

# BANGLAPROTHA: Evaluating Vision Language Models in Underrepresented Long-tail Cultural Contexts

Md Fahim<sup>1,2\*</sup>, Md Sakib Ul Rahman<sup>3\*</sup>, Akm Moshir Rahman<sup>2\*</sup>, Md Farhan Ishmam<sup>4\*</sup>,  
Md Tasmim Rahman<sup>1</sup>, Fariha Tanjim Shifat<sup>1</sup>, Fabiha Haider<sup>1</sup>, Md Farhad Alam Bhuiyan<sup>1</sup>

<sup>1</sup>Penta Global Limited    <sup>2</sup>CCDS, Independent University, Bangladesh

<sup>3</sup>University of Maryland, Baltimore County    <sup>4</sup>University of Utah

\*Equal Contribution {fahimcse381, souroveskb, pdcsedu}@gmail.com

## Abstract

*The advanced multimodal processing of current vision language models (VLMs) has prompted rigorous benchmarking across multicultural settings, revealing a clear inclination toward Western culture. While the bias likely stems from the predominance of Western-centric images in the VLM pretraining data, the resulting long-tail distribution problem is only exacerbated in underrepresented cultural settings, such as Bengali. Our work explores this problem through an aspect-based evaluation of several classes of VLMs on the rich Bengali culture. Our BanglaProtha dataset is a VQA dataset, containing images that encapsulate Bengali cultural elements, questions in native Bengali, and semantically similar multiple-choice answer options. Our experiments provide behavioral insights into VLMs across prompting & fine-tuning strategies, cultural aspects, model size, and augmentation methods. Our work serves as a diagnostic tool for addressing and mitigating inequalities in multicultural and multilingual settings, thereby bringing efforts to democratize AI systems. Our code and data are available at <https://github.com/farhanishmam/BanglaProtha>.*

## 1. Introduction

The recent scaling of VLMs [10, 50] led to extensive improvement across several vision-language tasks, such as visual question answering (VQA) [7], visual grounding [37], and visual reasoning [28]. The visio-linguistic elements of the associated datasets vary widely, thereby challenging the exceptionally advanced systems [25]. One such variation arises in multicultural settings where the images contain cultural artifacts, e.g., region, event, architecture, *inter alia*, that are typically not observed in benchmark datasets [45]. The associated questions also center on these cultural aspects and are often multi-lingual, where non-English scripts

bring linguistic variations in the textual modality [53].

There have been several instances where VLMs exhibit substantial bias towards Western cultural concepts [45, 52]. The performance disparity can be attributed to the imbalance in pre-training data of the associated VLMs [47] and hence re-framing the multi-cultural evaluation problem to a long-tail distribution problem, *i.e.*, the niche cultural elements are present at the tail-end of the pre-training data distribution. As contemporary VLMs often require or rely on a training paradigm [24], the lack of resources makes it challenging to mitigate this skewed distribution problem.

Despite recent interest in multicultural and multilingual VLM evaluation [9, 53], the cultural depth of such benchmarks remains shallow. We exemplify this through the diverse Bengali<sup>1</sup> culture. To characterize the uniqueness of this culture, several aspects must be considered, *e.g.* food, events, landmarks, and art. However, existing benchmarks rarely explore beyond surface-level aspects of this culture [45]. Additionally, Bengali has a unique blend of Indic and Southeast Asian cultural elements, often requiring a multi-faceted understanding of cultural concepts [43].

With several culture-specific benchmarks in other non-Western cultures [46, 61], we found it crucial to construct a dataset and evaluate VLMs on culturally unique aspects of Bengali. Our contributions can be summarized as:

- We present BANGLAPROTHA, a VQA dataset with images encapsulating nine distinct Bengali cultural aspects, questions in native Bengali, and semantically similar multiple-choice answer options.
- We evaluate monolingual, multilingual, and large-scale VLMs on our dataset using five prompting and four fine-tuning strategies to assess their performance.

<sup>1</sup>In this work, *Bangla* and *Bengali* are used synonymously to denote the same language, cultural identity, and people, predominantly associated with Bangladesh and the West Bengal region of India.



Figure 1. Sample images from the BANGLAPROTHA dataset across different cultural aspects, where (i) the images are relevant to Bengali cultural aspects, *i.e.*, event and fashion/attire (ii) the question is in native Bengali scripts with the English translation, and (iii) the answer options are semantically similar. The English translation has been provided for non-Bengali speakers and is *not* part of our dataset.

- Our findings reveal key behavioral insights on the cultural aspects, model size, necessity of training data & answer options, and vision-language alignment.

## 2. Related Works

**Multilingual & Bengali VQA.** While VQA research has predominantly been conducted in English [15], efforts have been made to develop non-English and multilingual VQA datasets, *e.g.*, FM-IQA [19], MCVQA [22], Multi30K [17], xGQA [49], and MaXM [15]. For Bengali, initial benchmarks like Bengali-VQA-v1 [26], derived from VQAv1 [7], and Bengali CLEVR [26], derived from CLEVR [28], were created via machine translation of English datasets. However, machine translation often struggles with low-resource languages like Bengali, as it fails to capture the nuances of the language and introduces linguistic artifacts [18, 40].

Rafi *et al.* [51] introduced a manually annotated Bengali VQA dataset derived from VQAv2 [21], but limited to binary questions and Western-centric images that fail to represent Bengali region-specific contexts. ChitroJera [8] and BVQA [13] addressed these issues using images relevant to the Bengali region. However, their QA pairs are generated using LLMs, which limits the questions' ability to reflect cultural nuances. Furthermore, both datasets serve as standard VQA benchmarks, without any categorization based on cultural concepts.

**Western Bias & Multicultural VQA.** Recent studies have highlighted the performance disparity across cultural and

social norms, with VLMs exhibiting bias towards the Western counterparts [14, 44]. A lack of Bengali cultural understanding has also been evident [52]. Multicultural VQA benchmarks [11, 53] serve as a diagnostic tool in identifying such cultural biases, albeit with limited samples from each country, culture, or geographic region.

The CVQA dataset [53] evaluates VQA models on multi-lingual and multi-cultural contexts from 30 countries across 10 distinct cultural aspects. However, it simplifies the Bengali culture to the Indian region only, overlooking the diversity and traditions of the Bengali culture from Bangladesh. CultureVQA [45] establishes a similar benchmark across 11 countries, but categorized geographically. Thus, the Bengali culture was blended with the rest of the rich Indian culture. CultureVerse [32] provides the most diverse benchmark across 188 regions, but has limited samples in Bengali. Closest to our work is ALMBench [57], which includes Bengali cultural aspects from the whole region but lacks the nuances in answer options.

## 3. BANGLAPROTHA Dataset

**Cultural Concepts Categorization.** We systematically categorize BANGLAPROTHA into nine diverse categories encapsulating several Bengali cultural aspects as seen in Fig. 1. Our categorization draws inspiration from prior works in VQA [32, 39, 53], while adapting to better align with the Bengali cultural context. We aggregated several fine-grained categories into generalized ones: (1) vehicles & transportation, people, and everyday life to *Social*

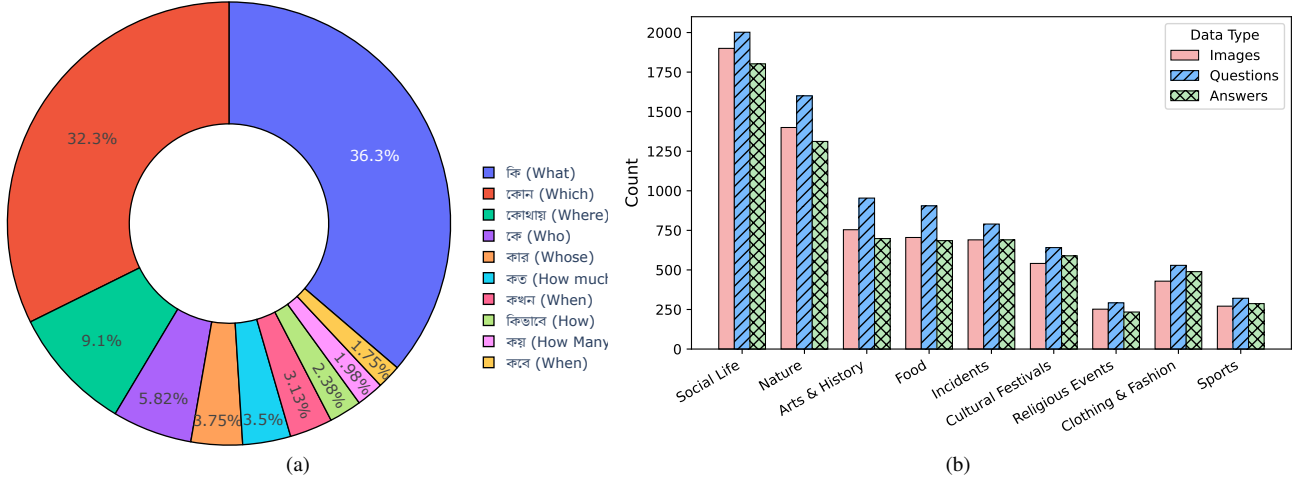


Figure 2. (a) Question type Distribution, (b) Category-wise count of images, questions, and unique answers across cultural concepts.

Life, (2) plants, animals, landscape, and buildings to Nature, (3) tradition, art, history, public figure, and pop-culture to Art and History. Detailed category descriptions are provided in §C.

**Image Sourcing.** We initially source 21,254 images from previous Bengali image datasets: BORNON [42], BANGLALEKHA-IMAGECAPTIONS [52], and BNATURE [4]. We further diversify the dataset by sourcing 2,800 publicly available Bengali images from websites. After rigorous filtering, the final distribution of image sources is provided in Tab. 1 and Fig. (§A.2a).

**Data Annotation.** We recruited 17 native Bengali annotators (12 male, 5 female), who are university undergraduates with strong expertise in Bengali culture and literature. The annotators received standard monetary compensation on a per-sample basis. To ensure annotation quality, we provided (i) detailed annotation guidelines (§A.1,A.2) and (ii) a dedicated annotation tool (§A.3).

For each image, the annotators created (i) a culturally relevant question and (ii) four semantically similar and question-relevant answer choices. They were instructed to carefully examine the image objects and their context to produce semantically similar answer options. For instance, if an image depicts a rural game, the annotators should include the names of other rural games as options. These semantically similar options were designed to test a model’s ability to evaluate cultural elements. The annotators were also instructed to discard any image that did not belong to any of the nine culture categories.

**Annotation Verification.** Our dataset went through rigorous manual verification using a multi-stage filtering process. Samples were excluded based on several criteria: (1) images lacking sufficient cultural relevance, (2) imprecise

or ambiguous questions, (3) misalignment between images and questions, (4) incorrect categorization, and (5) questions focusing solely on object detection without cultural context. From the initial 9,479 annotated image-question samples, 8,034 ( $8034/9479 = 84.76\%$ ) met all quality standards and were retained, while the remaining 1,445 samples were discarded.

Data Source	#I	#Q	#Q:#I
BNATURE	2079	2517	1.21
BORNON	1402	1638	1.17
BANGLALEKHA	2541	2816	1.11
Bengali Websites	920	1063	1.16
<b>Total</b>	<b>6942</b>	<b>8034</b>	<b>1.16</b>

Table 1. Source distribution in BANGLAPROTHA. #I: No. of Images, #Q: No. of Questions, and #Q:#I: Question-Image ratio.

**Dataset Statistics & Analysis.** From Fig. 2b and Tab. (§A.3), we observe a higher sample count from the Social Life and Nature categories, whereas the count of unique images, questions, and answers remains somewhat uniform across the cultural aspects. “What” and “Which” type questions are predominantly more than other types, taking a substantial 68.6% of the total question types (Fig. 2a). Tab. 1 shows that BANGLAPROTHA has a strong representation of all four data sources while maintaining a good question-to-image ratio. Fig. (§A.2a) illustrates the distribution of cultural aspects across sources, e.g., website images have a high number of Incidents, but fewer Religious Events. Finally, BANGLAPROTHA is divided into the standard 80:20 train-test splits.

Prompt Strategy	Prompt Overview
Zero-Shot/Base Prompt	<i>Find the most accurate option for the given image, question, and answer options.</i>
Chain-of-Thought (CoT)	<i>Think step by step before selecting the answer.</i>
Translation-based	<i>Translate the question and options into English.</i>
Culture-Specific	<i>The image is culturally relevant to the Bengali culture across nine key concepts.</i>
Description-based	<i>First describe the contents of the image and then use it to generate the answer.</i>

Table 2. The prompting strategies used in our benchmarks. The base prompt is appended at the end of the latter four prompting strategies.

