

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 Streamlit². 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.

²<https://streamlit.io>

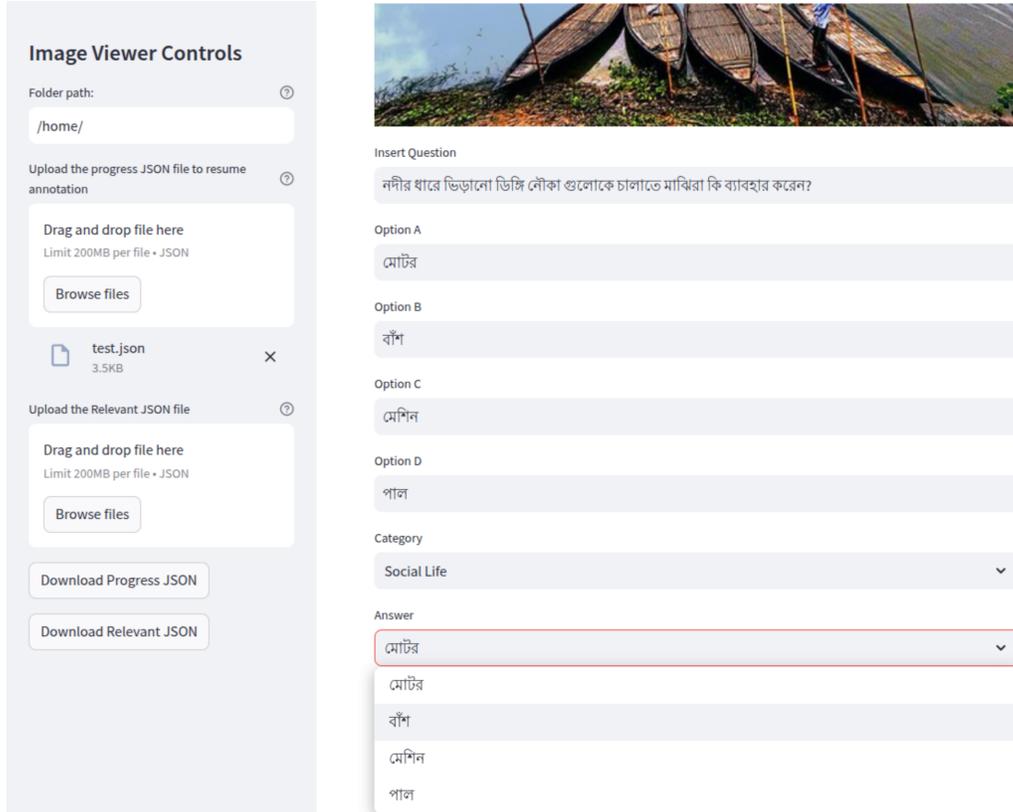


Figure A.1. Demo of our annotation tool.

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 κ 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-6.7B³, LLaVA-1.5-7B [30],

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

and LLaVA-Next-Mistral-7B⁴ [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 framework⁵ 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 .

