Evaluating Visual and Cultural Interpretation: The K-Viscuit Benchmark with Human-VLM Collaboration

Anonymous ACL submission

Abstract

To create culturally inclusive vision-language models (VLMs), developing a benchmark that tests their ability to address culturally relevant questions is essential. Existing approaches typically rely on human annotators, making the process labor-intensive and creating a cognitive burden in generating diverse questions. To address this, we propose a semi-automated framework for constructing cultural VLM benchmarks, specifically targeting multiple-choice QA. This framework combines human-VLM collaboration, where VLMs generate questions based on guidelines, a small set of annotated examples, and relevant knowledge, followed by a verification process by native speakers. We demonstrate the effectiveness of this framework through the creation of K-Viscuit, a dataset focused on Korean culture. Our experiments on this dataset reveal that open-source models lag behind proprietary ones in understanding Korean culture, highlighting key areas for improvement. We also present a series of further analyses, including human evaluation, augmenting VLMs with external knowledge, and the evaluation beyond multiple-choice QA.[1]

1 Introduction

Recent advances in vision-language models (VLMs) have demonstrated remarkable capabilities in tasks ranging from image captioning to visual question answering. However, these models are predominantly trained on Western-centric datasets (Lin et al., 2014; Young et al., 2014; Antol et al., 2015), leading to significant performance disparities when applied to non-Western contexts (Liu et al., 2021; Yin et al., 2021, 2023; Romero et al., 2024). This cultural bias is particularly problematic as visual interpretation often depends heavily on cultural context, necessitating the development of more culturally aware VLMs.

[1] Our dataset and code will be publicly available.

Several benchmarks have been proposed to evaluate cultural understanding in VLMs (Liu et al., 2021; Yin et al., 2021; Romero et al., 2024; Nayak et al., 2024). These approaches primarily rely on manual question generation, which, while valuable, faces certain practical challenges. The manual process can be time-consuming and resource-intensive when scaling to new cultural contexts. Additionally, as noted in cognitive science research (Ramos, 2020), human annotators may experience cognitive fixation, potentially limiting the diversity of generated questions. These practical considerations motivate the need for more efficient benchmark construction approaches.

Inspired by recent successes in human-LLM collaborative data generation (Liu et al., 2022; Bartolo et al., 2022; Kamalloo et al., 2023), we propose a semi-automated framework for constructing cultural VLM benchmarks that enhances both the efficiency and diversity of culture-relevant visual question and answer generation, as shown in Fig. 1. Our framework incorporates human-VLM collaboration, where the VLM generates and recommends questions and answers based on carefully crafted guidelines, a small set of human-annotated examples, and image-specific knowledge. Native speakers then verify these recommended questions to ensure quality and cultural relevance.

Using this framework, we develop K-Viscuit (Korean Visual and Cultural Interpretation Test), a benchmark dataset for Korean culture that can be adapted for other cultural contexts. K-Viscuit features two distinct evaluation types: visual recognition and visual reasoning. Also, the benchmark employs carefully designed multiple-choice questions with highly similar distractors to prevent models from exploiting superficial patterns.

Our evaluation with K-Viscuit reveals a significant performance gap between open-source and proprietary VLMs in understanding Korean culture. We provide insights into the current limitations

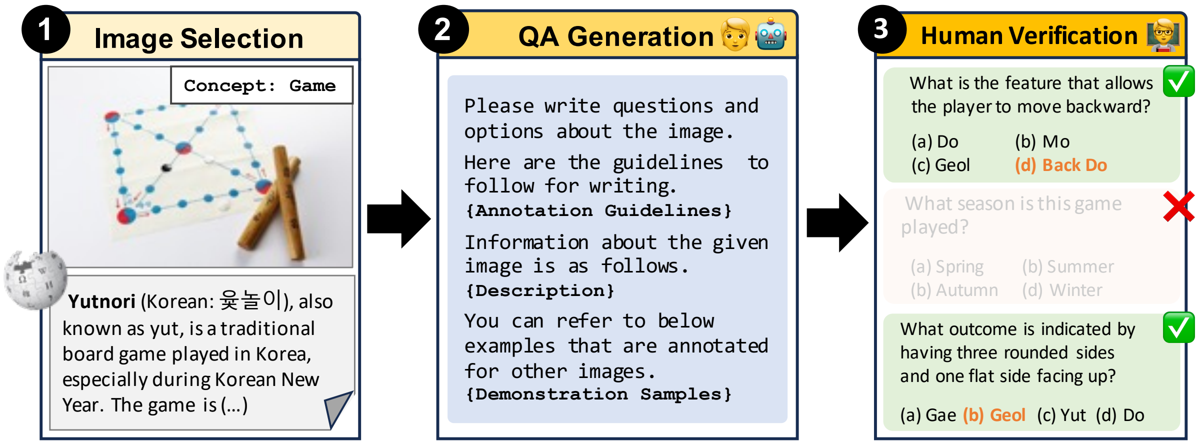

Figure 1: Framework Overview.

and potential improvements in VLMs' cultural understanding capabilities through detailed analyses, which include human evaluation, external knowledge integration, and extended evaluation beyond a multi-choice question-answering setup.

Our contributions are summarized as follows:

- We propose a semi-automated framework for constructing benchmarks to evaluate the cultural understanding capabilities of VLMs.
- We develop `K-Viscuit`, a Korean culture-focused VQA benchmark using our proposed framework.
- We present comprehensive experimental results and analyses of both open-source and proprietary VLMs evaluated on `K-Viscuit`.