## 4. Experiment Design

We classify our experiments into two categories: (i) prompt-based and (ii) fine-tuning experiments. The model details have been reported in (§D.1).

### 4.1. Prompt-based Experiments

We consider five prompting strategies (Tab. 2), the vanilla zero-shot prompting, Chain-of-Thought (CoT) [59], Translation-based [23], Culture-specific [32], and Description-based prompting. Description-based prompting takes inspiration from previous works where passing a textual description helped in visual classification [8, 32]. Tab. 2 provides an overview of all prompts, with detailed descriptions available in (§G).

### 4.2. Fine-tuning Experiments

**Full Fine-tuning.** To perform full fine-tuning on the pre-trained multimodal model  $\phi_{\text{mm}}$ , we frame the task as a classification problem, following previous approaches [39, 54]. Given an image  $\mathbf{I}$ , a question  $\mathbf{Q}$ , and a set of answer choices  $\{a_1, a_2, \dots, a_n\}$ , where  $n$  denotes the number of candidate answers, the model  $\phi_{\text{mm}}$  is trained to predict the most relevant answer  $a_*$  from this set. We optimize the model using the standard **cross-entropy loss** over the answer choices, incentivizing the model to assign higher probabilities to the correct answers and vice versa.

**Partial Fine-tuning.** Most open-source VLMs adopt a modular Vision-Encoder + Adapter + LLM architecture. An input image  $\mathbf{I}$  is first processed by a vision encoder  $\phi_{\mathbf{I}}$ , followed by an adapter module  $\phi_{\mathbf{A}}$ , which transforms image features into a sequence of visual tokens. In parallel, the input text  $\mathbf{T}$  is encoded by a text embedding layer  $\mathcal{E}(\cdot)$  to obtain textual tokens. These visual and textual tokens are concatenated and passed to a large language model  $\phi_{\text{LLM}}$ , which performs multimodal reasoning and language generation. The output text prediction is given by:

$$\hat{\mathbf{T}} = \phi_{\text{LLM}}([\phi_{\mathbf{A}}(\phi_{\mathbf{I}}(\mathbf{I})); \mathcal{E}(\mathbf{T})]).$$

The entire model is trained using an autoregressive next-token prediction loss over the textual sequence  $\mathbf{T}$ . In our partial fine-tuning experiments, we explore three settings:

- ***L-LoRA***: Applies LoRA [24] fine-tuning to the language model component  $\phi_{\text{LLM}}$  within the multimodal model  $\phi_{\text{VLM}}$ , aiming better alignment of the visual and textual representations.
- ***L-LoRA + Adapter***: Fine-tunes both the adapter module  $\phi_{\mathbf{A}}$  and language model  $\phi_{\text{LLM}}$  using LoRA, to investigate whether updating the visual token transformation improves performance.
- ***L-LoRA + Adapter + VE***: Fine-tunes the vision encoder  $\phi_{\mathbf{I}}$  along with the previous components  $\phi_{\mathbf{A}}$  and  $\phi_{\text{LLM}}$ , to examine whether learning fine-grained visual representations leads to better performance.

Training settings for both full and partial fine-tuning are provided in §D.2 and §D.3 respectively.

## 5. Results Analysis

The results of prompt-based experiments on eight open-source and three closed-source VLMs under five prompting strategies are presented in Tab. 3. Similarly, Tab. 4 reports the performance of five VLMs for full fine-tuning and another five for partial fine-tuning, each employing three distinct strategies as outlined in Sec. 4.2.

**Open Source vs. Closed Source Models.** The accuracy of monolingual open-source VLMs remains below 42% for both zero-shot and CoT prompting (Tab. 3). In contrast, multilingual open-source VLMs generally exceed 50% accuracy, except for the smaller Phi-3.5-V. The performance difference is obvious, as the monolingual models are not trained to comprehend Bengali text.

Among the open-source models, only Gemma-3 12B matches or exceeds the performance of the closed-source models. This can be attributed to the *supposedly* larger size or pretraining data of the closed-source models. GPT-4o consistently outperforms the other models, with the exception of being slightly surpassed by Gemma in zero-shot prompting and Claude in description-based prompting. Overall, GPT-4o takes the crown using the culture-specific prompting, achieving an average accuracy of 83.42%.

**How should we prompt?** From Fig. 3, we observe that closed-source and larger models, *e.g.*, llama-3.2-V 11B,

		Models	Cultural Concepts									
			Food	Fest	Rel	Nature	Fash	Sport	Life	Art/Hs	Incid	Avg
Zero-Shot Prompting	O-Mono	BLIP-2 OPT 6.7B [29]	35.21	33.38	31.47	43.77	32.36	46.13	38.42	34.18	40.66	37.29
		LLaVa-1.5 7B [30]	33.31	33.32	40.00	40.00	25.06	38.28	51.68	38.36	33.30	37.04
		LLaVa-Next 7B [31]	28.32	36.74	40.05	43.36	43.26	36.66	44.96	51.78	45.13	41.14
	O-Multi	LLaMa-3.2-V 11B [41]	61.23	69.78	60.52	56.30	72.67	73.34	70.20	70.05	83.32	68.60
		Phi-3.5-V [1]	38.31	31.72	23.28	25.08	29.94	45.05	36.71	30.10	38.34	33.17
		Phi-4 Multimodal [2]	57.64	47.53	44.14	42.37	53.56	54.42	53.27	50.88	50.20	50.45
		Qwen-2.5 7B [58]	50.10	50.15	60.20	51.66	73.26	60.08	61.72	73.32	80.05	62.28
		Gemma-3 12B [56]	68.28	86.73	71.72	76.74	86.72	78.26	83.36	83.28	78.37	79.27
	Closed	Claude-3.5 Sonnet [6]	81.67	70.24	81.90	71.92	75.08	82.26	81.38	72.12	82.56	77.68
		Gemini-2.0 Flash [20]	86.24	74.71	65.88	55.72	80.91	58.62	64.48	86.42	87.93	73.43
GPT 4o [3]		78.68	68.52	78.93	76.74	75.35	91.32	83.41	72.28	79.38	78.29	
Chain-of-Thought (CoT)	O-Mono	BLIP-2 OPT 6.7B	34.45	35.62	30.05	31.67	36.74	31.08	24.62	42.41	31.66	33.14
		LLaVa-1.5 7B	36.38	35.72	27.46	35.51	44.05	38.77	23.28	38.31	40.14	35.51
		LLaVa-Next 7B	43.08	38.27	42.44	38.31	30.53	28.80	60.10	25.84	46.57	39.33
	O-Multi	LLaMa-3.2-V 11B	65.05	68.31	58.27	63.32	76.26	78.28	80.04	65.03	76.27	70.09
		Phi-3.5-V	36.72	30.02	23.28	18.52	28.32	25.10	30.48	33.28	35.20	28.99
		Phi-4 Multimodal	47.49	48.22	49.94	23.24	56.64	48.28	46.72	48.22	44.92	45.96
		Qwen-2.5 7B	61.68	61.64	60.10	60.15	68.34	73.28	73.38	71.70	71.73	66.89
		Gemma-3 12B	75.04	88.28	73.34	75.06	90.04	75.02	83.25	78.32	81.69	80.00
	Closed	Claude-3.5 Sonnet	84.78	78.23	85.04	75.78	80.56	81.78	84.10	76.04	84.77	81.23
		Gemini-2.0 Flash	87.14	79.32	73.89	58.04	86.22	62.23	67.55	89.57	90.42	77.15
GPT 4o		80.18	84.43	82.62	78.34	81.48	83.44	82.56	75.10	86.67	81.65	
Translation-based	O-Mono	BLIP-2 OPT 6.7B	37.04	34.78	33.45	45.15	34.22	48.12	40.20	35.76	42.57	39.03
		LLaVa-1.5 7B	25.00	38.60	28.81	35.59	26.79	41.67	49.15	47.27	37.93	36.76
		LLaVa-Next 7B	33.33	35.00	37.29	31.67	25.42	36.67	38.33	31.67	33.33	33.63
	O-Multi	LLaMa-3.2-V 11B	62.71	75.00	66.10	65.00	74.58	63.33	70.00	63.33	85.00	69.45
		Phi-3.5-V	28.33	35.59	37.29	20.00	18.33	28.33	40.00	21.67	35.00	29.39
		Phi-4 Multimodal	55.17	45.00	45.76	42.37	56.67	51.67	48.33	60.00	51.67	50.74
		Qwen-2.5 7B	56.67	66.67	58.33	63.33	78.33	65.00	70.00	66.67	75.00	66.67
		Gemma-3 12B	71.67	85.00	75.00	76.27	86.67	76.67	83.33	85.00	85.00	80.52
	Closed	Claude-3.5 Sonnet	85.91	78.00	85.45	77.20	75.52	85.42	84.38	76.67	84.04	81.40
		Gemini-2.0 Flash	88.04	78.56	71.15	57.62	85.14	60.48	67.90	90.22	91.44	76.73
GPT 4o		87.52	77.62	80.30	79.42	85.25	84.38	84.42	78.48	87.41	82.75	
Culture-specific	O-Mono	BLIP-2 OPT 6.7B	33.92	32.29	30.14	42.64	31.32	44.63	37.22	32.90	39.66	36.08
		LLaVa-1.5 7B	45.00	38.33	30.51	27.12	30.00	23.73	35.59	36.67	40.00	34.11
		LLaVa-Next 7B	28.33	36.67	33.33	30.00	25.42	31.03	41.38	30.00	31.67	32.98
	O-Multi	LLaMa-3.2-V 11B	60.00	70.00	66.67	78.33	71.67	75.00	71.67	75.00	86.67	72.78
		Phi-3.5-V	37.93	40.68	31.48	24.14	32.76	38.98	35.71	29.82	27.59	33.23
		Phi-4 Multimodal	53.45	45.76	53.33	38.60	56.67	55.93	48.21	52.54	40.68	49.46
		Qwen-2.5 7B	58.62	66.10	55.93	56.90	72.88	61.67	69.49	70.00	80.00	65.73
		Gemma-3 12B	76.67	85.00	78.33	75.00	86.67	75.00	83.33	78.33	80.00	79.81
	Closed	Claude-3.5 Sonnet	86.62	79.23	87.32	77.67	78.14	87.11	85.78	78.20	86.15	82.91
		Gemini-2.0 Flash	90.52	80.48	74.60	58.78	87.04	63.67	70.42	91.05	92.88	78.83
GPT 4o		81.92	86.71	84.23	80.24	84.05	83.10	85.30	75.94	89.25	83.42	
Description-based	O-Mono	BLIP-2 OPT 6.7B	36.70	34.62	32.78	45.42	33.23	47.53	39.64	35.32	42.05	38.59
		LLaVa-1.5 7B	35.59	29.31	30.51	30.51	33.33	32.14	38.98	48.33	40.00	35.41
		LLaVa-Next 7B	31.67	42.37	31.03	29.31	28.57	36.67	37.29	40.68	50.00	36.40
	O-Multi	LLaMa-3.2-V 11B	62.71	77.59	66.10	67.80	79.66	66.10	71.43	70.69	83.05	71.68
		Phi-3.5-V	33.33	46.55	24.56	33.90	31.58	38.33	30.51	30.00	37.29	34.01
		Phi-4 Multimodal	54.39	44.07	55.93	37.93	56.67	50.00	46.67	50.00	49.15	49.42
		Qwen-2.5 7B	55.00	65.52	50.85	60.34	82.14	64.41	61.11	72.88	77.59	65.54
		Gemma-3 12B	70.00	86.67	75.00	76.67	83.33	75.00	83.33	78.33	80.00	78.70
	Closed	Claude-3.5 Sonnet	84.74	81.56	85.80	78.15	76.62	88.10	86.80	74.23	84.89	82.32
		Gemini-2.0 Flash	89.90	81.05	69.13	54.32	85.14	58.72	63.66	89.94	94.25	76.23
GPT 4o		80.55	75.52	80.72	80.83	84.85	86.18	84.74	66.70	85.68	80.64	

Table 3. Model benchmarking results across different **Prompting Strategies** on the test split of BANGLAPROTHA. *O-Mono* and *O-Multi* refer to the open-source monolingual and multilingual models, respectively, while *Closed* refers to the closed-source models. **Cyan** highlights the highest score for each cultural concept, and **Bold** indicates the overall best-performing model.