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

⁵<https://pytorch.org/>

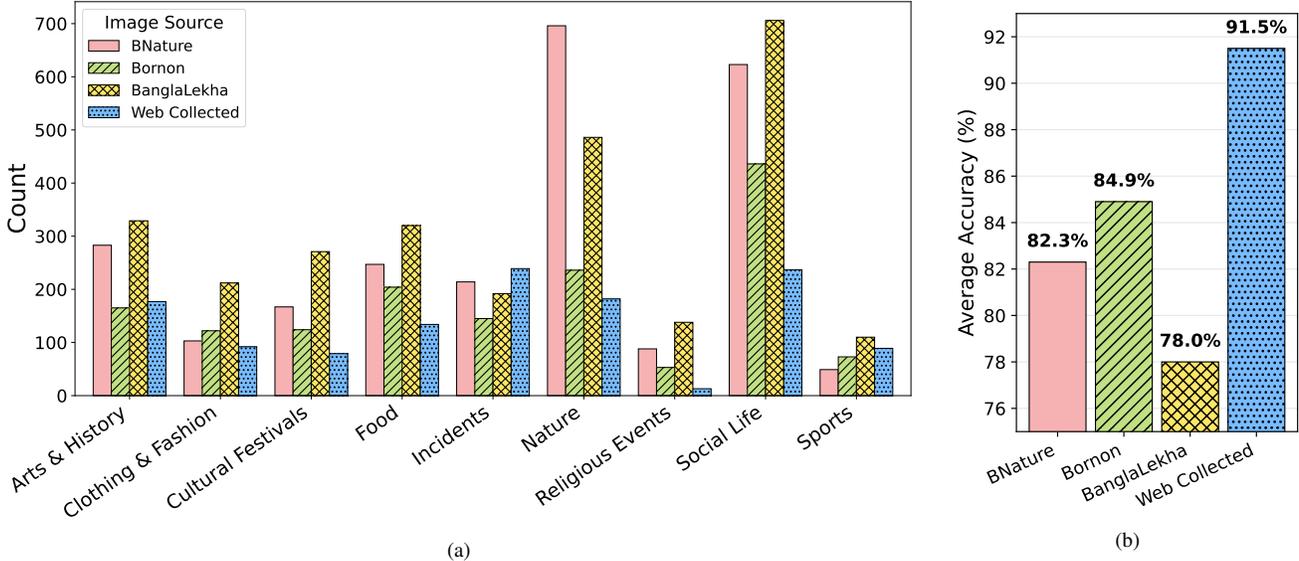


Figure A.2. (a) Distribution of image sources of BANGLAPROTHA across cultural aspects, (b) Performance of GPT-4o model using zero-shot prompting across image sources. Performance can be correlated to the quality of images from that source.

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
Total	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.

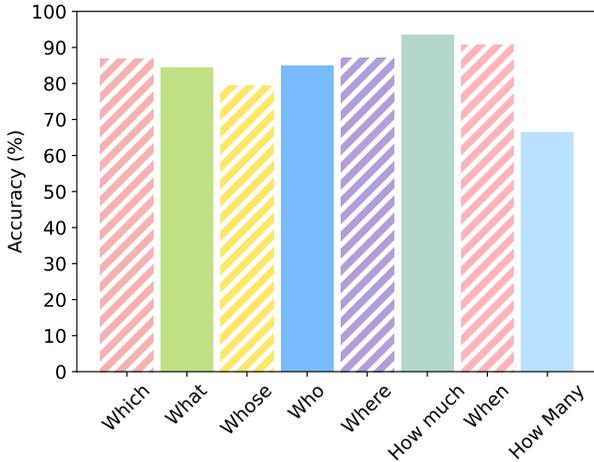
D.5. Hyperparameter

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.

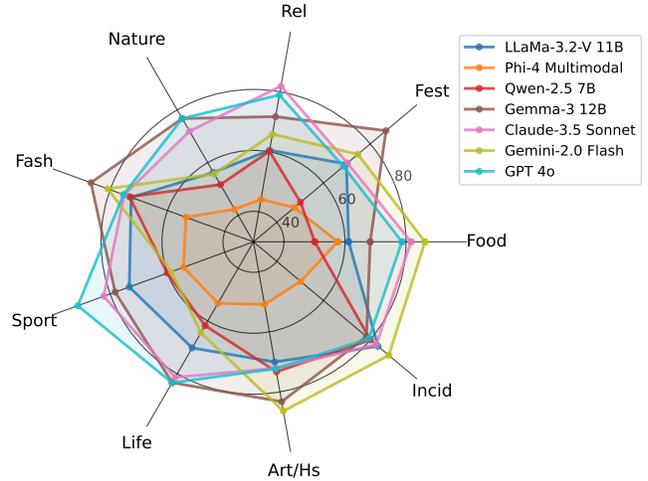
E. Additional Experimental Results

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
LLaVa-1.5 7B										
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	62.22
Qwen2.5-VL 7B										
L-LoRA	60.00	76.67	63.33	66.67	78.33	76.67	76.67	78.33	76.67	72.59
Augmentation	43.33	61.67	50.00	63.33	80.00	65.00	66.67	71.67	73.33	63.89
Paligemma-2 10B										
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	78.15
LLaMa-3.2V 11B										
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	75.74
Gemma-3 12B										
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	80.74