2 Related Work

Recent research has made significant strides in developing benchmarks to assess AI models' cultural understanding capabilities. These efforts are particularly important as many existing models, including VLMs and Large Language Models (LLMs), are trained predominantly on Western-centric datasets (Young et al., 2014; Lin et al., 2014; Antol et al., 2015), limiting their effectiveness in non-Western contexts. For LLMs, several notable cultural benchmarks have emerged: Kim et al. (2024) introduced `CLIcK`, a benchmark for evaluating Korean language models' cultural knowledge through carefully designed QA pairs, while Wibowo et al. (2023) developed `COPAL-ID` to capture Indonesian cultural nuances in text-based commonsense reasoning.

In the multimodal domain, VLMs require consideration of both visual and textual inputs to effectively reflect cultural contexts. Liu et al. (2021) addressed this challenge with `MaRVL`, a multilingual visually-grounded reasoning dataset spanning five languages and cultures, demonstrating the importance of cross-cultural visual-linguistic understanding. Building on this foundation, Yin et al. (2021) introduced `GD-VCR` to evaluate geographical and cultural aspects of visual commonsense reasoning, while Romero et al. (2024) developed `CVQA`, a comprehensive multilingual VQA benchmark for assessing VLMs across diverse cultural contexts.

While these cultural benchmarks have provided valuable insights, their reliance on manual annotation can constrain both the diversity and efficiency of dataset creation (Liu et al., 2021; Yin et al., 2021; Ramaswamy et al., 2024; Romero et al., 2024). Recent work has demonstrated the potential of AI-assisted dataset generation when combined with human expertise. Liu et al. (2022) successfully applied this approach to natural language inference tasks, and similar strategies have been widely adopted in creating language-only datasets (Taori et al., 2023; Kim et al., 2023; Dubois et al., 2024; Kim et al., 2024). Our work extends this paradigm to multimodal cultural benchmarks, leveraging AI models to enhance dataset diversity while maintaining high quality through human verification.

3 Data Construction Framework

We present our human-AI collaborative framework for constructing datasets to evaluate VLMs' understanding of specific cultural domains. In this work, we focus on Korean culture as our target domain. First, we provide an overview (§3.1) and imple-

Figure 2: **Dataset Examples**. We present an image and two questions of different types for each concept category.

mentation details. Then, we present the analysis of our resulting dataset, `K-Viscuit` (**K**orean **Vis**ual and **Cu**ltural **I**nterpretation **T**est) (§3.2).

3.1 Framework Overview

Our framework is designed to create a multiple-choice visual question answering (VQA) task, where each evaluation sample consists of an image, a question, and four options with one correct answer. Native Korean speakers participate in the dataset construction process, while a powerful proprietary VLM is employed to mitigate unintended human biases, such as cognitive fixation (Ramos, 2020), and streamline the annotation process. The generated samples cover various aspects of the target culture derived from daily life and require multimodal reasoning to interpret both visual and textual information accurately. The dataset construction consists of four stages: 1) concept selection, 2) image selection, 3) question and options annotation, and 4) human verification. Fig. 1 illustrates the overall framework.

3.1.1 Concept Categorization

Inspired by recent studies on multicultural evaluation datasets (Liu et al., 2021; Wibowo et al., 2023; Kim et al., 2024), we aim to assess knowledge of various concepts encountered in daily life by Korean natives. While each concept should have some degree of universality, its manifestation often varies across cultures. Following Liu et al. (2021), we reference semantic concepts from the Intercontinental Dictionary Series (IDS) (Key and Comrie, 2015) to define our concept list. Our dataset encompasses ten core concepts: FOOD, BEVERAGE, GAME, CELEBRATIONS, RELIGION, TOOL, CLOTHES, HERITAGE, ARCHITECTURE, and AGRICULTURE.

3.1.2 Image Selection

Korean native annotators collected web images corresponding to the selected concepts. To ensure diverse representation, we limited each specific object to appearing no more than twice within any single category. Following Liu et al. (2021), we selected only images depicting concepts that could physically exist in everyday life. Annotators were encouraged to source diverse and suitable images from various web resources. Wikimedia Commons[2] served as the primary source, and only CC-licensed images were selected.

3.1.3 Question Generation

Question Type Based on the selected images, we annotate questions in multiple-choice QA format. To comprehensively evaluate understanding of Korean culture, we categorize questions into two types: visual recognition (TYPE 1) and reasoning (TYPE 2). Visual recognition questions assess basic visual information such as object identification, while reasoning questions require fine-grained cultural knowledge or deeper reasoning processes re-

[2] https://commons.wikimedia.org/wiki

# of samples	657
- Type 1/ Type 2	237/420
# of unique images	237
# of options	2628
# of unique options	2129
Avg. question length	13.5
- Type 1/ Type 2	10.1/15.5
Avg. option length	1.7

Table 1: **Dataset statistics**. The length of questions and options denotes the number of words.

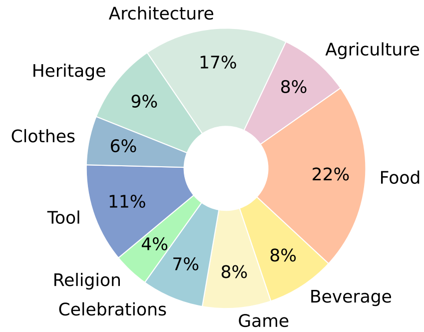

Figure 3: Concept distribution of our dataset.

lated to the image. For each image, we created one Type 1 question and between one to four Type 2 questions. This categorization offers two key advantages: First, Type 1 questions enable assessment of a model's basic visual understanding of culturally embedded concepts. Second, Type 2 questions comprehensively evaluate cultural understanding beyond simple object recognition.