Models	Cultural Concepts									
	Food	Fest	Rel	Nature	Fash	Sport	Life	Art/Hs	Incid	Avg
<b>Full Fine-Tuning</b>										
BanglaBERT [12] + ViT [5]	33.24	33.27	37.18	36.08	31.36	38.14	40.17	33.82	32.21	35.05
BanglaBERT [12] + Swin [34]	35.45	36.25	38.78	37.20	32.45	40.17	41.27	35.92	33.64	36.79
CLIP [50]	31.25	32.90	36.78	34.64	28.62	33.15	34.70	32.34	30.48	32.76
LXMERT [55]	34.38	35.89	39.67	37.22	33.52	36.10	37.93	34.87	32.90	35.83
ALIGN [27]	38.91	39.83	41.62	40.63	35.44	43.39	44.72	39.60	35.84	40.00
SmolVLM2 [38]	32.34	48.09	42.56	41.65	45.76	41.87	41.00	52.48	55.89	44.63
Intern-VL3-2B [16]	65.12	40.24	45.08	53.36	45.13	61.60	56.58	56.53	46.67	<b>52.22</b>
<b>Partial Fine-Tuning</b>										
<b>LLaVa-1.5 7B [30]</b>										
L-LoRA	56.67	55.00	55.00	45.00	48.33	50.00	68.33	46.67	60.00	53.89
L-LoRA + Adapter	63.33	51.67	50.00	53.33	50.00	58.33	45.00	51.67	56.67	53.33
L-LoRA + Adapter + VE	66.67	58.33	36.67	56.67	53.33	60.00	60.00	63.33	73.33	<b>58.70</b>
<b>Qwen2.5-VL 7B [58]</b>										
L-LoRA	60.00	76.67	63.33	66.67	78.33	76.67	76.67	78.33	76.67	72.59
L-LoRA + Adapter	68.33	76.67	61.67	66.67	71.67	78.33	70.00	80.00	83.33	72.96
L-LoRA + Adapter + VE	65.00	70.00	63.33	75.00	71.67	81.67	80.00	85.00	78.33	<b>74.44</b>
<b>Paligemma-2 10B [10]</b>										
L-LoRA	73.33	78.33	66.67	66.67	86.67	75.00	70.00	70.00	68.33	72.78
L-LoRA + Adapter	81.67	76.67	63.33	80.00	81.67	76.67	70.00	81.67	76.67	<b>76.48</b>
L-LoRA + Adapter + VE	68.33	83.33	70.00	65.00	81.67	73.33	66.67	76.67	78.33	73.70
<b>LLaMa-3.2V 11B [41]</b>										
L-LoRA	63.33	76.67	61.67	63.33	75.00	71.67	73.33	71.67	76.67	70.37
L-LoRA + Adapter	81.67	70.00	68.33	60.00	78.33	71.67	75.00	73.33	81.67	<b>73.33</b>
L-LoRA + Adapter + VE	56.67	71.67	61.67	66.67	70.00	66.67	65.00	75.00	76.67	67.78
<b>Gemma-3 12B [56]</b>										
L-LoRA	80.00	81.67	75.00	68.33	86.67	68.33	71.67	81.67	83.33	77.41
L-LoRA + Adapter	81.67	83.33	70.00	71.67	85.00	78.33	73.33	83.33	76.67	78.15
L-LoRA + Adapter + VE	82.22	92.12	80.64	81.90	88.45	84.52	75.72	85.13	86.44	<b>84.13</b>

Table 4. Model benchmarking results across different *Finetuning Strategies* on the test split of BANGLAPROTHA. *O-Mono* and *O-Multi* refer to the open-source monolingual and multilingual models, respectively, while *Closed* refers to the closed-source models. **Cyan** highlights the highest score for each cultural concept, and **Bold** indicates the overall best-performing model in each category.

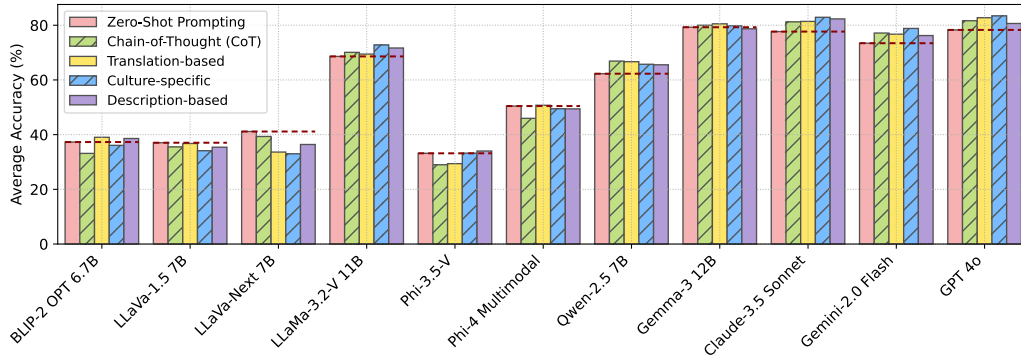
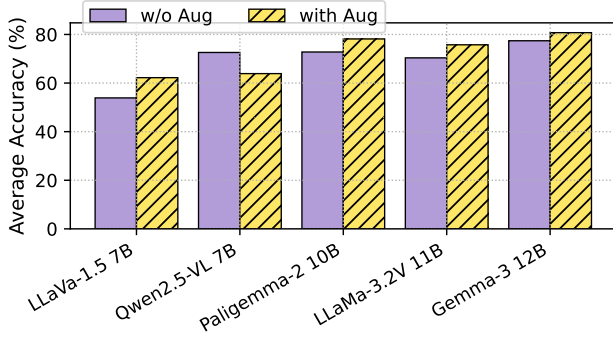


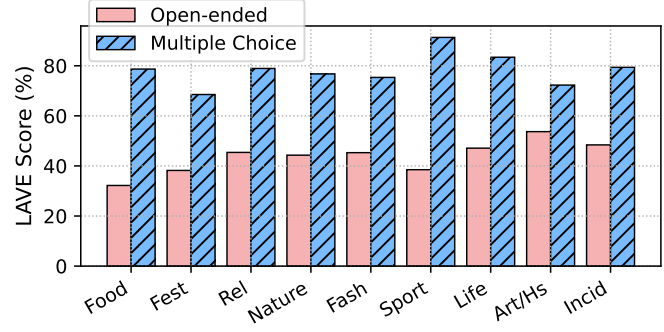
Figure 3. Average accuracy of models across prompting strategies on the test split of BANGLAPROTHA.

Gemma-3 12B, and Qwen 2.5 7B, achieve the highest performance gains under culture-specific prompting, emphasizing the need for culturally tailored instructions for reasoning and generating culturally relevant responses. Zero-shot prompting, however, remains the weakest of the

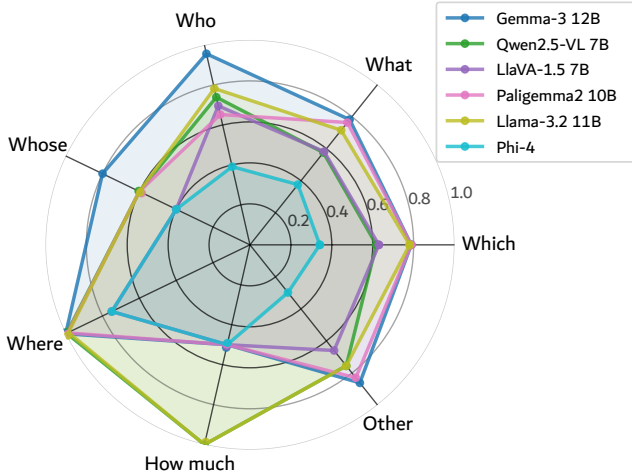
prompting strategies. In sharp contrast, zero-shot prompting consistently outperforms the other strategies for the smaller open-source models, suggesting that the additional prompting instructions tend to overwhelm the limited capacity of the smaller architectures.



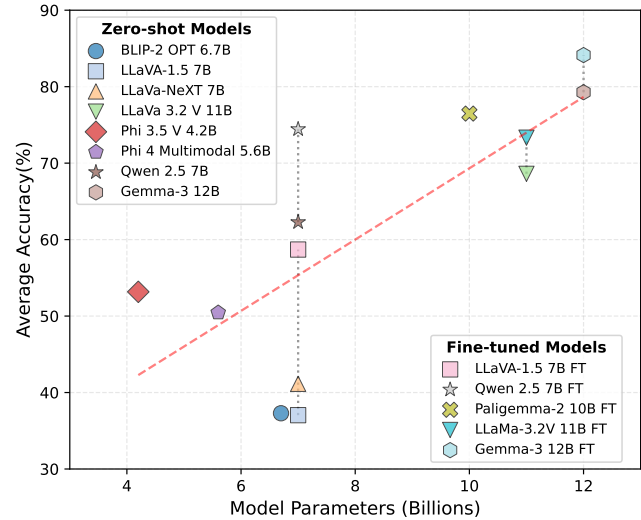
(a)



(b)



(c)



(d)

Figure 4. (a) L-LoRA accuracy of models with vs. without data augmentation, (b) LAVE scores of open-ended vs. multiple-choice questions across cultural concepts for GPT-4o, (c) Accuracy of fine-tuned models across question types, and (d) Average accuracy vs. model parameters of zero-shot and fine-tuned models (using the best-performing strategy), with the accuracy trendline across parameters (in red).

**Prompting vs. Fine-tuning.** Following Tabs. 3 and 4, fully fine-tuned dual encoder and modality alignment models achieved performance comparable to the monolingual open-source models ( $\sim 40\%$  average accuracy). Partial fine-tuning showed a substantial bump in performance over prompting strategies, *e.g.*, the best fine-tuning strategy for LLaVa-1.5 7B outperformed its best prompting strategy by 21.66%. Similarly, fine-tuned Gemma-3 12B achieved the highest average accuracy, 84.13%, on our dataset, slightly surpassing GPT-4o using culture-specific prompting.

**What’s the best way to finetune?** Tab. 4 highlights ALIGN outperforming other fully fine-tuned models across all aspects, but lagging behind the larger partially fine-tuned LLMs. Fine-tuning the language component using L-LoRA usually improved the performance, *e.g.*, for LLaVa-1.5 by roughly 17% vs. its best prompting strategy. However, we observed several instances of performance drop, *e.g.*,

Gemma-3’s average accuracy dropped roughly 3% vs. its best prompting strategy.

**Should we fine-tune the visual component?** Fine-tuning the visual components, *i.e.*, adapters and vision encoders, generally leads to improved performance (Tab. 4). The best results are usually achieved when *both*, visual and textual, components are fine-tuned, improving the alignment with the underrepresented Bengali cultural images. However, models, such as Paligemma-2 10B and LLaMa-3.2V 11B, underperform when adapters and vision encoders are fine-tuned, likely due to overfitting. Adapter fine-tuning tends to outperform applying L-LoRA fine-tuning only, though occasionally leading to a marginal performance decline (*e.g.*, -0.65% accuracy drop for LLaVa-1.5 7B).

**Performance variation across Cultural Aspects.** Most models tend to perform better on Fashion, Social

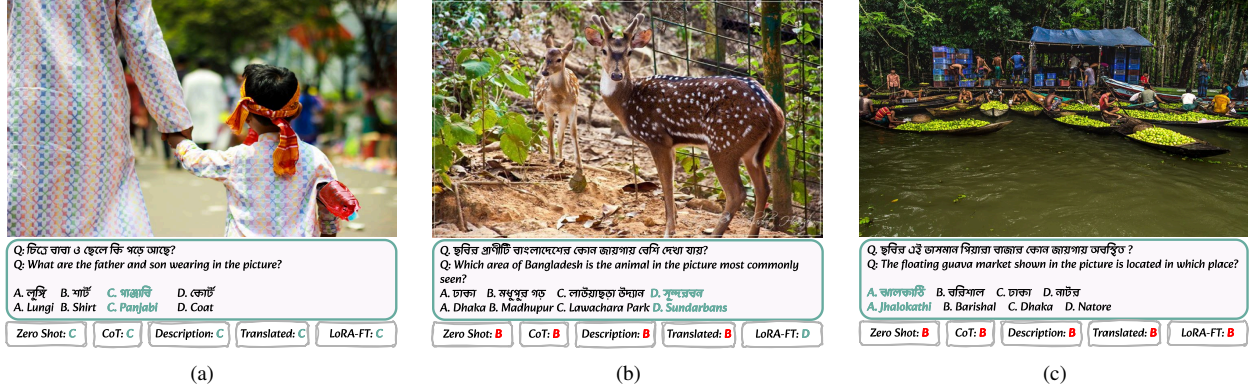


Figure 5. Performance comparison across evaluation methods of Gemma-3 12B: (a) unanimous correct predictions, (b) only LoRA-FT predicts correctly, and (c) unanimous incorrect predictions. Red and Green highlight the incorrect and correct predictions respectively.

