressing need for more comprehensive datasets that encapsulate a wider array of cultural nuances and artifacts, ensuring that VLMs can reason on culturally specific queries. We must emphasize that cultural reasoning involves not only the recognition of cultural artifacts but also the realization of their significance within specific contexts. For instance, considering our example in Figure 1, certain clues need to be taken into account, such as that *barong dance* belongs to a specific culture to differentiate it from other visual options, as well as the various characters that symbolize different meanings.

Creating such adequate benchmarks on cultural reasoning is challenging due to the various factors influencing cultural representation, such as the selection of images, the formulation of questions, and the data collection process. Despite providing essential insights, the present benchmarks exhibit significant limitations. For instance, (Urailertprasert et al., 2024; Baek et al., 2024; Liu et al., 2025; Schneider et al., 2025) focus on the cultural reasoning VQA; however, many of the images do not have any distractors, focusing on only the cultural concept, while the questions are AI-generated, which may lack authenticity in cultural representation. Additionally, textual answers to the traditional VQA approaches may be influenced by spurious correlations (Fu et al., 2023; Liu et al., 2023; Zhang et al., 2023; Wang et al., 2023; Zhang et al., 2024) regardless of their design, as addressed by recent works. Furthermore, benchmarks specific to the segmentation task in this context have yet to be developed.

To this end, we propose SCB, a novel benchmark to assess the cultural reasoning of VLMs in Southeast Asia countries, providing diversity in culture, given its low resources in cultural representation within existing datasets. SCB includes complex images with rich and varied cultural contexts, paired with thoughtfully crafted questions that challenge the model's understanding and reasoning of cultural specifics in two stages: i) The multiple-choice options contain images representing diverse cultural artifacts, ii) The segmentation of cultural artifacts plays a role as evidence of reasoning. Advancement to the subsequent stage takes place only by following an accurate visual selection. Moreover, by integrating input from native speakers and cultural experts, we ensure that the questions reflect authentic cultural narratives and avoid biases in AI-generated content. Thus, our approach provides a more holistic view of the context and requires VLMs to reason about the relationships between different cultural elements, enhancing the depth of cultural reasoning. Our benchmark consists of five main categories, 138 cultural concepts, 1,065 images, and 3,178 questions from seven Southeast Asian countries as depicted in Figure 2.

Further, we systematically evaluate several state-of-the-art VLMs on three distinct types. Type 1 consists of options originating from the same country, while Type 2 encompasses options from different countries in relation to the correct answer. Type 3 consists of a blend of Type 1 and Type 2 options. The sole commonality among these types is category consistency for all options (e.g., dance). The results indicate that VLMs perform the least on Type 1 questions, display the highest performance on Type 2 questions, and exhibit intermediate performance on Type 3 questions. This suggests that cues within the questions regarding the country or specific regional cultures can aid in discerning the correct answer. Moreover, there is a notable discrepancy between visual reasoning and spatial grounding, suggesting that although VLMs may select the correct option, they frequently lack the capacity to substantiate their reasoning through grounding. Consequently, the SCB is vital for fostering cross-modal reasoning within a culturally sensitive framework, simultaneously shedding light on the disparity between visual reasoning and grounding. Our research will aid in developing more culturally conscious models, thereby improving their functionality in reasoning across diverse cultural contexts.

2 Related Work

2.1 Benchmarks for Cultural Understanding

The domain has seen the emergence of various recent multicultural vision-language datasets and benchmarks that incorporate explicit cultural taxonomies and tailored tasks (e.g., culture-aware VQA, grounding, and captioning), as shown in Table 1. For example, Crossmodal-3600 (Thapliyal et al., 2022), MOSAIC (Burda-Lassen et al., 2025), and MosAIC (Bai et al., 2025) are primarily centered on image captioning tasks. In contrast, while SEA-VL (Cahyawijaya et al., 2025) includes an image captioning component, its predominant emphasis is on image generation, similar to the approach taken by MosAIG (Bhalerao et al., 2025). Numerous studies examine VQA in various settings. For example, MTVQA (Tang et al., 2024), CulturalVQA (Nayak et al., 2024), and a part of CVLUE (Wang et al., 2025) have open-ended questions, while CROPE (Nikandrou et al., 2025) employs binary (True/False) questions. More relevant to our work, GD-VCR (Yin et al., 2021), CVQA (Mogrovejo et al., 2024), a part of CultureVerse (Liu et al., 2025), and a part of GIMMICK (Schneider et al., 2025) feature multiple-choice questions within the framework of cultural understanding. Unlike these studies that utilize textual options, our research incorporates visual alternatives. It is important to note that we present SCB in a single row, while some other studies are reported separately according to their specific tasks, as our evaluation combines two tasks, unlike the others that evaluate each separately. Besides, GlobalRG (Bhatia et al., 2024) and a part of CVLUE (Wang et al., 2025) address visual grounding of cultural artifacts using bounding boxes (BB), relying on straightforward prompts that include the keyword concept. In contrast, our research tackles questions that necessitate reasoning and employs a semantic segmentation mask that emphasizes fine-grained details.

2.2 Benchmarks for Cultural Reasoning

Cultural reasoning is a critical aspect that distinguishes mere cultural understanding from deeper cognitive engagement with cultural contexts. From this point of view, various studies bridge the gap in the VQA task. For instance, MaRVL (Liu et al., 2021) is the first dataset focusing on cultural reasoning; however, its objective is restricted to determining the truth value of specific image captions. SEA-VQA (Urailertprasert et al., 2024), K-Viscuit (Baek et al., 2024), and a few parts of CultureVerse (Liu et al., 2025) and GIMMICK (Schneider et al., 2025) focus on cultural reasoning through multiple-choice VQA. However, the multiple-choice responses in these studies are textual, and the questions are generated by AI, subsequently refined by human annotators, as seen in other related works. Additionally, unlike our study, these datasets lack a defined framework for selecting complex images, as discussed in Section 3. Only FoodieQA (Li et al., 2024) has visual options akin to our research and features human-