AI-assisted Question Annotation We create questions and their options (one correct answer and three distractors) by leveraging a powerful proprietary VLM (GPT-4-Turbo). For each concept category, human annotators first create exemplar questions and options for at least three images. These manually annotated examples serve as demonstrations for the VLM to generate additional questions and options.

Specifically, the VLM receives: 1) the target image, 2) human-annotated demonstration examples, 3) detailed annotation guidelines, and 4) image-specific knowledge descriptions. We include relevant contextual knowledge for each image to enhance question diversity and relevance, ensuring VLM-generated questions are grounded in real-world understanding. Notably, following Wang et al. (2023), our guidelines emphasize maintaining high similarity among all four multiple-choice options, a principle also reflected in human-annotated examples. All information is provided to the VLM through natural language prompts, with distinct annotation guidelines for visual recognition and reasoning questions. The detailed prompts are presented in Appendix A.

It should be noted that all text in the dataset is written in English to isolate the evaluation of multicultural comprehension from multilingual aspects. However, as culture and language are not entirely orthogonal (SUSAN, 1996; Kramsch, 2014), completely separating them can be challenging. For instance, we observed that certain Korean terms lack exact English equivalents. In such cases, we romanize these Korean terms following standard transliteration rules.

3.1.4 Human Verification

Our preliminary studies showed that while the VLM often produced factually correct samples, some did not fully align with the cultural nuances we sought to capture. Rather than discarding only incorrect content, our human verification process prioritizes selecting question-option sets that best reflect intended cultural subtleties. Although this leads to setting aside numerous plausible samples, it does not imply unreliability; instead, it refines the dataset to ensure deep cultural resonance. Approved samples require minimal revision before inclusion. This approach emphasizes nuanced cultural alignment and indicates the need for further work to better synchronize VLM outputs with human cultural intentions.

3.2 `K-Viscuit`

Statistics Table 1 presents detailed statistics of our benchmark dataset. Our dataset comprises 657 total examples (237 Type 1 and 420 Type 2 questions) based on 237 unique images across 10 concept categories. The average word counts are 10.11 and 15.46 for Type 1 and Type 2 questions respectively, with an overall average of 13.53 words. Each question includes four options, totaling 2,628 options with an average length of 1.74 words. Fig. 3 shows the distribution of concept categories in our dataset. We also compare our dataset size with CVQA (Romero et al., 2024), a recently proposed cultural VQA benchmark, in Appendix E.

Required Knowledge to Solve Questions To characterize our dataset, we analyze the types of knowledge required to solve questions. Following Tong et al. (2024), we used `GPT-4` to analyze Type 2 questions. We provided all Type 2 QA pairs and summarized the required knowledge for each

Model	All	Food	Beverage	Game	Celeb.	Religion	Tool	Clothes	Heritage	Arch.	Agri.
InstructBLIP-7B (Dai et al., 2024)	50.84	40.85	42.31	38.46	53.19	40.74	50.67	62.16	51.61	60.55	72.22
instructBLIP-13B (Dai et al., 2024)	55.56	45.77	50.00	46.15	59.57	55.56	54.67	64.86	66.13	60.55	64.81
mPLUG-Owl2-7B (Ye et al., 2023)	48.25	42.25	42.31	30.77	63.83	55.56	48.00	54.05	45.16	49.54	66.67
LLaVA-1.6-7B (Liu et al., 2024)	56.32	43.66	48.08	40.38	57.45	51.85	54.67	67.57	59.68	72.48	72.22
LLaVA-1.6-13B (Liu et al., 2024)	57.08	45.07	53.85	36.54	68.09	40.74	53.33	70.27	66.13	69.72	70.37
InternLM-XC2-7B (Dong et al., 2024)	59.67	50.70	48.08	40.38	65.96	55.56	58.67	64.86	69.35	69.72	75.93
Molmo-7B-D (Deitke et al., 2024)	61.04	58.45	71.15	44.23	68.09	44.44	61.33	70.27	56.45	64.22	68.52
Idefics2-8B (Laurençon et al., 2024)	63.62	51.41	50.00	50.00	74.47	66.67	69.33	75.68	74.19	73.39	62.96
Llama-3.2-11B (Dubey et al., 2024)	68.04	61.27	65.38	50.00	72.34	70.37	72.00	75.68	72.58	69.72	81.48
Claude-3-opus (Anthropic, 2024)	70.02	62.68	73.08	59.62	72.34	77.78	74.67	78.38	75.81	67.89	75.93
GPT-4-Turbo (Achiam et al., 2023)	80.82	73.94	80.77	78.85	85.11	**85.19**	81.33	86.49	85.48	79.82	**87.04**
Gemini-1.5-Pro (Reid et al., 2024)	81.58	80.28	78.85	71.15	85.11	77.78	82.67	83.78	83.87	84.40	85.19
GPT-4o (OpenAI, 2024)	**89.50**	**88.73**	**82.69**	**86.54**	**95.74**	**85.19**	**90.67**	**91.89**	**91.94**	**91.74**	**87.04**

Table 2: **VLMs Evaluation Results.** *Celeb.*, *Arch.*, and *Agri.* denote *Celebration*, *Architecture*, and *Agriculture*. The highest and the second highest scores in each column are highlighted in bold and underlined.