Life, and Incidents categories, while struggling with Religious and Nature Fig. (§A.3b). Under zero-shot setting, GPT-4o excels in Sports, Gemini-2.0 in Art/Hs, Incidents, and Food, Gemma-3 in Fashion and Festivals, and Claude-3.5 in Religion. We see similar performance strengths of Gemini-2.0 and Gemma-3 on Festival, Fashion, and Art/Hs, likely due to overlap in training corpora as the models share the same parent company. Gemma-3, using L-LoRA + Adapter + VE, consistently outperforms most fine-tuned models across all cultural aspects.

**Performance across Question Types.** Following Fig. 4c, the fine-tuned Llama-3.2 11B performs better on *How much* questions, while Gemma-3 12B leads on rest of the question types. Similarly, Fig. (§A.3a) shows the best prompting model, GPT-4o, performing better on *How much* (93.4%), *When* (90.6%), and *Where* (87.2%) questions, demonstrating the model’s excellence on qualitative, temporal, and spatial reasoning. In contrast, the lowest performance was observed for *How Many* (66.5%) questions, exposing the model’s weakness in counting.

**Impact of Data Augmentation.** We applied LLaVA-style augmentation [30] during L-LoRA finetuning by augmenting each question-answer(QA) pair  $k = 4$  times,  $k$  representing the number of options per question, resulting in  $6.5k \times 4 = 26k$  samples (detailed in §D.4). This compensates for the scarcity of multiple-choice variations by shuffling the answer options to create new training samples. From Fig. 4a, we observe a 3-9% boost in average accuracy across all models except Qwen-2.5 (cultural aspect-wise breakdown in Tab. §A.4). As the augmentation only permutes the position of the correct option, the results expose a positional textual bias in the models. Similar experiments were conducted using the circular evaluation strategy [33], reported in §E.3 and Tab. A.8.

**What if we remove answer options?** We investigate this by evaluating GPT-4o in an open-ended setting using the LAVE metric [36] (details in §E.1). In this setup, GPT-4o attained a LAVE score of 43.68%, a significant drop from its 81.64% accuracy in the multiple-choice format. Fig. 4b shows performance dropping notably for Food, Festival, and Sport categories. Similar experiments were conducted on open-source VLMs (Tab. A.6), with models exhibiting a consistent decline in performance under the open-ended setting.

**Error Analysis.** Fig. 5 shows predictions of the best performing Gemma-3 12B model across prompting and fine-tuning strategies. In Fig. 5a, all evaluation settings correctly identify the traditional attire (*Panjabi*). In Fig. 5b, only the LoRA-FT setting correctly recognizes the correct habitat of the animal. In Fig. 5c, all settings mispredict the location of the floating guava market, which requires a high level of cultural knowledge even for Bengali natives. While some classes of culturally-grounded visual questions can be handled by the fine-tuned model, others that require in-depth knowledge remain challenging. §F expands qualitative error analysis across models and cultural concepts.

## 6. Conclusion

We introduced BANGLAPROTHA, the first Bengali culturally grounded Visual Question Answering dataset encompassing nine diverse cultural domains. Through comprehensive experiments with both open- and closed-source, monolingual and multilingual VLMs under various fine-tuning and prompting setups, we provide valuable insights into the current capabilities and limitations of Bengali multimodal understanding. We hope our dataset and findings will foster future research toward more culturally aware and linguistically inclusive vision-language models.

## References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. [5, 3, 8](#)
- [2] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. [5, 3, 8](#)
- [3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. [5, 3](#)
- [4] Hasan Al Faraby, Md Muzahidul Azad, Md Riduyan Fedous, Md Kishor Morol, et al. Image to bengali caption generation using deep cnn and bidirectional gated recurrent unit. In *2020 23rd international conference on computer and information technology (ICCIT)*, pages 1–6. IEEE, 2020. [3](#)
- [5] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. [6, 3](#)
- [6] Anthropic. Model card addendum for Claude 3. [https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model\\_Card\\_Claude\\_3\\_Addendum.pdf](https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf), 2024. Accessed: 2025-06-04. [5, 3](#)
- [7] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. [1, 2](#)
- [8] Deeparghya Dutta Barua, Md Sakib Ul Rahman Sourove, Md Fahim, Fabiha Haider, Fariha Tanjim Shifat, Md Tas-mim Rahman Adib, Anam Borhan Uddin, Md Farhan Ishmam, and Md Farhad Alam. Chitrojera: A regionally relevant visual question answering dataset for bangla. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 473–491. Springer, 2025. [2, 4](#)
- [9] Federico Becattini, Pietro Bongini, Luana Bulla, Alberto Del Bimbo, Ludovica Marinucci, Misael Mongiovì, and Valentina Presutti. Viscounth: a large-scale multilingual visual question answering dataset for cultural heritage. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6):1–20, 2023. [1](#)
- [10] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. [1, 6, 3](#)
- [11] Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. *arXiv preprint arXiv:2407.00263*, 2024. [2](#)
- [12] Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States, July 2022. Association for Computational Linguistics. [6, 3](#)
- [13] Md Shalha Mucha Bhuyan, Eftekar Hossain, Khaleda Akhter Sathi, Md Azad Hossain, and M Ali Akber Dewan. Bvqa: Connecting language and vision through multimodal attention for open-ended question answering. *IEEE Access*, 2025. [2](#)
- [14] Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*, 2023. [2](#)
- [15] Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. Maxm: Towards multilingual visual question answering. *arXiv preprint arXiv:2209.05401*, 2022. [2](#)
- [16] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. [6](#)
- [17] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*, 2016. [2](#)
- [18] Mohamed Atta Faheem, Khaled Tawfik Wassif, Hanaa Bayomi, and Sherif Mahdy Abdou. Improving neural machine translation for low resource languages through non-parallel corpora: a case study of egyptian dialect to modern standard arabic translation. *Scientific Reports*, 14(1):2265, 2024. [2](#)
- [19] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28, 2015. [2](#)
- [20] Google DeepMind and Sundar Pichai. Introducing Gemini 2.0: Our new AI model for the agentic era. <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/>, December 2024. Accessed: 2025-06-04. [5, 3](#)
- [21] 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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. [2](#)
- [22] Deepak Gupta, Pabitra Lenka, Asif Ekbal, and Pushpak Bhattacharyya. A unified framework for multilingual and code-mixed visual question answering. In *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing*, pages 900–913, 2020. [2](#)
- [23] Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Md Sakib Ul Rahman Sourove, Deeparghya Dutta Barua,

- Md Fahim, and Md Farhad Alam Bhuiyan. BanTH: A multi-label hate speech detection dataset for transliterated Bangla. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7217–7236, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. 4
- [24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1, 4, 3
- [25] Md Farhan Ishmam, Md Sakib Hossain Shovon, Muhammad Firoz Mridha, and Nilanjan Dey. From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. *Information Fusion*, 106:102270, 2024. 1
- [26] SM Shahriar Islam, Riyad Ahsan Auntor, Minhajul Islam, Mohammad Yousuf Hossain Anik, ABM Alim Al Islam, and Jannatun Noor. Note: Towards devising an efficient vqa in the bengali language. In *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, pages 632–637, 2022. 2
- [27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 6, 3
- [28] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 1, 2
- [29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 5, 3
- [30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 5, 6, 8, 3, 9
- [31] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. 5, 3
- [32] Shudong Liu, Yiqiao Jin, Cheng Li, Derek F Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. CultureVLM: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv preprint arXiv:2501.01282*, 2025. 2, 4
- [33] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 8, 6
- [34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6, 3
- [35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3
- [36] Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4171–4179, 2024. 8
- [37] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 1
- [38] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 6
- [39] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3190–3199, 06 2019. 2, 4
- [40] Raphael Merx, Adérito José Guterres Correia, Hanna Suominen, and Ekaterina Vylomova. Low-resource machine translation: what for? who for? an observational study on a dedicated tetun language translation service. In Atul Kr. Ojha, Chao-hong Liu, Ekaterina Vylomova, Flammie Pirinen, Jonathan Washington, Nathaniel Oco, and Xiaobing Zhao, editors, *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 54–65, Albuquerque, New Mexico, U.S.A., May 2025. Association for Computational Linguistics. 2
- [41] Meta AI. Llama 3: Connecting the next generation of ai with vision, edge, and mobile devices at Connect 2024. <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>, May 2024. Accessed: 2025-06-04. 5, 6, 3, 8, 9
- [42] Faisal Muhammad Shah, Mayeesha Humaira, Md Abidur Rahman Khan Jim, Amit Saha Ami, and Shimul Paul. Bornon: Bengali image captioning with transformer-based deep learning approach. *SN Computer Science*, 3:1–16, 2022. 3
- [43] Ghulam Murshid. *Bengali culture over a thousand years*. Niyogi Books, 2018. 1
- [44] Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*, 2023. 2
- [45] 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. *arXiv preprint arXiv:2407.10920*, 2024. 1, 2

- [46] Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. Jmmmu: A japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. *arXiv preprint arXiv:2410.17250*, 2024. [1](#)
- [47] Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12988–12997, 2024. [1](#)
- [48] ChaeHun Park, Yujin Baek, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. Evaluating visual and cultural interpretation: The k-viscuit benchmark with human-vm collaboration. *arXiv preprint arXiv:2406.16469*, 2024. [5](#)
- [49] Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. xgqa: Cross-lingual visual question answering. *arXiv preprint arXiv:2109.06082*, 2021. [2](#)
- [50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. [1](#), [6](#), [3](#)
- [51] Mahamudul Hasan Rafi, Shifat Islam, SM Hasan Imtiaz Labib, SM Sajid Hasan, Faisal Muhammad Shah, and Sifat Ahmed. A deep learning-based bengali visual question answering system. In *2022 25th International Conference on Computer and Information Technology (ICCIT)*, pages 114–119. IEEE, 2022. [2](#)
- [52] Matiur Rahman, Nabeel Mohammed, Nafees Mansoor, and Sifat Momen. Chitron: An automatic bangla image captioning system. *Procedia Computer Science*, 154:636–642, 2019. [1](#), [2](#), [3](#)
- [53] David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*, 2024. [1](#), [2](#)
- [54] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge, 2022. [4](#)
- [55] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. [6](#), [3](#)
- [56] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. [5](#), [6](#), [3](#), [8](#), [9](#)
- [57] Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadgign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, et al. All languages matter: Evaluating llms on culturally diverse 100 languages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19565–19575, 2025. [2](#)
- [58] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. [5](#), [6](#), [3](#), [8](#), [9](#)
- [59] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. [4](#)
- [60] Thomas Wolf, Lysandre Debut, Victor Sanphih, 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. 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, Oct. 2020. Association for Computational Linguistics. [3](#)
- [61] Pengju Xu, Yan Wang, Shuyuan Zhang, Xuan Zhou, Xin Li, Yue Yuan, Fengzhao Li, Shunyu Zhou, Xingyu Wang, Yi Zhang, et al. Tcc-bench: Benchmarking the traditional chinese culture understanding capabilities of mllms. *arXiv preprint arXiv:2505.11275*, 2025. [1](#)
- [62] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024. [3](#)

# BANGLAPROTHA: Evaluating Vision Language Models in Underrepresented Long-tail Cultural Contexts

## Supplementary Material

### A. Annotation Guidelines

The annotation guidelines include the Question and Option Creation Guidelines (§A.1) and Categorization Guidelines (§A.2). For the ease and convenience of the annotators, we have kept the annotation guidelines concise to the essentials, eliminating redundant details. For further clarification or context, the annotators may consult other materials, *e.g.*, formal category definitions (§C).

