Enhancing Cultural Awareness in Vision-Language Models: The Power of Multimodal Few-Shot Prompting

Pitikorn Khlaisamniang[1,2]

[1]Looloo Technology
[2]Artificial Intelligence Association of Thailand

Abstract

Visual-Language Models (VLMs) have demonstrated significant capabilities in multimodal understanding, yet their awareness of diverse cultural contexts remains a critical area for evaluation. Standard benchmarks often fall short in assessing culturally specific knowledge. This work investigates the impact of prompt engineering strategies on VLM performance, particularly focusing on techniques relevant to evaluating nuanced understanding, potentially including cultural awareness. We compare zero-shot performance with few-shot prompting using both text-only and multimodal (image-text) examples on the CulturalVQA benchmark. Our findings indicate that few-shot prompting leads to a notable improvement over the zero-shot baseline. Text-based few-shot prompts show a clear increase in performance, while multimodal few-shot prompts that incorporate both text and images achieve the best results. These outcomes underscore the power of few-shot prompting—especially with multimodal examples—in enhancing VLM performance on tasks requiring specific contextual understanding and suggest that prompt engineering is a valuable tool for probing and improving the capabilities of VLMs in specialized domains, including cultural contexts addressed by benchmarks like CulturalVQABench.

1. Introduction

Visual-Language Models (VLMs) represent a significant advancement in artificial intelligence, demonstrating remarkable capabilities in understanding and generating content that bridges visual and textual modalities [8, 11]. These models, often built upon large-scale pre-training, excel at tasks like image captioning and visual question answering (VQA). However, as VLMs become more integrated into diverse global applications, ensuring they possess nuanced, culturally sensitive understanding is paramount.

While standard benchmarks evaluate general VQA capabilities [6], they often lack the focus required to assess a model's grasp of cultural contexts, symbols, and practices depicted in images. Recognizing this gap, recent efforts have focused on developing specialized benchmarks. One notable example is CulturalVQABench [10], specifically designed to probe the cultural awareness of VLMs by presenting questions that require understanding culturally specific visual elements. Evaluating models on such benchmarks is crucial for identifying limitations and guiding the development of more culturally competent AI systems.

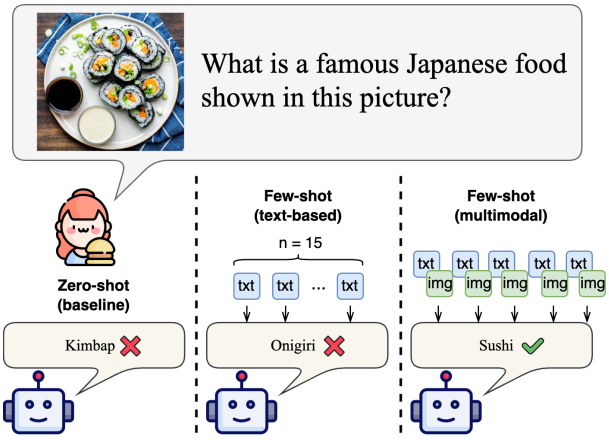

Figure 1. Comparison of **Zero-shot Prompting**, **Few-shot (text-based) Prompting** and **Few-shot (multimodal) Prompting** for answering questions based on image-text pairs.

Furthermore, the performance of VLMs can often be significantly influenced by how they are prompted. Prompt engineering, particularly few-shot prompting where the model is given a small number of examples within the prompt itself, has emerged as a powerful technique to guide model behavior [4]. Exploring different few-shot strategies (Fig. 1), such as using only text examples versus combining text with relevant image examples (multimodal prompting), offers a pathway to potentially enhance or better evalu-

Figure 2. Samples from the CulteralVQA dataset [10], which consists of images depicting cultural concepts from 11 countries across five facets: traditions, rituals, food, drink, and clothing. The dataset also includes questions designed to assess cultural understanding of the depicted concepts, along with corresponding answers. These samples are also used in **Few-shot (text-based) Prompting**.

ate the cultural reasoning abilities of VLMs when faced with benchmarks like CulturalVQABench. This work investigates the intersection of these areas, examining the cultural awareness of contemporary VLMs using CulturalVQABench and exploring the impact of different few-shot prompting techniques on their performance.

2. Related Work

2.1. Visual-Language Models (VLMs)

The field of VLMs has rapidly evolved, moving from models focused on specific tasks to large-scale, versatile architectures. Early approaches often involved separate encoders for vision and language, combined through attention mechanisms [2]. The advent of transformer architectures and large-scale pre-training on image-text pairs led to breakthroughs like CLIP [11], which learns joint embeddings, and ALIGN [7]. Subsequent models like BLIP [8], Flamingo [1], and GPT-4V(ision) [12] have further pushed the boundaries, demonstrating impressive zero-shot and few-shot capabilities across a wide range of multimodal tasks, including complex reasoning and VQA. These models typically leverage vast datasets, learning rich representations but potentially inheriting biases present in the data.