Dataset	Country	Category	Concept	Image	Question	Image Complexity	Input	Question Type	Task Format	Question Creation	Segment Creation
Crossmodal-3600 (Thapliyal et al., 2022)	36	-	100	3,600	-	Normal	Prompt + An Image	CU	Image Captioning	-	-
MOSAIC (Burda-Lassen et al., 2025)	-	-	336	1,500	-	Normal	Prompt + An Image	CU	Image Captioning	-	-
MosAIC (Bai et al., 2025)	3	14	700	2,832	-	Normal	Prompts + An Image	CU	Image Captioning	-	-
SEA-VL (Cahyawijaya et al., 2025)	11	-	-	1.3M	-	Normal	Prompts + An Image	CU	Image Generation and Captioning	-	-
MosAIG (Bhalerao et al., 2025)	5	-	25	9,000	-	Normal	Prompt	CU	Image Generation	-	-
GD-VCR (Yin et al., 2021)	4	-	10	328	886	Normal	Question + An Image + Textual Choices	CU	MCVQA	Human	-
MTVQA (Tang et al., 2024)	10	20	-	2,116	6,778	Normal	Question + An Image	CU	Open-ended VQA	Human	-
CVQA (Mogrovejo et al., 2024)	30	10	-	5,239	10,374	Normal	Question + An Image + Textual Choices	CU	MCVQA	Human	-
CulturalVQA (Nayak et al., 2024)	11	5	13	2,328	2,328	Normal	Question + An Image	CU	Open-ended VQA	AI + Human	-
CROPE (Nikandrou et al., 2025)	5	-	158	1,060	1,060	Normal	Question + An Image + Textual Choices	CU	Binary VQA	Human	-
CVLUE-VQA (Wang et al., 2025)	1	15	92	7,169	7,169	Normal	Question + An Image	CU	Open-ended VQA	Human	-
CultureVerse-IR (Liu et al., 2025)	188	15	11,085	11,085	11,085	Normal	Question + An Image + Textual Choices	CU	MCVQA	AI + Human	-
CultureVerse-SR (Liu et al., 2025)	188	15	11,085	11,085	11,085	Normal	Question + An Image + Textual Choices	CU	MCVQA	AI + Human	-
GIMMICK-COQA (Schneider et al., 2025)	144	5	728	6,857	982	Normal	Question + # of Images + Textual Choices	CU	MCVQA	AI + Human	-
MaRVL (Liu et al., 2021)	5	18	447	4,914	5,670	Normal	Statement + # of Images + Textual Choices	CR	Binary VQA	Human	-
FoodieQA (Li et al., 2024)	1	14	-	389	403	Normal	Question + # of Images as Visual Choices	CR	MCVQA	Human	-
SEA-VQA (Urailertprasert et al., 2024)	8	-	53	515	1,999	Normal	Question + An Image + Textual Choices	CR	MCVQA	AI + Human	-
K-Viscuit (Baek et al., 2024)	1	10	-	237	420	Normal	Question + An Image + Textual Choices	CR	MCVQA	AI + Human	-
CultureVerse-CK (Liu et al., 2025)	188	15	11,085	11,085	11,085	Normal	Question + An Image + Textual Choices	CR	MCVQA	AI + Human	-
GIMMICK-CIVQA (Schneider et al., 2025)	144	5	635	1,928	2,233	Normal	Question + An Image + Textual Choices	CR	MCVQA	AI + Human	-
GIMMICK-CKQA (Schneider et al., 2025)	144	5	635	6,857	728	Normal	Question + An Image + Textual Choices	CR	MCVQA	AI + Human	-
GlobalRG (Bhatia et al., 2024)	15	20	220	3,591	-	Normal	Prompt + An Image	CU	Visual Grounding	-	Human, BBox
CVLUE-VG (Wang et al., 2025)	1	15	92	7,169	5,385	Normal	Prompt + An Image	CU	Visual Grounding	-	Human, BBox
Seeing Culture Benchmark (SCB)	7	5	138	1,065	3,178	Complex	I) Question + An Image + Textual Choices II) Question + An Image	CR	I) MCVQA, II) Visual Grounding	Human	Human, Polygon

Table 1: Comparison between SCB and related works is divided into three distinct sections. The initial section addresses works not concentrating on VQA or visual grounding tasks. The subsequent portion focuses on VQA-related studies, while the final section pertains to visual grounding-related research. Here, "CU" stands for cultural understanding, and "CR" signifies cultural reasoning. "MCVQA" refers to multiple-choice VQA. We filter out images that depict only a single object or lack distractor objects, making our images complex compared to the others. This analysis underscores the distinctive contributions of SCB in furthering the development of cultural visual reasoning and grounding within the field.

constructed questions; nonetheless, it has a limited scope, focusing exclusively on Chinese cuisine. Moreover, the concept of visual grounding, which involves extracting evidence from an image to substantiate reasoning, has not been previously examined.

3 SCB Benchmark

Existing cultural benchmarks for Vision-Language Models (VLMs) exhibit several limitations, as detailed in Table 1. In terms of these limitations, we observe the following: 1) the **questions** fail to foster both cultural reasoning and spatial grounding, 2) there is a scarcity of humanized **questions**, leading to a reliance on mechanical, AI-generated queries, 3) the **images** provided are often not sufficiently complex to challenge VLMs, e.g. lack of distractors. To address these challenges, the SCB provides a more nuanced approach by incorporating culturally rich images and authentic questions that reflect diverse cultural narratives. Further elaboration is provided in the respective sections.

Taxonomy. We adopt a hierarchical framework to categorize cultural elements. Each national culture is subdivided into five principal categories: music, game, dance, celebration, and wedding. Within these categories, specific cultural concepts are delineated, allowing for a structured representation that can be expressed in the format of country/category/concept, *e.g. Cambodia/music/Khaen*. It is important to note that these categories are mutually exclusive; for instance, the *music* category pertains solely to musical instruments, whereas the *wedding* category encompasses garments and other cultural artifacts associated with the wedding ceremony. Additionally, some concepts may incorporate multiple characters or objects. For example, in Figure 1, the concept of the *barong dance* includes two characters, *barong* and *monkey*. This approach facilitates a comprehensive understanding of cultural diversity and its manifestations across different societies.