#### A.1. Question and Option Creation Guidelines

As an annotator, you need to compose clear, meaningful, and culturally relevant questions and corresponding answer choices based on the provided images. Ensure that the questions are unambiguous and align with the theme of the dataset. For each question, provide four closely related answer options. Please follow the guidelines below to formulate the questions and answer choices:

1. Choose images that authentically reflect the everyday life, traditions, or notable customs of Bengali-speaking communities.
2. Write all questions in Bengali script to maintain linguistic authenticity.
3. Ensure each question is directly related to a cultural element of the given image and reflects Bengali tradition, lifestyle, and customs.
4. Avoid vague or general questions that do not refer to specific visual elements.
5. Keep the question clear and concise. Avoid complex sentence structures or ambiguous wording.
6. Provide exactly four options: one unequivocally correct answer and three plausible—but incorrect distractors. The distractors must be logically related to the question.
7. Match the length and grammatical structure of all options to avoid inadvertent cues.
8. Maintain a uniform tone and phrasing style to preserve coherence throughout the dataset.
9. If you are unsure about a specific annotation, refer to the provided sample annotation document for examples and further clarification.

#### A.2. Categorization Guidelines

Each question should be assigned to a relevant category to ensure the dataset remains well-structured. Categories help organize the data and improve its usefulness for future analysis. Carefully review the content of the question before selecting the most appropriate category. The following nine categories should be used:

- *Food* – Includes questions about traditional dishes, local cuisine, *e.g.*, Biryani, and food-related customs.
- *Cultural Festivals* – Covers national and regional celebrations, *e.g.*, Pohela Boishakh, Bijoy Dibosh.
- *Religious Events* – Related to religious celebrations, *e.g.*, Eid, Puja, Christmas, and associated customs.
- *Nature* – Questions about rivers, landscapes, weather, and other natural elements specific to Bengal.
- *Clothing and Fashion* – Includes traditional attire, textile heritage, and fashion trends in Bengal.
- *Sports* – Covers games and physical activities commonly played in the region, *e.g.*, kabaddi, swimming.
- *Social Life* – Focuses on daily life, social interactions, and community activities of the Bengalis.
- *Arts and History* – Questions about Bengali paintings, landmarks, historical events, and architectural heritage.
- *Incidents* – Recorded Bengali historical events, *e.g.* election day, landmark declarations, and public events, *e.g.*, river erosion.

#### A.3. Annotation Tool

To facilitate the annotation process, we developed an annotation tool using Streamlit<sup>2</sup>. We provided live, hands-on tutorials to the annotators and detailed software documentation. Fig. A.1 illustrates the tool’s interface and workflow. The annotation process begins with annotators specifying the path of the image directory. To support continuous annotation sessions, the tool allows annotators to resume their previous work by uploading a `progress.json` file. At any point during the annotation process, annotators can save their progress by downloading both `progress.json` and `relevant.json` files.

<sup>2</sup><https://streamlit.io>

**Image Viewer Controls**

Folder path:

Upload the progress JSON file to resume annotation

Drag and drop file here  
Limit 200MB per file • JSON

Browse files

test.json  
3.5KB

Upload the Relevant JSON file

Drag and drop file here  
Limit 200MB per file • JSON

Browse files

Download Progress JSON

Download Relevant JSON

Insert Question

নদীর ধারে ভিড়ানো ডিসি লৌকা গুলোকে চালাতে মাঝিরা কি ব্যবহার করেন?

Option A

মোটর

Option B

বাঁশ

Option C

মেশিন

Option D

পাল

Category

Social Life

Answer

মোটর

মোটর

বাঁশ

মেশিন

পাল

Figure A.1. Demo of our annotation tool.

For each image, annotators identified an image as either *Relevant* or *Irrelevant*. For *Relevant* images, annotators are required to: formulate a question based on the visual content, create four multiple-choice options, select the correct answer from these options, and identify the corresponding cultural concept represented by the sample. Upon completing these steps for a relevant image, annotators can save the data and proceed to the next image. Images classified as *Irrelevant* are automatically skipped.

## B. Data Collection & IAA Score

Source	License
Unsplash	Royalty-free commercial license
Pixabay	CC0 (Public Domain)
Wikimedia Commons	CC0 (Public Domain)
Rawpixel	CC0 (Public Domain)
Pexels	Royalty-free commercial license

Table A.1. Source name and license of the websites used for image collection.

We have collected a total of 2,800 publicly available Bengali images from various online sources. The websites

used for image collection, along with their corresponding license information, are listed in Tab. A.1. To assess the annotation quality, we also evaluated the Inter-Annotator-Agreement (IAA) score. For this, we randomly selected 450 samples (50 samples for each cultural concept) from our dataset. We instructed the annotator to pick the best option for a given image and the options. Each sample is annotated by three random annotators. Then we measure the Kappa Score as our IAA score and report it in Tab. A.2 for various cultural concepts. The results indicate a consistently high level of agreement across all categories, with an average  $\kappa$  of 0.848. This suggests strong reliability and consistency in the annotation process.

## C. Category Definitions

While §A.2 provides informal definitions and examples of each category for the annotators, in this section, we *formally* define each category to capture distinct concepts of Bengali culture and heritage.

- **Cultural Festival:** Celebrations rooted in Bengali tradition, and seasonal cycles, e.g., Pohela Falgun (First Day of Spring), Pohela Boishakh (Bengali New Year), and regional folk festivals.

Cultural Concept	Kappa
Food	0.812
Fest	0.879
Rel	0.844
Nature	0.899
Fash	0.861
Sport	0.785
Life	0.837
Art/History	0.893
Incident	0.826
Avg	0.848

Table A.2. Inter-Annotator Agreement (Kappa) for Cultural Concepts

- **Food:** Local cuisine and culinary items of the Bengal region, *e.g.*, rice, fish, lentils, and sweets like Rasgulla, Sandesh, and Mishti Doi.
- **Nature:** Landscape of the Bengali region including rich plains, meandering rivers, and lush green scenes, *e.g.*, scenes of farmers working in the mustard field.
- **Clothing & Fashion:** The style of dress and adornment of the Bengalis, *e.g.* handwoven saree, panjabi, and embroidery.
- **Social Life:** The interactions, communal gatherings, and shared activities of Bengalis, *e.g.*, women drying rice after harvesting.
- **Arts & History:** From Tagore’s literature and folk art to Mughal, Pal–Sen, and British colonial architecture.
- **Religious Events:** Major Bengali religious observances, *e.g.*, Eid, Durga Puja, and Christmas.
- **Sports:** Popular pastimes and recreational activities of Bengalis, *e.g.*, cricket, football, and traditional games like ha-du-du, bouci, and nouka baich.
- **Incidents:** Significant public events of Bengal, *e.g.*, natural disasters, river erosion, urban challenges, showcasing societal impact and the human dimension of change. Includes historical milestones, *e.g.*, election day occurrences, landmark declarations.

## D. Experimental Setup

### D.1. Models

We evaluate open-source monolingual (English-centric) and multilingual VLMs in our prompt-based and partial fine-tuning experiments. For the monolingual models, we include BLIP-2 [29] using OPT-6.7B<sup>3</sup>, LLaVA-1.5-7B [30],

<sup>3</sup><https://huggingface.co/advaitadasein/blip2-opt-6.7b>

and LLaVA-Next-Mistral-7B<sup>4</sup> [31]. For the multilingual open-source VLMs, our evaluation includes LLaMA-3.2-11B-Vision-Instruct [41], Phi-4-Multimodal-Instruct [2], Phi-3.5-Vision-Instruct [1], Gemma-3-12B [56], Qwen2.5-VL-7B-Instruct [58], and Paligemma2-10B [10]. In addition, we conduct evaluations using closed-source large VLMs, including GPT-4o [3], Claude-3.5 Sonnet [6], and Gemini-2.0 Flash [20].

For full fine-tuning, we evaluate dual encoder models: BanglaBERT [12] + ViT [5] and BanglaBERT + Swin [34]. Since all the text is in Bengali, BanglaBERT is used for textual encoding. We also fine-tune cross-modal models: CLIP [50], LXMERT [55], and ALIGN [27]. We fuse visual and textual representations through concatenation for the dual encoder and CLIP models.

### D.2. Full Fine-tuning Setting

For the full fine-tuning experiment, we used the Huggingface library [60] to load the pretrained weights and the PyTorch framework<sup>5</sup> for model training. The AdamW optimizer [35] was employed with a learning rate of 2e-5 and a batch size of 32. For the dual encoder and CLIP models, the visual and textual representations were concatenated, passed through a two-layer MLP with ReLU activation, and a linear layer for classification. Each option, along with the question and image, was passed through the model, and the results were stacked to calculate the cross-entropy loss. The models were fine-tuned until convergence.

### D.3. Partial Fine-tuning Setting

For all partial fine-tuning experiments on open-source Vision-Language Models (VLMs), we use the LLaMaFactory framework [62], known for its efficiency in fine-tuning VLMs. For LoRA fine-tuning [24] of the LLMs, we use a LoRA adapter with a rank of 64 and set alpha to 64. We also apply a LORA dropout rate of 0.05 to prevent overfitting. The training configuration includes a batch size of 32, a learning rate of 2e-4, and a training duration of one epoch.

All the hyperparameters are provided in D.5

### D.4. Data Augmentation Setup

Let each datum be a triplet  $(I, Q, O)$ , where  $I$  is an image,  $Q$  its associated question, and  $O = \{a_1, a_2, \dots, a_k\}$  is the set of  $k$  candidate answers, among which the correct answer is  $a^* \in O$ . In our multiple-choice VQA setting, we fix  $k = 4$ . To mitigate the paucity of native multiple-choice examples, we synthesize additional training samples by permuting the position of the correct answer within  $O$ .

<sup>4</sup><https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf>

<sup>5</sup><https://pytorch.org/>

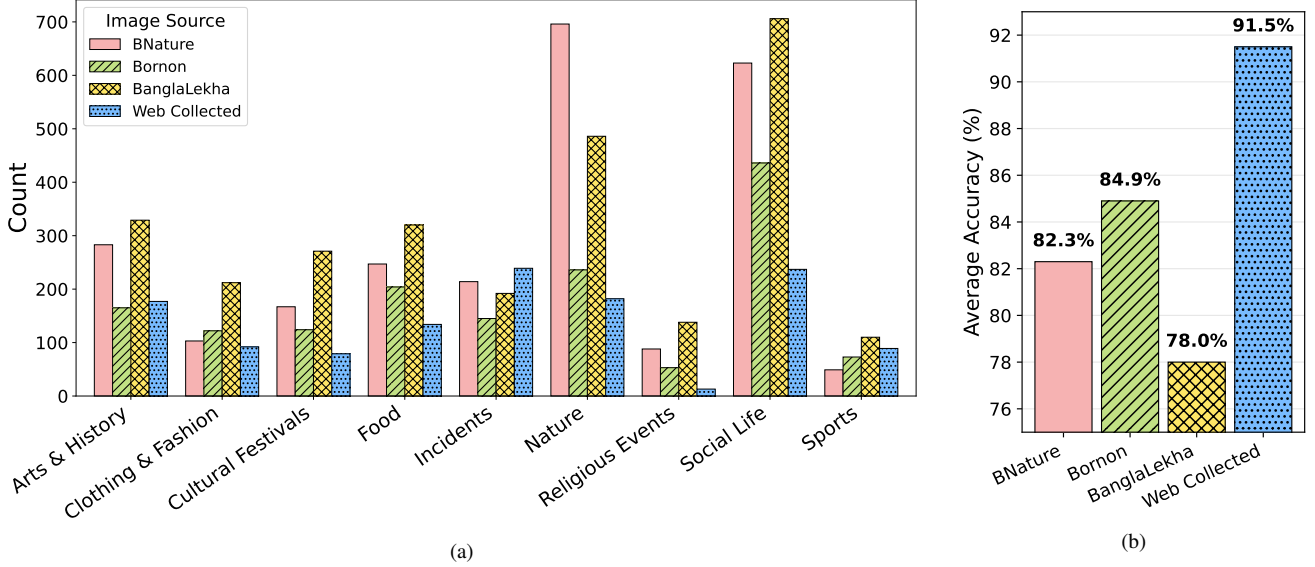


Figure A.2. (a) Distribution of image sources of BANGLAPROTHA across cultural aspects, (b) Performance of GPT-4o model using zero-shot prompting across image sources. Performance can be correlated to the quality of images from that source.