2.2. Cultural Awareness and Bias in AI

The potential for cultural bias in AI systems is a well-documented concern, spanning natural language processing [3] and computer vision[5]. Models trained predominantly on data from specific cultural contexts may fail to generalize or may exhibit biased behavior when deployed globally. Research has highlighted disparities in performance across different demographic groups and cultural settings. This underscores the need for AI systems, especially VLMs that interpret visual scenes, to possess cultural awareness – the ability to recognize and correctly interpret culturally specific objects, symbols, actions, and contexts. Evaluating and mitigating cultural bias is an active area of research, driving the development of culturally sensitive datasets and evaluation metrics.

2.3. Prompt Engineering for VLMs

Prompt engineering involves designing effective input prompts to elicit desired behaviors from large models without retraining them. For VLMs, this extends to multimodal prompts. Few-shot prompting, where the prompt includes a small number (k) of input-output examples, has proven effective in adapting large models to specific tasks or domains [4, 12]. In the context of VQA, few-shot prompts can provide examples of the desired question-answering format

or reasoning style. These prompts can be purely text-based (providing question: answer pairs) or multimodal (providing image, question: answer triplets). Comparing text-only versus multimodal few-shot prompting strategies is relevant for understanding how best to leverage context, especially when probing specialized knowledge domains like cultural understanding within VLMs [1, 12].

3. Methodology

The methodology employed in this study centers on leveraging a powerful base VLM and enhancing its performance on the CulturalVQA task through carefully designed few-shot prompts.

3.1. Base Model

The experiments utilize the publicly available API of GPT-4o, a state-of-the-art proprietary VLM recognized for its strong multimodal understanding and generation capabilities. While the closed-source nature of this model imposes limitations regarding transparency and reproducibility, its selection was motivated by its position as the top-performing VLM baseline reported in the CulturalVQA [10].

3.2. Prompting Strategy Rationale

Few-shot prompting was chosen as the primary intervention strategy. The rationale is that providing the model with a small number of illustrative examples within the prompt can effectively guide its reasoning process and supply relevant contextual information or reasoning patterns specific to the cultural domain. This approach aims to compensate for potential deficiencies in the model's training data regarding specific cultural knowledge, directly addressing the hypothesized cause of "lack of sufficient culturally relevant data", by providing such context at inference time. Furthermore, structured prompts might aid the model in better integrating visual cues with the required cultural interpretation, potentially mitigating the hypothesized "difficulty in combining cultural information from across vision and language modalities".

Figure 3 shows the prompts used in this paper for performance comparison.

3.2.1. Technique 1: Text-Only Few-Shot Prompting

The first approach involved constructing prompts containing 15 text-only few-shot examples (Fig. 2) preceding the actual target image-question pair from the CulturalVQA dataset. Each example within the prompt consisted of a [Question], its corresponding [Answer] and its simple image description [Info]. The selection principle for these 15 examples involved curating pairs that covered a diverse range of cultural concepts and reasoning types represented within the broader CulturalVQA domain, aiming to prime

Base Prompt

You will be given an image depicting a cultural concept and a question about the image.
Answer the question with a precise, culturally specific response (e.g., 'sushi' instead of 'food', 'Diwali' instead of 'festival') of 1-3 words.

Few-shot (text-based, n=15)

Here are some examples of the described task.

Examples:
  Image: {Info_i}
  Question: {Q_i}
  Answer: {A_i}
  ...

Few-shot (multimodal, n=5)

Here are some examples of the described task.

Examples:
  Image: {Image_i}
  Question: {Q_i}
  Answer: {A_i}
  ...

Question: {question}
Answer:

Figure 3. Prompts used in this paper: green backgrounds indicate **Few-shot (text-based)** prompting, while yellow backgrounds represent **Few-shot (multimodal)** prompting.

the model with relevant concepts, and answer structures. The hypothesis was that exposure to these text-based patterns of culturally relevant Q&A would improve the model's ability to generate accurate answers for new, unseen questions, even without direct visual input in the prompt examples themselves.

3.2.2. Technique 2: Multimodal Few-Shot Prompting

The second approach explored multimodal few-shot prompting, incorporating 5 examples (Fig. 4) where each example included an [Image], its associated [Question], and the [Answer]. These examples were formatted using the native multimodal input capabilities of the GPT-4o. The selection of these 5 examples focused on providing clear instances where specific visual elements needed to be linked to cultural knowledge to arrive at the correct answer. The smaller number of shot compared to the text-only approach (5 vs 15) was partly determined by API input constraints and the hypothesis that multimodal examples, being more information-dense by explicitly grounding the question in visual evidence, might yield significant improvements with fewer instances. This technique directly targets the challenge of modality integration, explicitly demonstrating the required reasoning process of connecting visual features to cultural interpretations.