Countries. To establish a benchmark that accurately encapsulates cultural diversity, we have selected seven underrepresented Southeast Asia countries, including Cambodia, Myanmar, Indonesia, Vietnam, the Philippines, Malaysia, and Thailand. This selection underscores the importance of recognizing and valuing the rich tapestry of cultural identities within this region.

Concepts. We solicit suggestions for cultural concepts based on the defined categories for each country using a Large Language Model (LLM), ChatGPT (OpenAI, 2014). Following this, we conduct a survey to gather insights from local individuals representing each culture, either in English or their local language, to reach authentic images during the image crawling process. The survey aims to refine and validate the concepts proposed by the LLM, with two to three respondents from each country. Ultimately, we distill the results to identify concepts that receive unanimous agreement among the participants. A similar approach is applied to potential characters or objects associated with these concepts. A range of statistical visualizations regarding concepts and questions is presented in Figures 3 and 4.

Images. We crawl via Google Images based on the concepts we identify, collecting 150 images for each concept. Subsequently, we enlist human annotators to carry out manual filtration to ensure the quality of the images. This filtration process assesses whether the retrieved images: i) are relevant to the concept keyword, ii) depict real-world scenarios, iii) are free from duplication, iv) do not have the cultural artifact completely or predominantly obscured, meaning images that are excessively focused on the cultural artifact with a blurry background are excluded, v) contain various distracting objects or scenes, preferably related other cultural artifacts, which may cause conflict to other cultural concept(s) vi) yet sufficiently clear to identify the cultural artifact. The initial three steps, which are standard practice in other datasets, reduce the image count from approximately 20,000 to 4,000. Nonetheless, the final three steps distinguish our image-collecting process. We also incorporate 32 images from the SEA-VL (Cahyawijaya et al., 2025) dataset. Ultimately, through meticulous review, we ensure that the SCB consists of 1,065 unique images.

Segmentation. Upon the selection of images, annotators operate an online segmentation tool (Skalski, 2019) to segment the corresponding concept keywords or their associated cultural artifacts, such as characters in a local dance, or objects used for certain celebrations. This can be illustrated in Figure 2, particularly in the segments denoted as *Indonesia/dance/balinese legong* and *Thailand/celebration/songkran festival*. Note that segmentation is performed using polygons instead of bounding boxes to ensure the capture of intricate details.

Question Formulation. We instruct annotators to formulate unique questions that are culturally aligned with the specific artifacts segmented in the images, while refraining from using templates. Specifically, questions should not refer directly to

Figure 3: Word clouds illustrating the concepts of 1,093 unique questions in SCB are categorized into five cultural themes: wedding, game, music, celebration, and dance. The variation in font size within these clouds reflects the frequency of concept occurrences relevant to each theme.

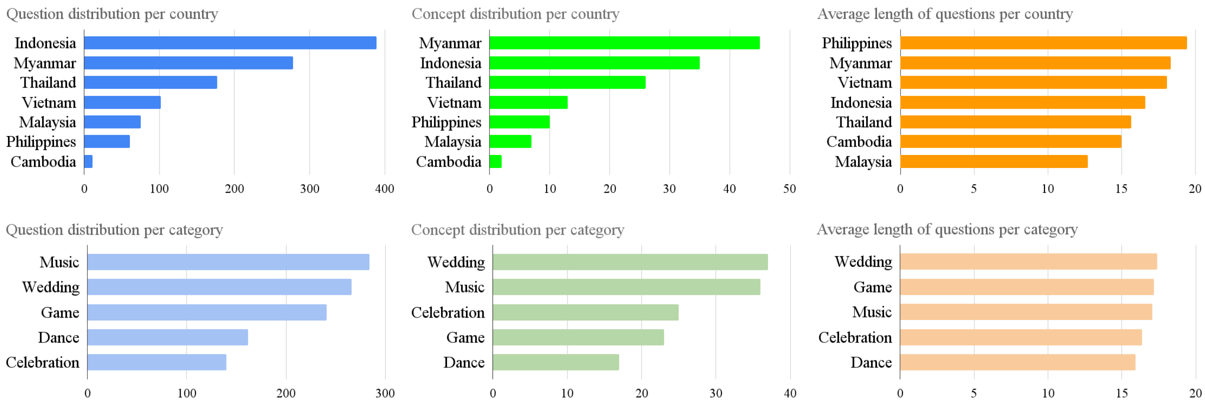

Figure 4: The figures encompass a comprehensive analysis of the distribution of unique questions, concepts, and the average length of questions, segmented by both country and category.

the artifact itself but rather to the symbols or cultural significance associated with it. Annotators are instructed to rely solely on their cultural knowledge, deliberately excluding any AI-generated sources. This ensures that each question requires a deeper reasoning of the culture authentically. For instance, the question, "In a traditional Thai wedding, what symbolizes the spiritual connection and blessings given to the couple by elders or religious figures?", pertains to the artifact represented by *Thailand/wedding/double auspicious headband*, which is accompanied by a prompt of "Locate the artifact in the image." as well. Subsequently, annotators adapt the questions into a VQA format. Following the same question, this can be seen as: "Which image is associated with a traditional Thai wedding artifact that symbolizes the spiritual connection and blessings given to the couple by elders or religious figures?". This is further refined by omitting the segmentation-oriented prompt. In addition, annotators are tasked with providing a rationale for the correct answer, drawing from either online resources or their own cultural knowledge.

Multi-Choice Questions and Visual Options. We extend these unique multiple-choice VQA questions into three types by utilizing varying visual options in our selection process. The foundation of this approach is to utilize the same question paired with its corresponding answer as the correct option, while the incorrect options are selected using three distinct pooling strategies derived from other data instances: Type 1, which sample concepts within the same category and country, Type 2, which sample concepts within same category but completely different country for all options, and Type 3, which consists of balanced mix of Type 1 and Type 2 through a rule-based choice-swapping. For instance, for each randomly chosen two options from the Type 1 question, including the ground truth (GT) choice, we randomly sample the other two options from Type 2 questions, balancing the options country-wise. To mitigate potential biases in this combination, each question is limited to a maximum of two repetitions for Type 3. Additionally, the number of images utilized for visual options is capped at 20 for all types. The breakdown algorithms for all pooling types can be found in Appendix A.2. The quantity of Type 1, Type 2, and Type 3 questions is 834, 840, and 1,504, respectively.