Category	Count				Ratio		Percentage	
	#Img	#Ques	#Ans	#Opt	#Q:#I	#O:#A	Img (%)	Ques (%)
Social Life	1900	2002	1802	4025	1.05	2.23	27.37	24.92
Nature	1400	1600	1312	3212	1.14	2.45	20.17	19.92
Arts & History	754	954	698	1482	1.27	2.12	10.86	11.87
Food	705	905	685	1681	1.28	2.45	10.16	11.26
Incidents	690	790	690	1924	1.14	2.79	09.94	09.83
Cultural Festivals	541	641	589	1315	1.18	2.23	07.79	07.98
Religious Events	252	292	234	601	1.16	2.57	03.63	03.63
Clothing & Fashion	429	529	489	1412	1.23	2.89	06.18	06.58
Sports	271	321	287	673	1.18	2.34	03.90	04.00
<b>Total</b>	6942	8034	6786	16325	1.16	2.41	100.00	100.00

Table A.3. Category-wise statistics: (i) counts of unique images, questions, answers, and answer options, (ii) question-to-image and option-to-answer ratio; higher indicates more uniqueness in the dataset, and (iii) image and question composition.

Concretely, we construct,

$$O^{(j)} = \{a_1, \dots, a_{j-1}, a^*, a_{j+1}, \dots, a_k\}$$

for each  $j = 1, \dots, k$ , where in the  $j$ -th permutation the correct answer  $a^*$  occupies the  $j$ -th slot. Each  $(I, Q, O^{(j)})$  defines a distinct training sample whose label remains  $a^*$ . Thus, every original datum yields exactly  $k$  augmented instances. If the dataset contains  $N$  original samples, this procedure expands it to,  $N_{\text{aug}} = N \times k = 4N$  total training examples, when  $k = 4$ . By exhaustively permuting the correct-answer position, we ensure uniform coverage of all answer slots, directly exposing and alleviating any positional biases present in the model.

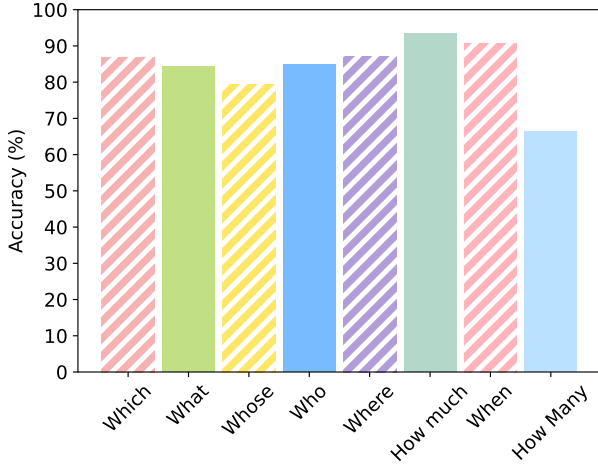
## D.5. Hyperparameter

The hyperparameters used in the experiments are listed in Tab. A.9. Across the fine-tuning strategies, we kept the learning rate constant for all components, including the vision encoder and vision projector. All experiments were conducted on a single NVIDIA A100 SXM4 80GB GPU.

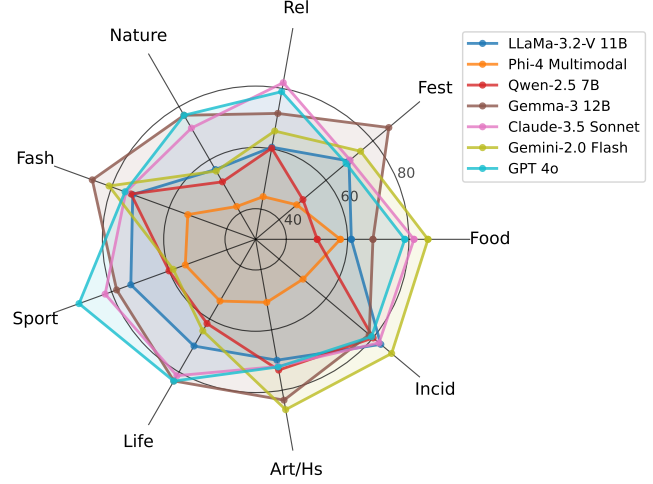
## E. Additional Experimental Results

### E.1. Open-Ended Experimentation

The prompts used for open-ended answer generation and LAVE evaluation are reported in §G. Figure 4b compares GPT-4o’s performance across cultural concepts in



(a) Accuracy over WH-questions for GPT-4o.



(b) Average accuracy across cultural concepts using zero-shot prompting.

Figure A.3. Accuracy of the top-performing model on WH-question types and cross-model comparison on cultural concept categories using zero-shot prompting.

Models	Cultural Concepts									
	Food	Fest	Rel	Nature	Fash	Sport	Life	Art/Hs	Incid	Avg
<b>LLaVa-1.5 7B</b>										
L-LoRA	56.67	55.00	55.00	45.00	48.33	50.00	68.33	46.67	60.00	53.89
Augmentation	66.67	66.67	61.67	55.00	63.33	60.00	58.33	58.33	70.00	<b>62.22</b>
<b>Qwen2.5-VL 7B</b>										
L-LoRA	60.00	76.67	63.33	66.67	78.33	76.67	76.67	78.33	76.67	<b>72.59</b>
Augmentation	43.33	61.67	50.00	63.33	80.00	65.00	66.67	71.67	73.33	63.89
<b>Paligemma-2 10B</b>										
L-LoRA	73.33	78.33	66.67	66.67	86.67	75.00	70.00	70.00	68.33	72.78
Augmentation	73.33	85.00	73.33	70.00	88.33	70.00	76.67	78.33	88.33	<b>78.15</b>
<b>LLaMa-3.2V 11B</b>										
L-LoRA	63.33	76.67	61.67	63.33	75.00	71.67	73.33	71.67	76.67	70.37
Augmentation	76.67	80.00	75.00	65.00	78.33	68.33	71.67	78.33	88.33	<b>75.74</b>
<b>Gemma-3 12B</b>										
L-LoRA	80.00	81.67	75.00	68.33	86.67	68.33	71.67	81.67	83.33	77.41
Augmentation	85.00	86.67	71.67	76.67	88.33	71.67	76.67	85.00	85.00	<b>80.74</b>

Table A.4. Results of fine-tuning using vanilla L-LoRA and using data augmentation with L-LoRA across cultural concepts.

both open-ended and multiple-choice formats. Open-ended accuracy varies widely, from 32.2% for Food to 53.7% for Art & History, highlighting the model’s uneven grasp of different cultural concepts. Multiple-choice questions consistently yield higher accuracy across all categories, peaking at 86.7% for Incidents and 84.4% for Festivals. Accuracy gains from open-ended to multiple-choice can be as high as roughly 50%, e.g., for Food (32.2% → 80.2%). The largest improvements appear in factual and event-based categories, while domain-specific concepts remain more challenging in the open-

ended setting. Further experiments on open-source VLMs in open-ended settings (Tab. A.6) show Gemma-3 12B and Llama-3.2 11B achieving higher scores than their peers, while LLaVA-NeXT-Mistral-7B and Phi-3.5 lag behind.

## E.2. Does language impact performance?

Inspired by K-Viscuit [48], we analyze the impact of input languages on the performance of open-source VLMs. Specifically, we evaluate three configurations: original Bengali texts, Machine-Translated (MT) English texts, and a bilingual setup combining both. Following Tab. A.7, MT

Error Type	Image
<b>Geographic Localization</b> (fails to identify specific places and their regional context)  <b>Ref Question:</b> Which region of this scene/place in the image belongs to?	 <div> A: Jhalokathi P: Barishal    A: Rangpur P: kushtia    A: Agartala P: krishnanagar    A: cox's bazar P: kuakata </div>
<b>Architectural Understanding</b> (poor distinction between sacred, traditional, and traditional architectural styles)  <b>Ref Question:</b> What type of place is shown in the image?	 <div> A: Madrasah P: Mosque    A: Pagoda P: Temple    A: Temple P: Mosque    A: lalbagh kella P: Mosque </div>
<b>Cultural Event</b> (fails to identify festivals and national celebrations)  <b>Ref Question:</b> Based on the time, attire, and activity, which celebration is taking place?	 <div> A: Spring Festive P: BN New Year    A: Martyrs' Day P: Wedding    A: Language Day P: BN New Year    A: Spring Festive P: BN New Year </div>
<b>Regional Food Identification</b> (unable to distinguish local and regional fish/vegetables species)  <b>Ref Question:</b> What is the name of the fish/vegetable shown in the image?	 <div> A: Pangasius P: Tilapia    A: Shrimp P: Dried Fish    A: Arigula P: Water spinach    A: Tilapia P: Rohu </div>
<b>Traditional Sports Recognition</b> (lacks awareness of local sporting activities)  <b>Ref Question:</b> What kind of toys or games the children are engaged with?	 <div> A: Pambaram P: Peg Top    A: Bouchi P: Blind Man's Bluff    A: Tomtom Car P: F1 Car </div>

Table A.5. Types of error made by the best-performing finetuned model, showcasing failures in cultural and regional understanding. Generalized reference questions are used across image sets instead of individual questions per image. (A = Answer; P= Prediction).

English texts lead to better performance for the LLaVA variants, Qwen, and Phi-3.5, while the bilingual configuration yields the best performance for Phi-4, LLaMA-3.2V, and Gemma-3. Except for Gemma-3, all the models underperform in Bengali compared to the MT English or bilingual configuration, highlighting the linguistic limitation of VLMs in handling low-resource languages like Bengali.

### E.3. Circular Evaluation

We implement the Circular Evaluation strategy [33], where each problem is augmented multiple times with answer choices shifted circularly. Models are both trained and evaluated using this strategy. Tab. A.8 shows L-LoRA finetuned models struggling with circular evaluation, implying that models exhibit bias on answer positions instead of semantically understanding the text.

Model	LAVE
Llama-3.2 11B	0.3622
LlaVA-1.5 7B	0.1083
LlaVA-NeXT Mistral-7B	0.0594
Phi-4	0.1097
Phi-3.5-Vision	0.0633
Gemma-3-12B	0.5778
Qwen2.5-VL-7B	0.2639

Table A.6. LAVE Scores of Open Source VLMs for Open Ended Questions

## F. Error Analysis

We continue the error analysis in more detail in this section across prompting strategies and cultural concepts.

### F.1. Qualitative Analysis of Prompting

Figs. A.6 and A.14 compares the predictions and additional instructional tokens produced by the prompting strategies mentioned in Sec. 4.1. The first example (Fig. A.6) shows that the model’s initial zero-shot prediction in Bengali was accurate, correctly identifying the term *Ail* (the narrow, raised earthen path between agricultural plots). However, an error was introduced when this culturally specific word was inaccurately translated into English as *canals* or *boundary*. The English-based Chain of Thought (CoT) reasoning recognized the flawed translation, rejected it, but defaulted to a broader, less accurate description: *over the fields*. This reveals a critical vulnerability where a model’s reasoning can be undermined by its inability to preserve linguistic and cultural nuance during translation.

On the contrary, the second example (Fig. A.6) demonstrates the model’s capacity for precise semantic differentiation, with both its initial prediction and its subsequent reasoning proving successful. The zero-shot answer correctly identified the scene as *mrutshilpo* (pottery art) due to the strong, unambiguous visual evidence of people working with clay. Chain of Thought reasoning reinforces this by meticulously analyzing the given options, realizing that the activity is a type of *handicraft* and could also belong to the *cottage industry*. This case highlights the model’s ability to move beyond simple recognition and correctly navigate layers of meaning to select the most precise and contextually appropriate answer.

The next example (Fig. A.6) showcases Chain of Thought as a crucial self-correction tool. The model’s initial zero-shot and translation-based predictions were flawed, misidentifying the Bengali *Dinga Nouka* as a

*Sampan* (rowboat), a common error where a generic or more widely known label is incorrectly applied to a specific regional artifact. However, the subsequent reasoning process successfully rectifies this mistake. Evaluating each answer option with deeper, culturally-specific knowledge, the CoT correctly disqualifies *Sampan* as Chinese, not Bengali, boat design. It then accurately identifies *Dinga Nouka* as the appropriate term, showing the prowess of reasoning to access nuanced facts and override initial errors.

In Fig. A.14, the analysis transitions from *Zero-Shot/CoT* approaches to a *Description/Cultural* prompting strategy to investigate its impact on model prediction. The floating guava market serves as a compelling example for these strategies. In this case, the Description prompt accurately captured the visual elements of the scene, *numerous small boats, piles of green guavas, and a riverine environment*, but lacked the geographic specificity required to identify the exact location. Conversely, the Cultural prompt correctly contextualized the setting as a *market particularly characteristic of the Barishal region*, yet this valid cultural inference led to the selection of the broader regional label *Barishal* instead of the precise and correct answer, *Jhalokathi*. Both prompts contained accurate information but failed to integrate their insights to achieve the necessary level of specificity.