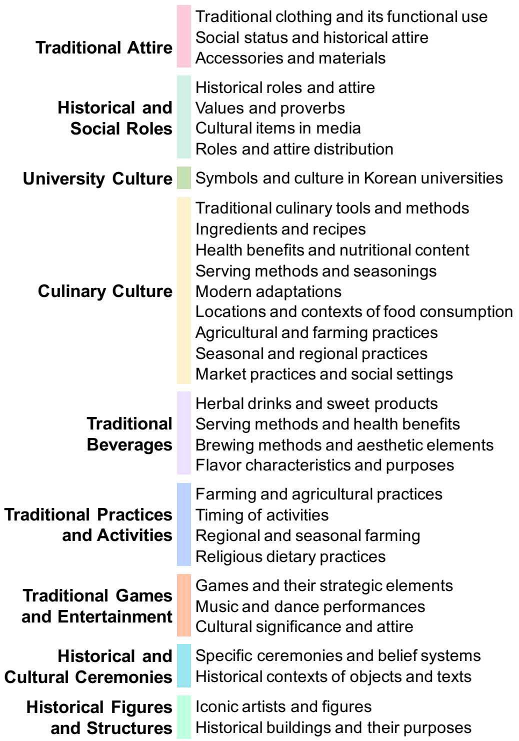

Figure 4: Required Cultural Knowledge.

question. As shown in Fig. 4, the analysis reveals diverse cultural elements. Detailed categorization instructions are provided in Appendix B.1.

Qualitative Examples Fig. 2 showcases sample images, questions, and options along with their concept categories and question types.

4 Experiments

We conduct experiments to evaluate various VLMs on our constructed dataset.

4.1 Models

The following open-source VLMs are used for experiments: InstructBLIP-7B/13B (Dai et al., 2024), LLaVA-v1.6-7B/13B (Liu et al., 2024), mPLUG-Owl2-7B (Ye et al., 2023), InternLM-XComposer2-VL-7B (Dong et al., 2024), Molmo-7B-D (Deitke et al., 2024), Idefics2-8B (Laurençon et al., 2024), and Llama-3.2-11B-Vision-Instruct (Dubey et al., 2024). We also use the following proprietary models: Claude-3-opus (Anthropic, 2024), GPT-4-Turbo (Achiam et al., 2023), Gemini-1.5-Pro (Reid et al., 2024), and GPT-4o (OpenAI, 2024). All models are evaluated in the conventional multiple-choice setup, where a model is prompted to choose its answer from one of four options. The input text is constructed by concatenating (1) a question, (2) each option with option letters in alphabetical order, and (3) the instruction about output format (i.e., *"Answer with the option's letter from the given choices directly."*). We use accuracy as an evaluation metric.

4.2 Results

Table 2 demonstrates the evaluation results of different VLMs on our dataset. We first observe that proprietary models usually show higher accuracies. The GPT-4o and Gemini-1.5-Pro achieve the highest and second-highest scores, respectively. Among the open-sourced models, Llama-3.2-11B and Idefics2-8B usually perform better than other models. The accuracy of models in different question types is shown in Table 3. We observe that most models show higher accuracy in TYPE 2 questions compared to TYPE 1 questions. Regarding these trends, we suspect visual recognition with diverse cultural contexts poses inherent challenges for VLMs. Our dataset enables evaluating and identifying such aspects where models can be improved.

Model	Type 1	Type 2	Total
InstructBLIP-7B	45.57	53.81	50.84
InstructBLIP-13B	51.48	57.86	55.56
mPLUG-Owl2-7B	43.04	51.19	48.25
LLaVA-1.6-7B	50.21	59.76	56.32
LLaVA-1.6-13B	54.01	58.81	57.08
InternLM-XC2-7B	56.12	61.67	59.67
Molmo-7B-D	55.27	64.29	61.04
Idefics2-8B	63.71	63.57	63.62
Llama-3.2-11B	69.20	67.38	68.04
Claude-3-opus	69.62	80.24	70.02
GPT-4-Turbo	78.90	81.90	80.82
Gemini-1.5-Pro	83.97	70.24	81.58
GPT-4o	**92.41**	**87.86**	**89.50**

Table 3: **Results on Different Question Types.** Type 1 and Type 2 denote question types in visual recognition and visual reasoning, respectively.

Model	EN	KO*	EN + KO*
Claude-3-Opus	70.02	65.30	70.32
GPT-4-Turbo	80.82	75.34	80.37
Gemini-1.5-Pro	81.58	80.82	83.41
GPT-4o	89.50	76.56	79.76

Table 4: **Results with Different Input Languages**. KO* is machine-translated texts. For EN+KO*, we provide questions and options in both languages to models.

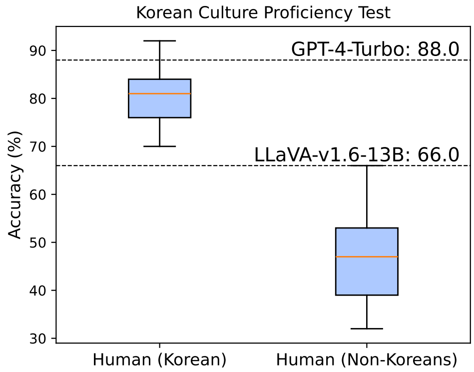

Figure 5: **Human Evaluation.** Accuracy comparison between Koreans and non-Koreans on 50 samples of `K-Viscuit`. The performance of selected models on this subset is also displayed. The average scores for Koreans and non-Koreans are 80.2 and 47.0, respectively.