Model	Type 1		Type 2		Type 3		Overall	
	Acc	Mean IoU	Acc	Mean IoU	Acc	Mean IoU	Acc	Mean IoU
InstructBLIP	11.07	–	10.31	–	11.04	–	10.86	–
Idefics2	13.21	0.19	11.03	0.05	12.30	0.18	12.21	0.15
Llama-3.2	23.57	–	25.66	–	23.80	–	24.23	–
LLaVA-Onevision	26.43	–	25.18	–	23.47	–	24.70	–
MiniCPM-2.6	28.33	–	34.65	–	32.85	–	32.13	–
InternVL2.5-4B	30.83	28.37	30.34	28.88	32.18	28.49	31.34	28.56
Qwen2.5-VL-7B	44.17	**44.90**	61.51	**48.22**	54.85	**47.60**	53.78	**47.20**
GPT-4.1	68.33	13.31	90.17	14.32	85.04	13.60	81.97	13.74
Gemini-2.5-Pro	71.07	16.56	90.17	16.67	85.44	15.79	82.88	16.22
GPT-o3	**73.69**	31.10	**91.13**	32.50	**88.23**	31.69	**85.15**	31.78

Table 2: Detailed performance benchmark with several VLMs on our Visual Reasoning and Grounding task. The upper section focuses on open-source VLMs, whereas the lower section pertains to closed-source models.

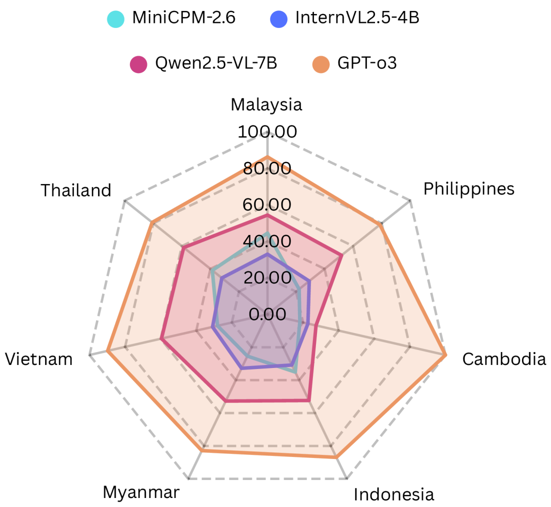

Figure 5: The overall multiple-choice VQA accuracy of certain VLMs across different countries.

4 Experiments

4.1 Visual Reasoning and Grounding Task

We perform a zero-shot evaluation utilizing the following prompt in the initial phase: a textual question for VQA alongside visual options. The output corresponds to one of the provided image options. To assess performance, we employ accuracy as the metric, in accordance with established methodologies in multiple-choice VQA tasks (Zhu et al., 2016; Nayak et al., 2024). In the initial phase, questions that are accurately addressed with the appropriate visual option advance to a subsequent stage to segment the cultural artifacts, while those that are not are excluded. In the following phase, given an image I and a question q that pertains to a cultural term, the objective is to generate a segmentation mask R that delineates the area in I relevant to q. We evaluate performance using bounding boxes (BB) rather than polygons, as current VLMs capable of both VQA and segmentation are restricted to grounding at the BB level. Consequently, the performance of the models is assessed by measuring the overlap between the predicted regions of interest and GT masks, employing Intersection over Union (IoU) as the evaluation metric: $IoU = \frac{R \cap R_{GT}}{R \cup R_{GT}}$. We then report it as the mean IoU.

VLMs used for evaluations. We conduct a comparative analysis of various advanced VLMs. This includes closed-source models such as GPT-4.1, GPT-o3 and Gemini Pro 2.5, alongside a diverse selection of open-source models that vary in size: IntructBLIP 7B (Dai et al., 2023), Idefics2 8B (Laurençon et al., 2024), LLama 3.2 11B (Dubey et al., 2024), LLaVa-OneVision 7B (Li et al., 2025), MiniCPM 2.6 8B (Yao et al., 2024), InternVL2.5 4B (Chen et al., 2024) and Qwen2.5-VL 7B (Yang et al., 2024). It is important to note that we do not employ VLMs capable of segmentation but not suited for multiple-choice VQA, given the requirements of our task.

4.2 Results

How do VLMs' performance vary across different question types? The findings presented in Table 2 reveal that VLMs, both open-source and closed-source, exhibit their poorest performance when the visual options originate from the same country, whereas they display the highest performance when the visual options come from different countries. This pattern can largely be explained by the contextual clues embedded in the questions that pertain to specific countries or cultures. As a result, VLMs are more adept at eliminating alter-

Figure 6: Two failure examples. All VLMs answer the multiple-choice VQA example on the left side incorrectly. The spatial grounding example on the right side is from GPT-o3, although it is the only VLM that correctly answers its multiple-choice VQA version. The blue character on the left identifies the accurate segment.

native visual options that may include indicators from diverse countries. Notably, the correct answer choices (a, b, c, and d) are evenly distributed in our multiple-choice VQA dataset, each accounting for approximately 24% to 26% of the total. This distribution remains consistent across all subsets. Based on this distribution, the expected accuracy of random guessing is approximately 25%. Furthermore, it is observed that 8.5% of the multiple-choice questions are consistently answered incorrectly by all three closed-source models. **Can VLMs validate their reasoning by segmenting the cultural artifact?** A notable discrepancy exists between visual reasoning capabilities and spatial grounding. For example, while GPT-o3 achieves an accuracy exceeding 90%, its mIoU score does not surpass 33%. This disparity is even more pronounced in other closed-source VLMs. Conversely, Qwen exhibits a smaller gap, considering its superior spatial grounding performance and lower efficacy in multiple-choice VQA. Overall, this suggests that, although VLMs may frequently select the correct answer, they often fail to underpin their reasoning with adequate grounding. **Do VLMs perform better in specific countries or categories?** As illustrated in Figure 5 regarding the multiple-choice VQA stage, Qwen demonstrates superior performance when compared to other open-source VLMs; however, it still significantly trails behind GPT-o3. Notably, GPT-o3 achieves its highest performance in Cambodia, whereas Qwen performs least effectively in the same country. The remaining open-source models are considerably less performant than Qwen and display relatively varied outcomes among themselves. **Qualitative results.** Figure 6 presents examples of failures. The left side image illustrates that all presented VLMs are unable to select the appropriate visual option within the same country. The prediction is easier for options involving multiple objects, as seen on the right side, due to more distinguishable image features. In contrast, visual grounding is more difficult because similar yet distinct candidates can confuse the model. Specifically, GPT-o3 correctly selects the correct option but fails to identify the supporting evidence. Overall, GPT-o3 achieves an MCQ accuracy of 94.79% on this query type—higher than its performance on all other query types—while its mIoU is 27.33%, the lowest among all. More results and details can be found in the Appendix, such as Table 3 and 4.