A different error is illustrated by the subsequent rickshaw example, where both prompts focused narrowly on the vehicle’s method of propulsion and misclassified it as *Hand-pulled*. This oversight neglected the image’s more visually dominant and culturally salient feature: the elaborate *Fine arts* decoration on the rickshaw, which was the correct answer. This example highlights a central challenge of cultural VQA: even when visual descriptions are accurate and cultural reasoning is appropriate, mispredictions may still arise. The failure often lies in the final stage of classification, where the model struggles to combine correct observations into a clear and accurate final answer.

### F.2. Error Types across Cultural Concepts

The examples provided in Tab. A.5 show several categories of errors made by the VLMs on BANGLAPROTHA. In the first set of images, VLMs fail to identify Bengali geographic locations precisely, likely due to insufficient pretraining samples. For instance, the model confuses Rangpur with Kushtia (both mustard-growing regions), misidentifies Agartala as Krishnanagar (both featuring similar colonial administrative architecture), and conflates Cox’s Bazar with Kuakata (both prominent Bengali coastal destinations). These errors reveal that while the model can recognize certain concepts, it lacks the deeper cultural and geographic knowledge necessary to differentiate between

Models	Cultural Concepts									
	Food	Fest	Rel	Nature	Fash	Sport	Life	Art/Hs	Incid	Avg
<b>LLaVa-1.5 7B [30]</b>										
BAN	33.33	33.33	40.00	40.00	25.00	38.33	51.67	38.33	33.33	37.04
ENG	45.00	60.00	45.00	50.00	53.33	53.33	65.00	61.67	50.00	<b>53.70</b>
Bilingual	30.00	41.67	43.33	46.67	53.33	48.33	40.00	51.67	46.67	44.63
<b>LLaVa-NeXT-Mistral-7b [30]</b>										
BAN	28.33	36.67	40.00	43.33	43.33	36.67	45.00	51.67	45.00	41.11
ENG	58.33	68.33	56.67	63.33	75.00	65.00	71.67	75.00	65.00	<b>66.48</b>
Bilingual	65.00	70.00	60.00	61.67	63.33	65.00	63.33	75.00	66.67	65.56
<b>Qwen2.5-VL 7B [58]</b>										
BAN	50.00	50.00	60.00	51.67	73.33	60.00	61.67	73.33	80.00	62.22
ENG	50.00	66.67	58.33	75.00	70.00	66.67	71.67	73.33	76.67	<b>67.59</b>
Bilingual	55.00	63.33	53.33	68.33	75.00	65.00	70.00	73.33	73.33	66.29
<b>Phi-3.5 Vision-Inst [1]</b>										
BAN	38.33	31.67	23.33	25.00	30.00	45.00	36.67	30.00	38.33	33.15
ENG	58.33	73.33	63.33	61.67	75.00	73.33	71.67	66.67	61.67	<b>67.22</b>
Bilingual	53.33	75.00	63.33	53.33	66.67	75.00	66.67	70.00	61.67	65.00
<b>Phi-4 Multimodal [2]</b>										
BAN	57.63	47.46	44.07	42.37	53.57	54.39	53.33	50.85	50.00	50.41
ENG	70.69	50.00	60.00	52.54	58.33	69.49	58.33	66.67	54.24	60.03
Bilingual	65.00	53.33	59.32	55.93	66.67	73.33	61.02	65.00	44.07	<b>60.41</b>
<b>LLaMa-3.2V 11B [41]</b>										
BAN	61.22	69.81	60.47	56.25	72.73	73.33	70.00	70.00	83.33	68.57
ENG	70.00	70.00	55.00	64.41	72.88	77.59	66.67	73.33	71.67	69.06
Bilingual	70.18	78.33	57.89	71.19	61.40	73.21	75.86	76.27	83.33	<b>71.96</b>
<b>Gemma-3 12B [56]</b>										
BAN	68.33	86.67	71.67	76.67	86.67	78.33	83.33	83.48	78.25	79.26
ENG	68.33	80.00	68.33	78.33	76.67	78.33	80.00	76.67	81.67	76.48
Bilingual	71.67	83.33	78.33	75.00	85.00	78.33	83.33	80.00	85.12	<b>80.03</b>

Table A.7. **Comparative Analysis on Different Text Input Languages.** BAN refers to original Bengali texts, ENG is machine-translated texts, and for Bilingual, we provide questions and options in both languages to models. Cyan highlights the highest score for each cultural concept, and **Bold** indicates the overall best-performing model for each model.

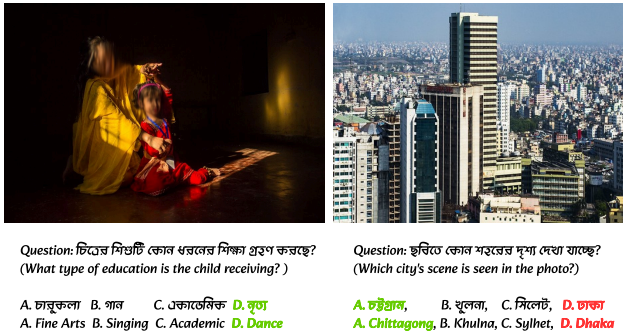


Figure A.4. Error Analysis on Arts & History



Figure A.5. Error Analysis on Clothing & Fashion

locations that appear visually similar.

Similarly, the model demonstrates a limited and often inaccurate understanding of non-Western architectural

Models	Cultural Concepts									
	Food	Fest	Rel	Nature	Fash	Sport	Life	Art/Hs	Incid	Avg
<b>LLaVa-1.5 7B [30]</b>										
Circular Eval	36.02	34.35	42.35	33.39	29.99	37.68	43.21	34.56	39.90	36.83
L-LoRA	56.67	55.00	55.00	45.00	48.33	50.00	68.33	46.67	60.00	<b>53.89</b>
<b>Qwen2.5-VL 7B [58]</b>										
Circular Eval	47.66	61.23	53.99	55.11	71.15	68.18	61.24	65.57	65.13	61.03
L-LoRA	60.00	76.67	63.33	66.67	78.33	76.67	76.67	78.33	76.67	<b>72.59</b>
<b>LLaMa-3.2V 11B [41]</b>										
Circular Eval	53.77	56.70	56.33	54.99	65.36	59.36	58.46	62.01	64.87	59.09
L-LoRA	63.33	76.67	61.67	63.33	75.00	71.67	73.33	71.67	76.67	<b>70.37</b>
<b>Gemma-3 12B [56]</b>										
Circular Eval	64.67	70.56	54.88	52.57	73.69	54.99	51.48	75.44	78.19	64.05
L-LoRA	80.00	81.67	75.00	68.33	86.67	68.33	71.67	81.67	83.33	<b>77.41</b>

Table A.8. Circular Evaluation on Finetune Models. LoRA fine-tuned models struggle with the circular evaluation.



Question: ছবিতে প্রদর্শিত মাটির জিনিসগুলো কোন শিল্পের অন্তর্গত?  
(The clay items shown in the picture belong to which craft?)

- A. পটশিল্প, B. কৃষিশিল্প, C. বস্ত্রশিল্প, D. ঔষধশিল্প  
A. Pottery, B. Agricultural craft, C. Textile craft, D. Herbal craft



Question: বাংলাদেশের ইতিহাসের কোন গুরুত্বপূর্ণ দিন উদ্‌যাপন করা হচ্ছে?  
(Which important day in Bangladesh's history is being celebrated?)

- A. বিজয় দিবস, B. আন্তর্জাতিক মাতৃভাষা দিবস ২১ ফেব্রুয়ারি, C. স্বাধীনতা দিবস, D. পহেলা বৈশাখ  
A. Victory Day, B. Int. Mother Language Day, C. Independence Day, D. Pohela Boishakh

Figure A.6. Error Analysis on Cultural Festivals

Hyperparameter	Finetune
batch size	32
lr	2e-4
lr schedule	cosine decay
lr warmup ratio	0.03
weight decay	0
optimizer	AdamW
lora_r	64
lora_alpha	64
epochs	4

Table A.9. Hyperparameters used for finetuning.

styles. In the second set of images in Tab. A.5, the models fail to recognize the distinct functional and religious purposes of different architectures due to inadequate knowledge of regional building design. For instance, the model misidentifies a madrasah (Islamic educational institution) as a mosque even when madrasah is written in

Bengali script on the building. It confuses a Hindu temple with a mosque despite the prominent display of a Hindu mantra on the temple wall in Bengali.

The above phenomenon exposes that not only does the model lack understanding of Bengali architecture, but it also fails to process Bengali textual cues that would enable correct identification. While it can detect general elements like domes and arches, it cannot integrate these with scene-text recognition or cultural context to determine a building's purpose accurately. The model also fails to interpret symbolic context, *i.e.*, it recognizes objects and people but misinterprets their meaning, *e.g.*, confusing a solemn memorial event with a celebration based on the presence of flowers.

Besides, the model exhibits a strong bias towards globally recognized food items and lacks knowledge of local biodiversity. It repeatedly misidentifies common regional fish and vegetable species by defaulting to the incorrect,

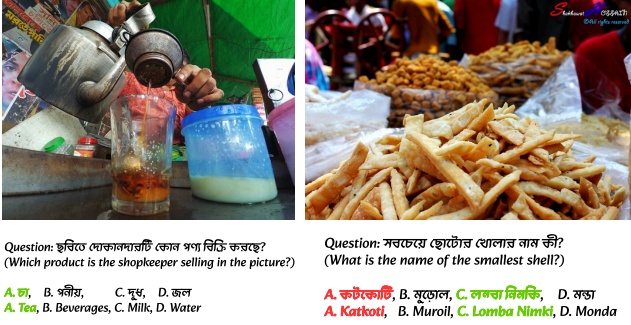


Figure A.7. Error Analysis on Food

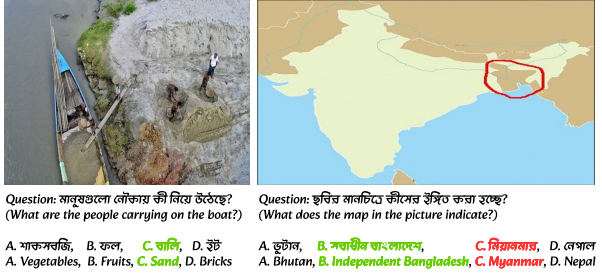


Figure A.8. Error Analysis on Incidents



Figure A.9. Error Analysis on Nature

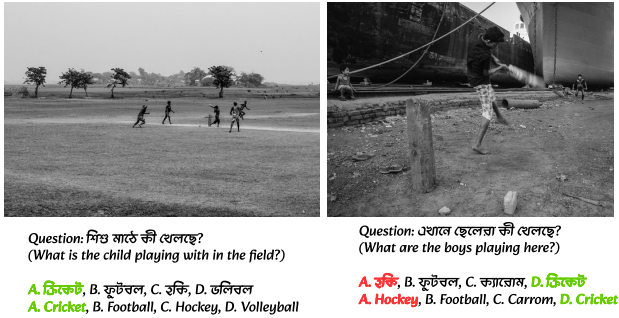


Figure A.10. Error Analysis on Sports

Western or global counterparts (e.g., labeling various local fish as *Tilapia*). This points to the underrepresentation of regional cuisine in the pertaining data.

Finally, the model shows a near-total lack of awareness of traditional folk games and handmade toys. It either fails to recognize the activity or misclassifies a local craft with an unrelated global equivalent (e.g., a handmade toy car for an F1 car), i.e., similar instances of underrepresentation in pretraining data. For each of the nine cultural categories, we contrast a successful prediction with an insightful failure, presented in Fig. A.4 through Fig. A.12

## G. Prompts

### Base/Zero-Shot Prompt

You are a highly capable multimodal AI assistant. You will be provided with an image as visual context, followed by a multiple-choice question related to that image. Each question will have four options labeled (A), (B), (C), and (D). Your task is to select the most accurate answer from the given options based on the image. Format your final answer as: (X), where X is the correct option letter from A, B, C, and D.

### Culture-specific Prompt

You are a culturally aware, expert multimodal AI assistant with a deep understanding of the Bengali culture. You will be provided with an image as visual context. The image will reflect one or more of nine key cultural aspects: Food, Cultural Festivals, Nature, Religious Events, Clothing and Fashion, Sports, Social Life, Incidents, and Arts & History. Following the image, you will receive a multiple-choice question with four options labeled (A), (B), (C), and (D). Your task is to select the most accurate answer from the given options based on the cultural context of the image. Format your final answer as: (X), where X is the correct option letter from A, B, C, and D.