	VQAv2	CVQA	K-Viscuit
LLaVA-v1.6-13B	82.8	57.9	57.1
Molmo-7B-D	85.6	65.5	61.0
Idefics2-8b	81.2	69.0	63.6
Llama-3.2-11B	75.2	72.4	68.0

Table 5: Evaluation results of selected open-source VLMs on various VQA datasets. The results on the VQAv2 (Goyal et al., 2017) are taken from the respective papers of each model. The performance on the CVQA dataset was measured on the Korean subset.

4.3 Analyses

Asking the VLM in Korean In this work, we primarily focused on evaluating the VLM's understanding of Korean culture and intentionally excluded their understanding of the Korean language. Thus, we designed the dataset to focus on the VLM's multiculturality, independent of its multilingualism. However, cultural information about certain groups often exists in the language that persons in the group frequently use. In other words, VLMs trained in multilingual corpora might have learned about Korean culture through texts written in Korean. Therefore, we probe whether asking questions in Korean could improve the performance of VLMs on our dataset. To translate our dataset into Korean, we use proprietary unimodal LM (`gpt-3.5-turbo`) as a machine translation system. For each sample in our dataset, the LM translates questions and four options into Korean. Three proprietary VLMs that can receive Korean text (i.e., GPT-4-Turbo, Claude-3-opus, and Gemini-1.5-Pro) are used for experiments.

Evaluation results with different input languages are presented in Table 4. We observe that solely providing translated texts in Korean to VLMs does not contribute to model performance. When English texts are also given to models, Gemini-1.5-Pro shows increased performance (i.e., 81.58 to 83.41). GPT-4-Turbo and Claude-3-opus usually do not take the benefits of using Korean texts.

Human Evaluation on the Benchmark We conducted a human evaluation to evaluate how well people from different backgrounds perform on `K-Viscuit`. We selected a subset of our dataset by randomly sampling 25 images, each with one Type 1 and one Type 2 question, totaling 50 questions. This test was administered to 20 Koreans and 14 non-Koreans. As depicted in Fig. 5, the results showed that participants with better knowledge of Korean culture achieved higher accuracy. However, even among Koreans, knowledge gaps exist between individuals and within a single person's knowledge across different domains, such as history. A Proprietary VLM (i.e., GPT-4-Turbo) shows comparable performance to human performance, indicating that utilizing VLMs for generating diverse, culturally nuanced questions is more effective than relying solely on individual human annotators. Details of the user study are provided in the Appendix B.2.

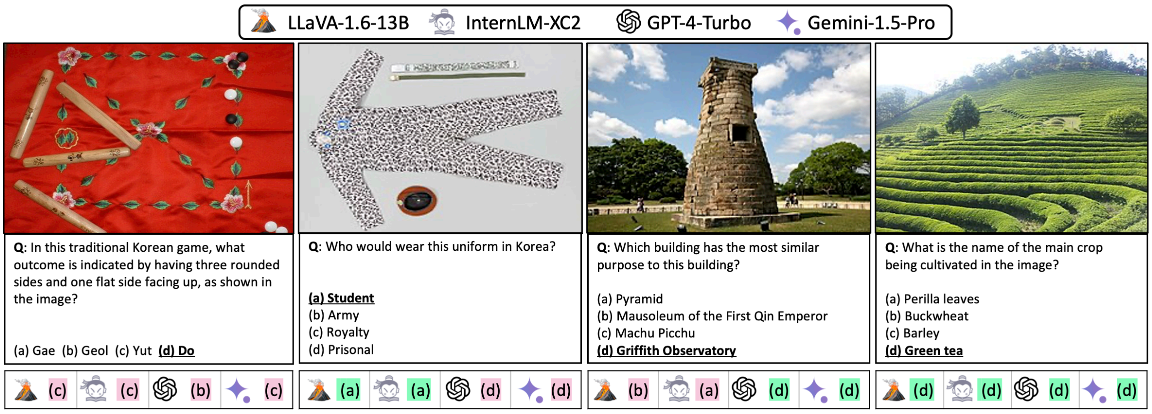

Figure 6: Qualitative examples of selected VLMs (i.e., LLaVA-1.6-13B, InternLM-XC2-7B, GPT-4-Turbo, and Gemini-1.5-Pro) on sampled questions. The answer option is highlighted in the underline. The correct and incorrect choice of models are highlighted in green and red.

Comparison of open-source VLMs on Various VQA Datasets We conducted experiments to measure the performance of VLMs on various VQA datasets. To this end, we compared the performance of four open-source VLMs on commonly used VQA benchmarks: the VQAv2 (Goyal et al., 2017) dev-test split, the Korean subset of CVQA (Romero et al., 2024), and our dataset, with results shown in Table 5. The VLMs demonstrated their best performance on VQAv2, while showing relatively lower accuracy on CVQA and our dataset. This suggests that the cultural questions we collected are indeed challenging for VLMs to solve.

Qualitative Results Fig. 6 presents the prediction of selected models on sampled examples. In the first example, all models fail to correctly answer the question about asking detailed game rules. In the second case, two proprietary models are wrong while open-source models make correct answers. The third problem requires the model to recognize the structure in the image as Cheomseongdae and infer that it serves a function similar to that of the Griffith Observatory in the United States. Both proprietary models make correct answers to this example. In the final example, all models successfully identify the correct answer about agriculture.