5 Conclusion

In conclusion, this paper presents the Seeing Culture Benchmark (SCB), which addresses the need for improved cultural reasoning in multimodal VLMs. By employing a two-stage approach incorporating VQA and cultural artifact segmentation, we provide a framework for assessing VLMs on culturally rich images from seven Southeast Asia countries. Our dataset includes 1,065 images and 3,178 curated questions, highlighting the underrepresented cultural diversity of the region. Our findings reveal the significant challenges of cross-modal cultural reasoning, emphasizing the need for enhanced visual reasoning and spatial grounding in culturally nuanced contexts. SCB is a vital resource for advancing research in this domain and addressing identified shortcomings in existing VLMs.

Limitation

We acknowledge several constraints in our approach.

Cultural Representation. Our objective was to encompass all countries in Southeast Asia; however, we faced challenges in sourcing sufficient cultural concepts through data crawling and in locating adequately qualified human annotators from specific nations, including Timor-Leste, Brunei, and Laos.

Long-tailed Distribution. The aforementioned issues related to the availability of qualified human annotators, along with difficulties in acquiring high-quality images, have resulted in a naturally occurring long-tailed distribution. Additionally, our methodology includes a process for filtering out less suitable crawled images.

Ethical Consideration

Cultural concepts overlap across cultures. Certain cultural artifacts are commonly found in multiple countries, albeit with nuanced differences, characterized by the use of either identical or distinct cultural concept terminology. To mitigate potential conflicts, we implemented an "avoid list" during the selection of visual options for the question types. This initial measure effectively decreased the total number of questions from over one thousand to more than 800 for both Type 1 and Type 2 questions; however, it also contributed to the overall stability of our research framework.

Annotators. We recruited annotators through Upwork, a global freelancing platform, following specific criteria. Firstly, participants were required to be natives of Southeast Asian countries, possessing a comprehensive understanding of the local culture, traditions, and customs. Secondly, they needed to have a basic proficiency in using computers or mobile devices, as they were expected to utilize specialized software for image labeling. We employed purposive sampling to identify freelancers on Upwork.com who fulfilled these inclusion criteria, focusing on their cultural expertise and experience with cultural content or research. Additionally, potential participants were evaluated based on their profiles, work history, reviews, and portfolio samples, prioritizing those with a strong grasp of local culture and relevant project experience. This methodology ensures that selected participants not only possess knowledge of their cultural background but also have the necessary skills to utilize the required tools and adhere to the research protocols. For our study, we engaged three annotators each for Philippines and Myanmar, and two annotators for the remaining countries. Participants were compensated monetarily at a rate of $5-10 per hour for their involvement in the research, with specific compensation structured at $5 for every 50 images labeled accurately.

Privacy Rights. We ensure that the intellectual property and privacy rights of the images collected are respected. We claim that the collected data will not be used commercially.

References

Yujin Baek, ChaeHun Park, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2024. Evaluating visual and cultural interpretation: The K-Viscuit benchmark with human-VLM collaboration. *Preprint*, arXiv:2406.16469.

Longju Bai, Angana Borah, Oana Ignat, and Rada Mihalcea. 2025. The power of many: Multi-agent multimodal models for cultural image captioning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2970–2993, Albuquerque, New Mexico. Association for Computational Linguistics.

Parth Bhalerao, Mounika Yalamarty, Brian Trinh, and Oana Ignat. 2025. Multi-agent multimodal models for multicultural text to image generation. *Preprint*, arXiv:2502.15972.

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, EunJeong Hwang, and Vered Shwartz. 2024. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6763–6782, Miami, Florida, USA. Association for Computational Linguistics.

Olena Burda-Lassen, Aman Chadha, Shashank Goswami, and Vinija Jain. 2025. How culturally aware are vision-language models? In *2025 IEEE 6th International Conference on Image Processing, Applications and Systems (IPAS)*, volume CFP2540Z-ART, pages 1–6.

Samuel Cahyawijaya, Holy Lovenia, Joel Ruben Antony Moniz, Tack Hwa Wong, Mohammad Rifqi Farhansyah, Thant Thiri Maung, Frederikus Hudi, David Anugraha, Muhammad Ravi Shulthan Habibi, Muhammad Reza Qorib, Amit Agarwal, Joseph Marvin Imperial, Hitesh Laxmichand Patel, Vicky Feliren, Bahrul Ilmi Nasution, Manuel Antonio Rufino, Genta Indra Winata, Rian Adam Rajagede, Carlos Rafael Catalan, and 73 others. 2025. Crowdsource, crawl, or generate? creating SEA-VL, a multicultural vision-language dataset for southeast asia. *Preprint*, arXiv:2503.07920.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The Llama 3 herd of models. *CoRR*, abs/2407.21783.

Xingyu Fu, Sheng Zhang, Gukyeong Kwon, Pramuditha Perera, Henghui Zhu, Yuhao Zhang, Alexander Hanbo Li, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Patrick Ng, Dan Roth, and Bing Xiang. 2023. Generate then select: Open-ended visual question answering guided by world knowledge. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2333–2346, Toronto, Canada. Association for Computational Linguistics.

Hugo Laurençon, Leo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*.

Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024. FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jin Liu, ChongFeng Fan, Fengyu Zhou, and Huijuan Xu. 2023. Be flexible! learn to debias by sampling and prompting for robust visual question answering. *Information Processing & Management*, 60(3):103296.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. CultureVLM: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv preprint arXiv:2501.01282*.