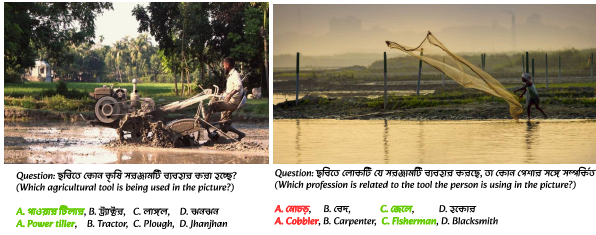


Figure A.11. Error Analysis on Social Life

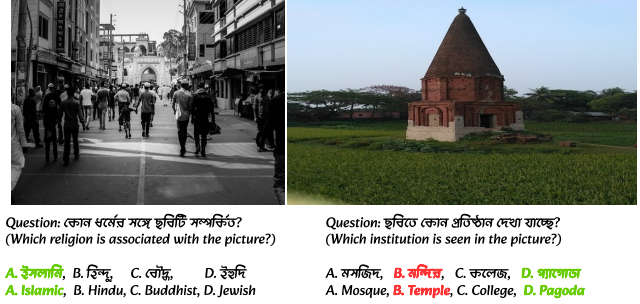


Figure A.12. Error Analysis on Religious Events

### Chain of Thought Prompt

You are a highly capable multimodal AI assistant. You will be provided with an image as visual context, followed by a multiple-choice question related to that image. Each question will have four options labeled (A), (B), (C), and (D). Your task is to analyze the image, question, and answer options carefully, then engage in step-by-step reasoning and eliminate the incorrect options. After completing the reasoning steps, select the most accurate answer from the given options. Format your final answer as: (X), where X is the correct option letter from A, B, C, and D.

### Open Ended Experimentation Prompt

You are a highly capable multimodal AI assistant. You will be provided with an image as visual context, followed by a question related to that image. Your task is to generate a suitable answer to the question based on the image content. You should always abide by the guidelines mentioned below:

1. The answer must be one or two words, at most three.
2. The answer must be in Bengali language and script.

(IMAGE), (QUESTION)

Generate your response in the following format:

A: (GENERATED ANSWER)

### Translation-based Prompt

You are a highly capable multimodal and multilingual AI assistant. You will be provided with an image as visual context, followed by a multiple-choice question related to that image. Each question will have four options labeled (A), (B), (C), and (D). Your task is to translate the four options from Bangla into English and select the most accurate answer from the given options based on the image. Format your final answer as: (X), where X is the correct option letter from A, B, C, and D.

### LAVE Evaluation Prompt

You are an expert cultural anthropologist tasked with evaluating the correctness of candidate answers for cultural Visual Question Answering (VQA).

Given an image as context, a question, a reference answer by an expert, and a candidate answer by a model, rate the candidate answer's correctness.

Use a scale of 0-1, where 0 indicates an incorrect, irrelevant, or imprecise answer, and 1 indicates a correct, precise answer according to the reference.

You have to provide the rationale for your rating and then provide a rating in the format: 'Rating: X', where X is either 0 or 1.

### Description-based Prompt

You are a highly capable multimodal AI assistant. You will be provided with an image as visual context, followed by a multiple-choice question related to that image. Each question will have four options labeled (A), (B), (C), and (D). Your task is to generate a suitable description of the given image and select the most accurate answer from the given options based on the image and the description. Format your final answer as: (X), where X is the correct option letter from A, B, C, and D.



Q: ছেলেগুলি কিসের উপর দিয়ে হেটে যাচ্ছে ?

Q: What are the boys walking on?

- A. আইল দিয়ে (boundary) B. ক্ষেতের উপর C. রাস্তা দিয়ে D. পানি দিয়ে  
A. On a narrow ridge (all) B. On the field C. On the road D. On water

Zero-Shot: A. আইল দিয়ে

Translated: Question: What are the children walking over?

Options:

- A. Through canals B. Over the fields  
C. Through the road D. Through the water

Answer: B

CoT: The question asks what the children are walking over. Looking at the image, it's clear they are walking through fields. Option A refers to "ail" which is not relevant here. Option C refers to roads, which are not visible in the image. Option D mentions water, which is not the case. Option B accurately captures that they are walking across fields. Therefore, B is correct. Answer: B

Q: ছবিতে প্রতিকলিত হওয়া শিল্পটি কোনটি?

Q: Which form of art is reflected in the picture?

- A. গৃহশিল্প B. হস্ত শিল্প C. পোশাক শিল্প D. কুটির শিল্প  
A. Pottery B. Handicraft C. Garment industry D. Cottage industry

Zero-Shot: A. গৃহশিল্প

Translated: Question: What kind of art is reflected in the picture?

Options: A. Pottery art B. Handicraft art C. Garment industry D. Cottage industry

Answer: A



CoT: The question asks "What art is reflected in the picture?". The image shows two people working with clay and making pottery.

Let's analyze the options:

A. মৃৎশিল্প (mrutshilpo) - This means "pottery art" or "clay modeling". This aligns perfectly with what's depicted in the image.

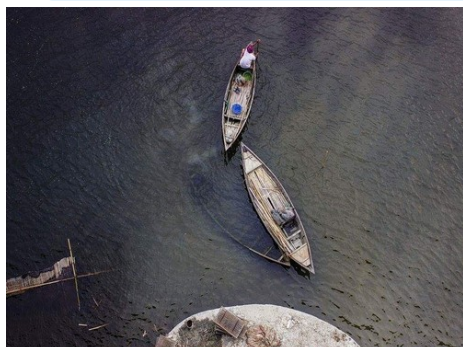
B. হস্ত শিল্প (hostshilpo) - This means "handicraft." While pottery is a handicraft, two people working with clay and making pottery, this is a broader term and doesn't specifically describe the art being practiced.

C. পোশাক শিল্প (poshoshshilpo) - This means "textile industry" or "clothing industry". This is clearly not relevant to the image.

D. কুটির শিল্প (kutirshilpo) - This means "cottage industry". This is a broader economic term and doesn't specifically address the art form.

Therefore, the most accurate answer is মৃৎশিল্প (pottery art).

Answer: A



Q: ছবিতে ভাসমান নৌকা দুটি কোন ধরনের?

Q: What type are the two floating boats in the picture?

- A. ডিঙ্গি নৌকা B. স্পিড বোর্ড C. সাম্পান D. ডেলা  
A. Dinghy boat B. Speed boat C. Sampan D. Raft

Zero-Shot: C. সাম্পান

Translated: Question: What type of boats are floating in the picture?

Options:

- A. Rowboat  
B. Speed board  
C. Sampan  
D. Raft

Answer: C

CoT: The question asks what kind of boats are shown in the image. We need to identify the type of boat that is typically used in Bangla (Bangladeshi/Bengali) culture depicted in the image. Looking at the image, the boats are small, traditional wooden boats. They appear to be simple, narrow, and likely used for fishing or transportation on rivers and canals. Let's evaluate the options: A. ডিঙ্গি নৌকা (Dinga Nouka): A dinghy is a small boat, often used as a tender to a larger vessel or for recreation. It fits the appearance of the boats in the image. B. স্পিড বোর্ড (Speed Board): This is a modern recreational item and doesn't align with the traditional context of Bangla culture, it does not even look like a speed board. C. সাম্পান (Sampan): Sampans are distinctively Chinese boats. While some influence might exist due to trade, they are not characteristic of Bangla boat design. D. ডেলা (Vela): A vela is a raft, usually made of logs tied together. These boats are clearly constructed vessels, not rafts. Therefore, 'Dinga Nouka' is the most appropriate description for the boats in the image reflecting a traditional Bangla setting. Answer: A

Figure A.13. Zero-Shot vs. Translated vs Chain-of-Thought approach error analysis. Detailed and accurate parts are emphasized in colored and bold letters. The red color is used to highlight the mistakes it made.



Q: ছবির টেরাকোটা চিত্রে বাংলাদেশের কোন জনগোষ্ঠীর দৈনন্দিন জীবন ফুটে উঠেছে?  
Q: In the terracotta depiction of the image, the daily life of which community of Bangladesh is portrayed?

- A. বাঙ্গালি B. ক্ষুদ্র নৃগোষ্ঠী C. মুসলিম D. অমুসলিম  
A. Bengali B. Ethnic minorities C. Muslim D. Non-Muslim

**Description:** The image shows a terracotta relief depicting scenes of daily life. The artwork is composed of individual terracotta tiles arranged to form a larger mural. The scenes portrayed appear to show rural activities, possibly agricultural practices, with people engaged in farming, carrying baskets, playing musical instruments, and other daily tasks. The figures are styled in traditional clothing, and the overall aesthetic suggests a representation of rural Bengali life. There's text written in Bengali script above the figures, potentially providing context about the artwork's title or subject. The base of the terracotta is mounted on a rough stone wall.

**Answer: A**

**Cultural:** The image showcases a terracotta relief depicting scenes of rural Bengali life. The people are dressed in traditional attire and engaged in activities like farming, playing musical instruments (like the \*dotara\* and \*tabla\*), and carrying baskets - all characteristic of the Bengali people's cultural heritage and daily routines. The depiction of agricultural practices, traditional clothing, and musical expressions strongly suggests a portrayal of the broader Bengali community rather than a specific minority group. The text "Shagorer Neelgori/del Rider" written on the terracotta panel hints at a specific location or event related to the Bengali culture.

**Answer: A**

Q: ছবির এই ভাসমান গিয়ারা বাজার কোন জায়গায় অবস্থিত?  
Q: The floating guava market shown in the picture is located in which place?

- A. জ্বালকাঠি B. বরিশাল C. ঢাকা D. নাটর  
A. Jhalokathi B. Barishal C. Dhaka D. Natore

**Description:** The image depicts a vibrant scene of a floating guava market on a river. Numerous small boats are laden with piles of green guavas, and men are actively sorting, loading, and selling the fruit. Some men are standing on a makeshift platform structure, likely a trading hub, with stacked crates of guavas nearby. The water is murky green, and the background is dense with lush tropical trees and vegetation, hinting at a riverine environment. The scene is bustling with activity and appears to be a significant economic operation centered around the harvest and trade of guavas. The overall atmosphere suggests a unique cultural practice integrated with the natural environment.

**Answer: B**



**Cultural:** The image depicts a floating guava market, a unique and culturally significant scene prevalent in the southern districts of Bangladesh. These markets are a traditional way of trading, particularly for agricultural goods like guavas, where vendors navigate waterways on boats, creating a bustling marketplace on the water. The presence of numerous boats laden with guavas, along with people actively involved in buying, selling, and loading/unloading, strongly suggests a floating market environment. Floating markets are particularly characteristic of the Barishal region in Bangladesh. **Answer: B**



Q: রিকশা তে আমরা কোন বৈশিষ্ট্য দেখি?  
Q: What feature do we observe in a rickshaw?

- A. চাকলা B. হাতে টানা C. ঘোড়া চালিত D. চাকাবিহীন  
A. Fine arts B. Hand-pulled C. Horse-drawn D. Wheel-less

**Description:** The image depicts a vibrant sunset scene in what appears to be a bustling, somewhat impoverished urban area. The sky is the dominant feature, filled with a dramatic display of clouds illuminated by the setting sun. The colors range from deep blues and grays to fiery oranges and yellows, creating a striking contrast. Sunbeams are visible piercing through the cloud cover. In the foreground, there's a street scene with buildings, trees, and several people. Notably, a decorated rickshaw is visible in the lower right corner, with a man standing nearby. Other vehicles and people are also present, suggesting a lively, everyday scene. **Answer: B**

**Cultural:** The image depicts a sunset scene in what appears to be a Bangladeshi urban or semi-urban setting. A common feature is the presence of a cycle rickshaw, a ubiquitous mode of transportation in Bangladesh. These rickshaws are manually pulled by a driver and are highly recognizable cultural symbols. The scene is bustling with everyday life, showing people and buildings in the background. The clothing worn by the people also suggests a South Asian context. The ornate decorations on the rickshaw itself are examples of local artistry. **Answer: B**

Figure A.14. Description vs Cultural evaluation approach error analysis. Detailed and accurate parts are emphasized in colored and bold letters. The red color is used to highlight the mistakes it made.