5 Discussion

We discuss several components and future directions to build VLMs grounded on diverse cultures.

5.1 Evaluation beyond Multiple-choice VQA

Our dataset is designed as a multiple-choice Visual Question Answering (VQA) task, where Vision-Language Models (VLMs) select one answer from given options. While this classification setup enables straightforward performance measurement through accuracy, it may not fully reveal the models' depth of cultural understanding. For instance, VLMs may correctly select an answer from the available options but fail in a generative VQA setup, where they must generate a free-form answer. As shown in Fig. 7, while the VLM accurately identifies the object as *bibimbap* in both multiple-choice and generative setups for the first example, it fails to provide accurate cultural context in the second example despite choosing the correct option in the multiple-choice format.

To quantitatively analyze this aspect, we randomly sampled 80 questions from the Food category and evaluated Llava-v1.6-13B's performance in a generative setting. The model was prompted with the instruction: *"This is a question about Korean culture. Answer the given questions briefly and concisely."* without access to the original multiple-choice options. Two native Korean evaluators assessed the generated responses using three categories: (1) *Correct*: The model's response fully contained the original answer or was reasonably accurate to the question, (2) *Hallucinated*: The model confidently provided incorrect information, and (3) *Abstained*: The model's response did not include a direct answer to the question but provided a relevant description of the image and did not attempt to generate an answer hastily (e.g., *"The specific type of food and its cultural context would determine when it is commonly served"*).

The human analysis results are shown in Table 6. We observed that the VLM's accuracy in the generative setup was relatively lower than in the multiple-choice setup (from 45.07% to 36.25%). This de-

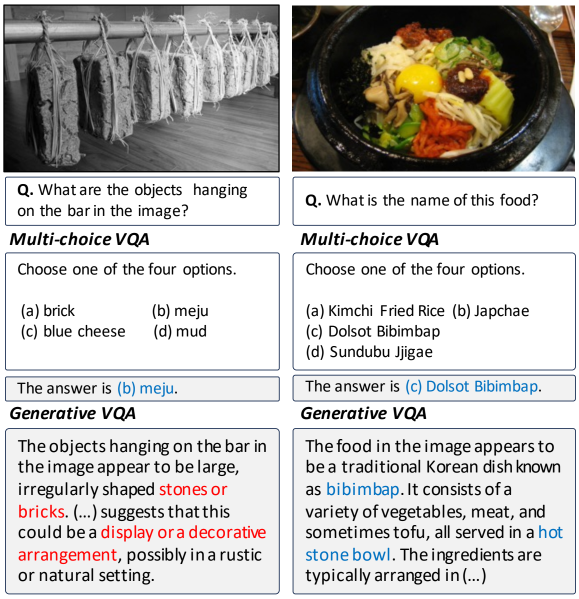

Figure 7: Qualitative results of a VLM with multi-choice and generative VQA setups. LLaVA-1.6-13B is used for the analysis. The correct and incorrect model generations are manually highlighted by the authors.

Correct	Hallucinated	Abstained	Total
29 (36.35%)	34 (42.5%)	17 (21.25%)	80 (100%)

Table 6: Human evaluation results of LLaVA-1.6-13B on *Food* category.

crease in accuracy is likely because the task is inherently more difficult without predefined options. Interestingly, a significant proportion of responses (21.25%) were classified as abstained, indicating that the VLM recognized the need for additional information to answer confidently and refrained from providing a definitive but potentially incorrect response. This behavior suggests that providing the necessary external knowledge could enhance performance. Our future research aims to propose various evaluation methods to assess whether VLMs genuinely possess the knowledge to answer questions accurately.

5.2 Augmenting VLMs with External Knowledge

In previous experiments, models struggled with questions requiring cultural knowledge. To address these gaps, we considered fine-tuning open-source models and augmenting models with external knowledge. Since the latter can be applied to both open-source and proprietary models, we focused on that approach. We augmented the models with relevant documents for each test image from the FOOD concept. Our retrieval method involved

Model	NONE	RETRIEVED	ORACLE
LLaVA-1.6-7B	43.66	68.31	78.87
LLaVA-1.6-13B	45.77	64.08	80.28
InternLM-XC2-7B	50.70	68.31	82.39
Idefics2-8B	51.41	67.61	82.39
Claude-3-opus	62.68	70.42	87.32
GPT-4-Turbo	73.94	78.17	88.73
Gemini-1.5-Pro	80.28	78.17	90.85
GPT-4o	88.73	83.10	92.25

Table 7: Retrieval-augmented Generation Results.

generating captions using GPT-4-Turbo, embedding these captions via `text-embedding-3-large` to build queries, and retrieving Top-1 document from 152 Wikipedia pages related to the objects mentioned in the TYPE 1 options. The `K-Viscuit` distractors were designed to closely resemble correct answers, creating a challenging retrieval setting with numerous hard negatives that mimic real-world conditions. We assess if the retrieval method remains effective under these conditions by examining performance changes when cultural knowledge is introduced.

As shown in Table 7, Providing retrieved documents can enhance model performance, as demonstrated by improved scores in the **RETRIEVED** setting compared to **NONE**. In cases where proprietary models did not benefit from certain retrieved documents, their performance still surpassed the baseline when given carefully curated **ORACLE** documents. This suggests that retrieval-based augmentation has the potential to strengthen cultural knowledge in VQA tasks, provided that the quality and relevance of the selected documents are refined. We provide the details in the Appendix D.