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, and 57 others. 2024. CVQA: Culturally-diverse multilingual visual question answering benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

Malvina Nikandrou, Georgios Pantazopoulos, Nikolas Vitsakis, Ioannis Konstas, and Alessandro Suglia. 2025. CROPE: Evaluating in-context adaptation of vision and language models to culture-specific concepts. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7917–7936, Albuquerque, New Mexico. Association for Computational Linguistics.

OpenAI. 2014. Chatgpt.

Florian Schneider, Carolin Holtermann, Chris Biemann, and Anne Lauscher. 2025. GIMMICK – globally inclusive multimodal multitask cultural knowledge benchmarking. *Preprint*, arXiv:2502.13766.

Piotr Skalski. 2019. Make Sense. `https://github.com/SkalskiP/make-sense/`.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. 2024. MTVQA: Benchmarking multilingual text-centric visual question answering. *Preprint*, arXiv:2405.11985.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Yuxuan Wang, Yijun Liu, Fei Yu, Chen Huang, Kexin Li, Zhiguo Wan, Wanxiang Che, and Hongyang Chen. 2025. CVLUE: A new benchmark dataset for chinese vision-language understanding evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8):8196–8204.

Zhecan Wang, Long Chen, Haoxuan You, Keyang Xu, Yicheng He, Wenhao Li, Noel Codella, Kai-Wei Chang, and Shih-Fu Chang. 2023. Dataset bias mitigation in multiple-choice visual question answering and beyond. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8598–8617, Singapore. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*.

Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the vision: Geo-diverse visual commonsense reasoning. In *EMNLP*.

Xi Zhang, Feifei Zhang, and Changsheng Xu. 2023. Reducing vision-answer biases for multiple-choice VQA. *IEEE Transactions on Image Processing*, 32:4621–4634.

Xi Zhang, Feifei Zhang, and Changsheng Xu. 2024. NExT-OOD: Overcoming dual multiple-choice VQA biases. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(04):1913–1931.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded question answering in images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004.

A Appendix

A.1 More Quantitative Results

Table 3 and Table 4 display the full details for the overall results for *country* and *category*. We see that the closed-source VLMs generally show higher accuracy performance, while the open-source ones reach higher mIoU results.

A.2 Seeing Culture Benchmark

A.2.1 Concepts

Figure 7 present all the concepts addressed in SCB. We share some examples in Figure 8.

A.2.2 Eliminated images and questions

In accordance with the details outlined in Section 3, we exclude certain images from consideration. Specifically, as shown in Figure 9, we remove the image on the left as its focus is solely on the target cultural artifact. The image on the right is also omitted due to the lack of a distracting object, although it contains a more complex scene than the left-hand image.

Certain questions are excluded due to their generic nature, potential overlap with other cultural artifacts, or lack of necessity for critical reasoning. For example, we dismissed the question concerning *Indonesia/game/kelereng*: "Which object in the image symbolizes childhood nostalgia, often played in schoolyards and neighborhoods in Indonesia?" because numerous games evoke similar childhood memories. Similarly, we rejected the question for *Myanmar/music/saung*: "Which Burmese object in the image has a hollow body made of wood, designed to enhance the richness of its sound?" as it merely describes the cultural artifact without engaging in reasoning or referencing a symbol.

A.2.3 Multiple-choice VQA Generation Algorithm

Algorithms 1, 2, and 3 explain how we choose visual options for each type. Additionally, we provide clarifications for the abbreviations utilized within the algorithms.

- $\mathcal{D}$: Dataset
- $\mathcal{V}$: Vectorstore index
- k: Number of similar items to retrieve
- $N_{\max}$: Maximum number of questions per name

Model	Malaysia		Philippines		Cambodia		Indonesia		Myanmar		Vietnam		Thailand	
	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU
InstructBLIP	3.85	–	8.96	–	4.55	–	10.6	–	12.97	–	10.19	–	13.37	–
Idefics2	9.89	0.03	2.83	–	9.09	0.22	12.42	0.11	15.20	0.19	14.51	0.19	10.17	0
Llama 3.2	26.37	–	21.23	–	13.64	–	24.40	–	22.73	–	25.93	–	26.45	–
LLaVA-Onevision	30.22	–	28.77	–	13.64	–	23.89	–	23.15	–	23.46	–	27.62	–
MiniCPM 2.6	44.51	–	22.17	–	18.18	–	35.08	–	25.38	–	28.09	–	38.66	–
InternVL2.5	32.97	33.35	29.25	32.47	22.73	22.27	30.79	29.42	32.78	26.61	30.86	32.19	31.98	21.57
Qwen2.5-VL	54.40	**56.89**	51.89	52.78	27.27	**60.50**	52.36	**46.47**	52.72	**48.55**	59.57	**45.56**	58.72	**40.65**
GPT-4.1	84.07	14.64	69.81	**100.00**	16.79	18.19	85.19	13.82	77.27	16.64	86.73	7.33	79.65	11.60
Gemini 2.5 Pro	85.71	17.10	68.87	15.40	90.91	13.36	85.48	15.76	79.08	20.21	88.27	12.53	**81.98**	13.98
GPT-o3	**86.26**	41.74	**78.77**	35.29	**100.00**	30.23	**86.93**	32.77	**82.85**	31.51	**89.81**	28.13	80.81	24.31

Table 3: The overall VLMs' performance presented by countries. "Acc" refers to *accuracy* while "mIoU" denotes *mean IoU*.

Model	Wedding		Dance		Music		Celebration		Game	
	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU
InstructBLIP	12.09	–	18.42	–	5.03	–	18.75	–	10.76	–
Idefics2	7.85	0.14	16.01	0.06	12.37	–	17.19	0.25	14.49	0.22
Llama 3.2	25.45	–	25.44	–	22.01	–	32.03	–	23.39	–
LLaVA	26.19	–	24.78	–	23.79	–	24.22	–	23.96	–
MiniCPM	33.40	–	36.40	–	35.43	–	21.88	–	24.96	–
InternVL	32.03	25.47	33.99	37.06	26.83	36.48	26.56	15.36	35.72	20.68
Qwen	54.08	**40.91**	57.02	**48.97**	49.37	**54.43**	42.19	**31.63**	59.40	**47.63**
GPT-4.1	81.12	13.29	85.53	15.54	81.76	14.53	62.50	15.36	84.65	11.87
Gemini	79.96	13.72	86.62	20.17	83.96	19.30	60.16	18.26	87.09	12.41
GPT-o3	**82.18**	28.33	**90.13**	39.51	**86.37**	37.06	**63.28**	21.68	**88.24**	25.23

Table 4: The overall VLMs' performance presented by categories. "Acc" refers to *accuracy* while "mIoU" denotes *mean IoU*.