4. Experimental Results

4.1. Benchmark

The evaluation was performed using the official CulturalVQA benchmark [10]. This benchmark consists of 2,378

Image Id:
f7716a9b-4dd1-411a-b4cd-a84d1f770dc6
Question:
This image shown can be use for what in Africa?
Answer:
For bathing and other traditional use.

Image Id:
c3a1ddb5-f12f-4806-bda6-4189d38228d7
Question:
In which occasion does the woman put salt in the hot beverage depicted in the item?
Answer:
Ask for blessing

Image Id:
919b59e5-3be5-4e60-ae8a-ee665e0f32da
Question:
In which type of glass do Turkish people consume the thing depicted in the image?
Answer:
Slim-waisted glass

Image Id:
39e397ed-fa8c-40ee-974d-b2f01e397e64
Question:
What is the animal in the image depicts?
Answer:
Garuda

Image Id:
1a5ba0ee-6445-4fcc-99df-9eceae01bcce
Question:
What is the traditional occupation name of this person?
Answer:
Naghali

Figure 4. CulturalVQA samples utilized in **Few-shot (multimodal) Prompting**.

image-question pairs designed to test cultural understanding across diverse global contexts (11 countries, 5 continents).

For the evaluation metric, we use LAVE (LLM-Assisted VQA Evaluation) [9], which leverages the in-context learning capabilities of instructiontuned LLMs.

4.2. Performance Comparison

Prompting Method	LAVE (Public Set)
Zero-shot	62.531645 (baseline)
Few-shot (text-based)	65.569620 ↑(4.86%)
Few-shot (multimodal)	69.620253 ↑↑(11.34%)

Table 1. Performance improvements of each prompting method on the CulturalVQA Benchmark (Public Set) using GPT-4o as the vision-language model (VLM).

The results (Tab. 1) demonstrate a clear benefit from using few-shot prompting techniques:

- Zero-shot (baseline) performance was score 62.53.
- Few-shot (text-based) prompting improved performance to 65.57, an increase of 4.86% over the baseline.
- Few-shot (multimodal) prompting achieved the highest performance at 69.62, representing a significant increase of 11.34% compared to the zero-shot baseline.

These findings indicate that providing examples significantly aids the model, and incorporating visual context alongside textual examples in multimodal prompts yields the most substantial performance gains on this benchmark.

4.3. CulturalVQA Challenge Scoreboards

We applied our proposed few-shot (multimodal) prompting technique to the CulturalVQA Challenge at CVPR 2025 and achieved first place on both the public and private scoreboards, as shown in Figure 5.

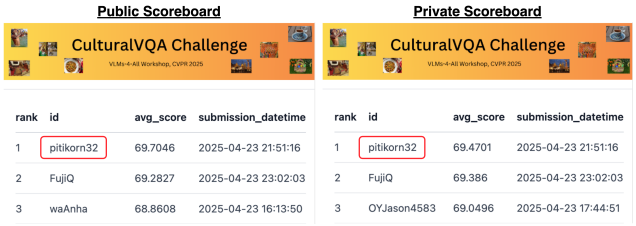

Figure 5. Scoreboards for the CulturalVQA Challenge (Public and Private) at CVPR 2025.

5. Conclusion

This work explored the capabilities of VLMs, focusing on cultural awareness evaluation and the role of prompt engineering. Our performance comparison on the CulturalVQA benchmark (Public Set) highlights the effectiveness of few-shot prompting strategies in enhancing VLM performance compared to a zero-shot baseline. Notably, multimodal few-shot prompting, which provides both image and text examples, demonstrated a superior improvement (+11.34%) over text-only few-shot prompting (+4.86%).

These results underscore the importance of context provided through prompt examples, particularly the richer context offered by multimodal examples, in guiding VLMs towards more accurate responses in complex VQA tasks. While benchmarks like CulturalVQABench [10] specifically target cultural understanding, the observed improvements from prompting on the CulturalVQA benchmark suggest that these techniques are valuable tools for probing and potentially enhancing VLM capabilities across various specialized domains. Further research can explore the optimization of multimodal prompts and their effectiveness across diverse cultural contexts and VQA benchmarks.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. 2, 3

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering, 2018. 2

[3] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, 2020. Association for Computational Linguistics. 2

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 1, 2

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018. 2

[6] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. 1

[7] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. 2

[8] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 1, 2

[9] Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. Improving automatic vqa evaluation using large language models, 2024. 4

[10] Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding, 2024. 1, 2, 3, 4

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 2

[12] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023. 2, 3