6 Conclusion

In this work, we proposed a semi-automated framework for constructing culturally aware benchmarks for vision-language models through human-VLM collaboration, focusing on visual recognition and reasoning in multi-choice QA. We demonstrated its effectiveness by creating `K-Viscuit`, revealing a significant performance gap between proprietary and open-source VLMs in understanding Korean culture. Through detailed analyses and knowledge augmentation experiments, we established the impact of cultural understanding on VQA performance and extended our investigation to open-ended answer generation. This work highlights the importance of cultural diversity in model evaluation and paves the way for more inclusive VLMs.

Limitations

Our current framework requires a manual selection of images that match specified concepts, preventing fully automated dataset generation. These human efforts can be alleviated by multimodal retrieval modules to some extent. To this end, it should be predetermined whether current multimodal encoders can sufficiently understand culturally nuanced images and texts. Enhancing retrieval models to better understand and match cultural contexts remains our exciting future work. The manual verification of automatically generated questions also can be a considerable burden. Developing a quality estimation module for generated questions could assist in this process by reducing the workload on human annotators.

Ethical Statement

In constructing K-Viscuit, we ensured all images were sourced from Wikimedia Commons under Creative Commons licenses, maintaining proper attribution and copyright compliance. The dataset was carefully curated to avoid harmful or inappropriate content, and all questions and answers were verified by native Korean speakers to ensure cultural relevance. While comprehensive, we acknowledge that our dataset captures only a subset of Korean cultural elements.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Max Bartolo, Tristan Thrush, Sebastian Riedel, Pontus Stenetorp, Robin Jia, and Douwe Kiela. 2022. Models in the loop: Aiding crowdworkers with generative annotation assistants. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3754–3767.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv preprint arXiv:2307.16883*.

Mary Ritchie Key and Bernard Comrie, editors. 2015. *IDS*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. Click: A benchmark dataset of cultural and linguistic intelligence in korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346.

Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Yoo, and Minjoon Seo. 2023. Aligning large language models through synthetic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13677–13700.

Claire Kramsch. 2014. Language and culture. *AILA review*, 27(1):30–55.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. llava. *Advances in neural information processing systems*, 36.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Steenkiste, Lisa 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.

OpenAI. 2024. Hello gpt-4o: Enhanced efficiency for gpt models. Accessed: 2024-05-13.

Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. 2024. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36.

Hector Ramos. 2020. Cognitive fixation and creativity. In *Encyclopedia of creativity, invention, innovation and entrepreneurship*, pages 319–320. Springer.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

D SUSAN. 1996. Language shock: understanding the culture of conversation.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*.

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.

Haryo Akbarianto Wibowo, Erland Hilman Fuadi, Made Nindyatama Nityasya, Radityo Eko Prasojo, and Alham Fikri Aji. 2023. Copal-id: Indonesian language reasoning with local culture and nuances. *arXiv preprint arXiv:2311.01012*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le

Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*.

Da Yin, Feng Gao, Govind Thattai, Michael Johnston, and Kai-Wei Chang. 2023. Givl: Improving geographical inclusivity of vision-language models with pre-training methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10961.

Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the vision: Geo-diverse visual commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2115–2129.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

A Dataset Construction Details

Guidelines to `GPT-4-Turbo` for Dataset Annotation The detailed prompts for the annotation with GPT-4-Turbo are presented in Table 8 and Table 9.

B Dataset Analyses Details

B.1 Required Knowledge Analysis

We analyzed how diverse the cultural knowledge required by the questions in our `K-Viscuit` dataset is. To this end, we used the following prompt to obtain responses from the `GPT-4` model. We delivered *all* TYPE 2 samples (including both the questions and the options) to the model by concatenating them into a single string.

B.2 Human Evaluation Details

We randomly selected images according to the proportion of each category to create the questionnaire for the human evaluation. If there were multiple TYPE 2 questions for a single image, we sampled them randomly. The number of selected images per category is as follows: FOOD (4), BEVERAGE (2), GAME (2), CELEBRATIONS (2), RELIGION (2), TOOL (3), CLOTHES (2), HERITAGE (2), ARCHITECTURE (4), and AGRICULTURE (2).

We released the survey on the Amazon MTurk platform, where non-Koreans with a relatively limited understanding of Korean culture were asked to complete the `K-Viscuit` questions within 20 minutes for a compensation of $5. The survey on the MTurk platform resulted in a demographic composed entirely of Americans. Their self-assessed proficiency levels were: *Very familiar* (35.7%), *Somewhat familiar* (50%), *Slightly familiar* (14.3%), and *Not familiar at all (0%)*. For Koreans, we administered the survey to 20 graduate students in their mid-to-late twenties. We received feedback that Koreans had the most difficulty with questions related to history.

C VLM Evaluation Details

Model Implementation Details We present further implementation details in VLMs used in our experiments. All open-source VLMs are implemented with the Transformers framework (Wolf et al., 2020), and the checkpoints are downloaded from Huggingface Hub[3]. For proprietary models, `gpt-4-turbo-2024-04-09`, `gemini-1.5-pro`, and `claude-3-opus -20240229` are used. The text

[3]`https://huggingface.co/models`

prompt used for proprietary models is presented in Table 11.

Prompt for TYPE 1 annotations:

`[System Prompt]`
You are a helpful Korean annotator to make visual question answering datasets.

`[User Prompt]`
Given an image, generate a question asking for the name of the main object or the main activity that people are engaged in, and generate one correct option (answer) and three wrong options (distractors).