- $U_{\max}$: Maximum allowed usage per choice
- $\mathcal{B}$: Set of banned IDs due to usage limit
- $\mathcal{Q}$: Output set of generated multiple-choice VQAs
- $\mathcal{C}$: Set of already-used choice combinations (as hashable sets)

A.2.4 Avoid list

The comprehensive *avoid list* is presented in Figure 10. This list has been meticulously compiled based on the insights provided by annotators to prevent overlap between countries.

Country	Cultural Concepts
Indonesia	jaipongan (sinden, klono sewandono, jathil, bujan gagong, warok), balinese legong (condong, princess rangkesari), barong dance (barong, rangda, monkey, airlangga's soldiers), engklek, gasing, permainan congklak, lompat tali, kuda lumping, engrang, kelereng, dodot javanese, siraman, janur, keris, siger, selendang, uang panai, paes ageng, tasbih, bunga nikah_wedding flowers, tumpeng, pakaian batik adat, lukisan penganting, gong, rebab, talempong, suling, gambang, gamelan, keroncong, angklung, kecapi, sape
Myanmar	taungbyon nat, သီတင်းကျွတ်မီးထွန်းပွဲတော်_thadingyut festival, နတ်ပွဲတော်ပွဲ_nat pwe, သဲပုံသဲပုံစေတီပွဲ_sand pagodas festival, bo tree watering, naga new year festival, သင်္ကြန်ကြင်ပွဲနှင့်ပွဲ_thingyan festival, shwedagon pagoda festival, regatta festival, ကဆုန်လနှင့်ပြည့်ဗုဒ္ဓနေ့_buddha jayanti, လှေပြိုင်ပွဲပြိုင်ပွဲ_boat racing, ချောတိုင်တိုတက်_greasy pole climbing, နံခေါင်းအုံးရိုက်_gaung ohn yite, ထုပ်ထုပ်စည်းတိုး_htoke si toe, lethwei, လွန်ဆွဲနံဆွဲပွဲ_tug-of-war, ပိုက်ကျော်ခြင်း_sepak takraw, ခြင်းလုံး_chinlone, မင်္ဂလာဝတ်စုံ_wedding dress, thanaka, ဇော်ဂျီအက_zawgyi dance, shan peacock dance, kyaukse dance, ကဗျာလွတ်လွအကတ်_ka-byar-lut, ဆီမီးဆီမီးကွက်ကွအကက်_oil lamp dance, တပင်တိုင်င်တိုအကင်_solo dance, ဇာတ်ပွဲ_zat pwe, ယိမ်း_yein, စည်းနှင့်ဝါးနှင့်_si and wa, စောင်း_saung, hsaing waing, နဲပတ်ဝိုတ်ဝိုင်းဝိုင်း_pat waing, ပလွေ_flute, နှဲ_hne
Thailand	ขันหมาก_engagement offering tray set, ชุดเจ้าบ่าว_groom's attire, สินสอด_dowry, พานทองเหลือง ใหญ่_large brass tray, ถุงเงิน_silver bag, หมาก_betel nut, ถุงทอง_gold bag, มงคลแฝด_double auspicious headband, การนับสินสอดทองหมั้น_counting dowry and engagement gold, พิธีรดน้ำสังข์_conch water pouring ceremony, เทศกาลสงกรานต์_songkran, มาฆบูชา_makha bucha, วันรัฐธรรมนูญแห่งประเทศไทย_constitution day, วันวิสาขบูชา_visakha bucha, วันออกพรรษา_ok phansa, วันอาสาฬหบูชา_asarnha bucha, โขน_khon, กลิ้งครกขึ้นภูเขา_rolling mortar uphill, หมากข่าง_spinning top, ชักเย่อ_tug of war, อังกะลุง_angklung, ระนาด_ranat
Vietnam	nhảy lò cò_hopscotch, lễ cưới truyền thống việt nam_traditional vietnamese wedding, tết đoan ngọ_duanwu festival, lễ hội đèn lồng hội an_hoi an lantern festival, múa cồng chiêng_gong dance, múa sạp_cheraw dance, bầu cua_gourd crab fish tiger, đàn tranh_vietnamese 16-string zither, đá cầu_shuttlecock kicking, sáo_flute, đàn bầu_vietnamese one-string zither, đàn đáy_vietnamese lute
Philippines	yugal, maria clara dress, candle lighting, arras, bouquet, barong tagalog, belo_veil, banduria_bandurria, kudyapi, kulintang
Malaysia	suling_bamboo flute, gambus, kompang, rebab, sape, baju kurung
Cambodia	chhing, khaen

Figure 7: Compilation of cultural concepts addressed in SCB.

Question: Which image shows the celebration artifact that is associated with how the Burmese clear their debt before the beginning of the new year?
Answer: (b)

(a) (b) (c) (d)

Question: Which image indicates the traditional Indonesian wedding artifact that is associated with the symbol of a new life beginning permitted by ancestors?
Answer: (c)

(a) (b) (c) (d)

Figure 8: Multiple-choice VQA examples from our SCB dataset. The red masks in the correct option demonstrate the supporting evidence.

Figure 9: Examples from the two images that we eliminated.