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

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 *mritshilpo* (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
LLaVa-1.5 7B [30]										
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	53.70
Bilingual	30.00	41.67	43.33	46.67	53.33	48.33	40.00	51.67	46.67	44.63
LLaVa-NeXT-Mistral-7b [30]										
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	66.48
Bilingual	65.00	70.00	60.00	61.67	63.33	65.00	63.33	75.00	66.67	65.56
Qwen2.5-VL 7B [58]										
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	67.59
Bilingual	55.00	63.33	53.33	68.33	75.00	65.00	70.00	73.33	73.33	66.29
Phi-3.5 Vision-Inst [1]										
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	67.22
Bilingual	53.33	75.00	63.33	53.33	66.67	75.00	66.67	70.00	61.67	65.00
Phi-4 Multimodal [2]										
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	60.41
LLaMa-3.2V 11B [41]										
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	71.96
Gemma-3 12B [56]										
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	80.03

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
LLaVa-1.5 7B [30]										
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	53.89
Qwen2.5-VL 7B [58]										
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	72.59
LLaMa-3.2V 11B [41]										
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	70.37
Gemma-3 12B [56]										
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	77.41

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. পুঁসিগু, 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.9. Error Analysis on Nature

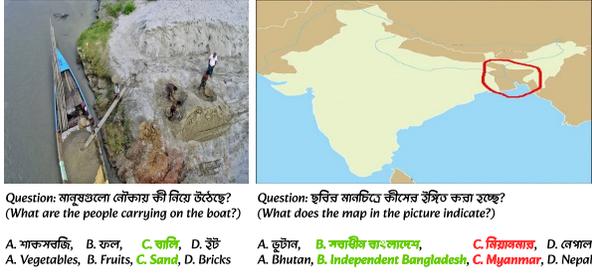


Figure A.8. Error Analysis on Incidents

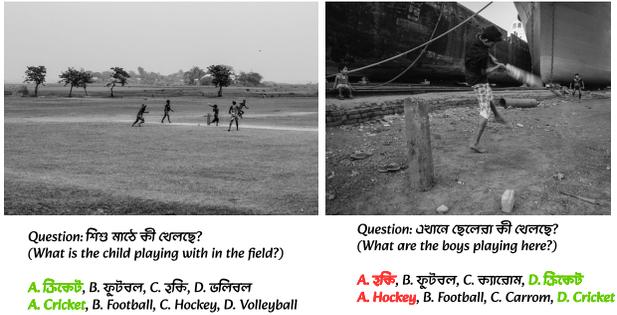


Figure A.10. Error Analysis on Sports

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

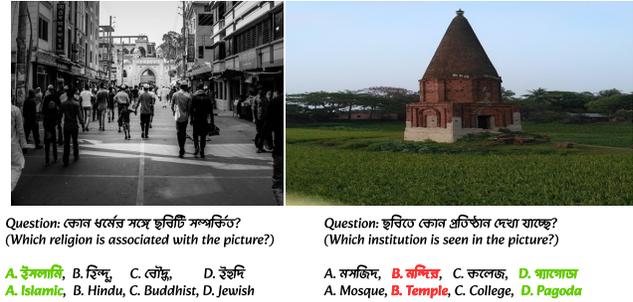


Figure A.12. Error Analysis on Religious Events

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. গানি 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 "all" 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. মৃৎশিল্প B. হস্ত শিল্প C. পোশাক শিল্প D. কুটির শিল্প
 A. Pottery B. Handicraft C. Garment industry D. Cottage industry



Zero-Shot: A. মৃৎশিল্প
 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 (mritshilpo) - 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 ধরণের?
 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 গুয়া 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. চাকুলা 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.