Detailed guidelines are as follows:
1. The objects shown in the image is called **“{object_name}”** in Korean. You can include this word into your correct options after translation into English.
2. All options should be written in up to 5 words.
3. Please struggle to make creative or challenging distractors so that they are not easily distinguished from the answer.
4. Distractors should seem similar to the correct answer and related to the category of the main object in image (e.g., Hanok - Agungi, Sarangchae, Anchae, Daecheongmaru).
Distractors are better when they have similar color, shape, or texture with the answer.
5. Separate each distractor with “;” symbol.
6. Don’t make any explanation.
7. Distractors should be culturally related to the image.
8. All the options should be either transliterated or translated. Never mix transliteration and translation.
9. Don’t be too specific (Avoid using a proper name: instead use `[University building]` instead of `[Ewha Campus Complex]`)

You can refer to below examples that are annotated for other images.
Question: What is the name of this place shown in the image?
Answer: Sarangchae
Distractors: Anchae ; Sadang ; Daecheongmaru

Question: What is the name of the structure seen in the image?
Answer: Ondol
Distractors: Agungi ; Jangdokdae ; Buttumak

Question: What is the name of this building shown in the image?
Answer: Gosiwon
Distractors: Officetel ; Apartment ; Share house

Please make four options (single answer and three distractors).

Table 8: A prompt of GPT-4-Turbo used in the annotation of TYPE 1 questions (i.e., *visual recognition*) in the ARCHITECTURE category.

Prompt for TYPE 2 annotations:

`[System Prompt]`
You are a helpful Korean annotator to make visual question answering datasets.

`[User Prompt]`
Please ask 5 questions and their options about the image. Here are the guidelines to follow for writing.

Detailed guidelines are as follows:
1. The objects shown in the image is **“{object_name}”**. Don’t include this word in your questions.
2. The question should require looking at the image to answer.
3. Questions should require some knowledge about Korean cultures.
4. Don’t make a simple question that does not require knowledge of Korean cultures, such as recognizing objects or counting objects.
5. It is desirable to generate questions that are difficult for foreigners who are unfamiliar with Korean culture.
6. After writing a question, please write a single correct option (answer) and three wrong options (distractors) for your above question.
7. All options should be written in up to 5 words.
8. Don’t ask traditional celebrations about the given image.
9. Try to ask questions that are more derived from the given image.
10. Any creative questions are very welcome.
11. Separate each distractor with “;” symbol.
12. [Description]
“{object_name}\n{description}”

You can refer to below examples that are annotated for other images.
Question: Seen in the image, what traditional Korean heated floor system is associated with the heat source from this feature?
Answer: Ondol
Distractors: Daecheongmaru ; Anchae ; Buttumak

Question: In the image, what kind of Korean roof finishing is visible, known for its multicolored patterns?
Answer: Dancheong
Distractors: Seoggarae ; Cheoma ; Maru

Question: Which mode of transportation is commonly used by tourists to ascend the mountain where the tower is located?
Answer: Cable Car
Distractors: Bus ; Bicycle ; Funicular Railway

Please make four options (single answer and three distractors).

Table 9: A prompt of GPT-4-Turbo used in the annotation of TYPE 2 questions (i.e., *visual reasoning*) in the ARCHITECTURE category.

Prompt for required knowledge analysis:

I created a multiple-choice quiz with four options per question, based on images related to Korean culture. Each question is designed to assess the understanding of one or more cultural elements. Please analyze which cultural element each question aims to measure and provide an overall summary.
`"{TYPE 2 samples}"`

Table 10: Prompt for required knowledge analysis.

Inference prompt for proprietary VLMs:

`[System Prompt]`
You will be given an image taken in Korea and a 4-way multiple-choice question. Answer the question based on the given image and your knowledge about Korean culture.

`[User Prompt]`
Question: **`"{question}"`**
Options:
a. **`"{option_a}"`**
b. **`"{option_b}"`**
c. **`"{option_c}"`**
d. **`"{option_d}"`**

Make sure to respond with the option's letter: 'a.', 'b.', 'c.', or 'd.'. Do not make any additional explanation.

Table 11: Inference prompt for proprietary VLMs.

D Retrieval Methodology Details

For external knowledge retrieval, we generated image captions using `GPT-4` with the prompt shown in Table 12.

Prompt for generating image captions:

You are an AI language model specializing in identifying and describing Korean food from photographs. When given a photograph of Korean food, your task is to accurately describe the food based on its visual characteristics and visible ingredients. Your description should include the name of the dish, main ingredients, common accompaniments, and notable features that help identify the food. Be detailed yet concise, providing clear and helpful information to those trying to understand Korean cuisine. Ensure that your description is within 150 words.

Table 12: Prompt for generating image captions.

E Comparison of `K-Viscuit` with CVQA dataset

Recent advances in vision-language models have sparked a growing interest in evaluating their cultural awareness across diverse global contexts. A notable contribution in this direction is CVQA (Romero et al., 2024), which introduces a comprehensive multilingual VQA benchmark spanning 28 countries with 9,044 total examples, averaging approximately 323 examples per country. Our `K-Viscuit` dataset, while focused solely on Korean culture, contains 657 examples—substantially exceeding CVQA's per-country average. Moreover, CVQA's Korean subset comprises 290 examples, less than half the size of `K-Viscuit`. This focused approach enables `K-Viscuit` to provide more comprehensive coverage of Korean cultural elements, offering deeper insights into VLMs' understanding of specific cultural contexts.