Category	Indonesia	Thailand	Philippines	Vietnam	Myanmar	Malaysia	Cambodia
Music	sape	-	kudyapi	-	-	sape	-
Music	angklung	อังกะลุง (angklung)	-	-	-	-	-
Music	suling (bamboo flute)	-	-	sáo (flute)	ပလွေ (flute)	-	-
Music	gambang	ระนาด (ranad)	-	-	-	-	-
Music	kecapi	-	-	đàn tranh (vietnamese 16-string zither)	-	-	-
Music	rebab	-	-	-	-	rebab	-
Music	talempong	-	kulintang	-	-	-	-
Game	engklek	-	-	nhảy lò cò	-	-	-
Game	-	ชักเย่อ (tug-of-war)	-	-	လွန်ဆွဲ န် ဆွဲပွဲ (tug-of-war)	-	-
Game	permainan congklak	หมากข่าง (spinning top)	-	-	-	-	-
Wedding	uang panai	สินสอด (dowry)	-	-	-	-	-
Wedding	bunga nikah	-	bouquet	-	-	-	-
Wedding	selendang	-	belo	-	-	-	-
Celebrations	သင်္ကြန် င်္ကြ ပွဲ န် ပွဲ (thingyan festival)	เทศกาลสงกรานต์ (songkran festival)	-	-	-	-	-

Figure 10: The avoid list for organizing visual options within the various question types indicates that cultural artifacts positioned within the same row are not included in the sampling process for visual options. For example, if we assume that the correct answer is an image from Indonesia/music/sape in the context of the VQA framework during the initial phase, then images associated with Malaysia/music/sape and Philippines/music/kudyapi are systematically excluded from consideration.

Algorithm 1 Type 1 ($\mathcal{D}, \mathcal{V}, N_{\max}, U_{\max}, k$)

Initialize usage counter $\mu : \mathbb{Z} \to \mathbb{N}$ for all IDs
Initialize $\mathcal{Q} \leftarrow \emptyset$, $\mathcal{B} \leftarrow \emptyset$, $\mathcal{C} \leftarrow \emptyset$
for each unique name n in $\mathcal{D}$ **do**
    Extract Country(n), Category(n)
    Let $\mathcal{D}_n \subset \mathcal{D}$ be the $N_{\max}$ samples with name n
    for each sample $q \in \mathcal{D}_n$ **do**
        Use $\mathcal{V}$ to retrieve top-k similar items $\mathcal{S}$ where Country(s) = Country(n), Category(s) = Category(n), Name$(s) \neq n$, and ID$(s) \notin \mathcal{B}$
        for each triple $(s_1, s_2, s_3) \subset \mathcal{S}$ **do**
            if each s_i has $\mu(s_i) < U_{\max}$ and $\{\text{ID}(s_i)\} \notin \mathcal{C}$ **then**
                Form choice set $\mathcal{A} = \{s_1, s_2, s_3, q\}$ with q as the correct answer
                Add ID($\mathcal{A}$) to $\mathcal{C}$, update μ
                Add $\mathcal{A}$ to $\mathcal{Q}$
                break
            end if
        end for
        if no valid triple found **then**
            Sample 3 random distractors $\mathcal{R}$ satisfying above constraints
            if $|\mathcal{R}| = 3$ **then**
                Form choice set $\mathcal{A} = \mathcal{R} \cup \{q\}$ and update μ, $\mathcal{C}$
                Add $\mathcal{A}$ to $\mathcal{Q}$
            end if
        end if
    end for
end for
return $\mathcal{Q}$

Algorithm 2 Type 2 multiple-choice questions

Initialize usage counter μ, banned ID set $\mathcal{B}$, choice hash set $\mathcal{C}$, and output $\mathcal{Q}$
for each unique name n in $\mathcal{D}$ **do**
    Extract Country(n), Category(n)
    Let $\mathcal{D}_n \subset \mathcal{D}$ be up to $N_{\max}$ rows with name n
    for each sample $q \in \mathcal{D}_n$ **do**
        Use $\mathcal{V}$ to retrieve $\mathcal{S}$ where Country$(s) \neq$ Country(n), Category(s) = Category(n), and ID$(s) \notin \mathcal{B}$
        for triplets (s_1, s_2, s_3) with distinct countries **do**
            if all $\mu(s_i) < U_{\max}$ and $\{\text{ID}(s_i)\} \notin \mathcal{C}$ **then**
                Form $\mathcal{A} = \{s_1, s_2, s_3, q\}$ with q correct
                Update μ, $\mathcal{C}$, add $\mathcal{A}$ to $\mathcal{Q}$
                break
            end if
        end for
        if no valid triplet found **then**
            Sample $\mathcal{R}$ from $\mathcal{D}$ such that country of r is not equal to country of n, category of r is equal to category of n, and name of r is not equal to n
            if $|\mathcal{R}| = 3$ **then**
                Form $\mathcal{A} = \mathcal{R} \cup \{q\}$ and update μ, $\mathcal{C}$
                Add $\mathcal{A}$ to $\mathcal{Q}$
            end if
        end if
    end for
end for
return $\mathcal{Q}$

Algorithm 3 Type 3 multiple-choice questions

Initialize choice usage μ, seen choice sets $\mathcal{C}$, output set $\mathcal{Q}$
Let $\mathcal{O}$ be original choice sets from $\mathcal{D}$ (to avoid duplicates)
$\mathcal{C} \leftarrow \mathcal{O}$
for each question $q \in \mathcal{D}$ **do**
  Set $used\_choices \leftarrow \emptyset$
  for $e = 1$ to E_{max} **do**
    Extract correct answer a^* with its country, category, and name
    Extract top distractor a' from q (highest score $\neq -1.0$)
    Collect banned triples from a^* and a': country, category, and name
    Initialize $choices \leftarrow \{a^*, a'\}$, and record used countries and names
    Let $\mathcal{P} \leftarrow$ opposite type pool (type1 if q is type2, else type2)
    Filter $\mathcal{P}$ to get eligible distractors satisfying: same category as q, distinct country and name, not in banned triples, usage $\mu < U_{\text{max}}$, and not in $used\_choices$
    if at least 2 eligible distractors found **then**
      Sample 2 distractors d_1, d_2 and add to $choices$
      Update μ and $used\_choices$
      Shuffle $choices$ and assign to q_e
      Set correct choice score to -1.0, others to -2.0
      Mark q_e.type $\leftarrow$ mixed, and update $\mathcal{C}$
      if $choices \notin \mathcal{C}$ **then**
        Add q_e to $\mathcal{Q}$ and to $\mathcal{C}$
      end if
    end if
  end for
end for
return $\mathcal{Q}$