Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

VisText-VLM2Plastic: A Bilingual Benchmark for Visual Grounding and Environmental Reasoning in Vision-Language Models

Tarek Rahman

trahman221182@bscse.uiu.ac.bd

United International University https://orcid.org/0009-0006-0224-6516

Research Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11717803/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

VisText-VLM2Plastic: A Bilingual Benchmark for Visual Grounding and Environmental Reasoning in Vision-Language Models

Tarek Rahman[1]
[1]United International University
trahman221182@bscse.uiu.ac.bd

Abstract

Vision-language models (VLMs) increasingly support open-ended visual reasoning, yet their behavior in low-resource languages and sustainability-oriented domains remains underexplored. We introduce a controlled English–Bangla plastic-waste VQA benchmark comprising 402 image records derived from 136 base images and 2,412 question–answer pairs across five waste categories and three task types: recognition, visual description, and environmental reasoning. We evaluate GPT-4.1 mini, Gemini 2.5 Flash, Qwen2.5-VL-7B-Instruct, and Gemma-3-4B zero-shot using task-specific answer-quality metrics, bilingual consistency, and image ablation. Gemini 2.5 Flash yields the highest recognition point estimate, with 93.5% accuracy and 91.3% macro-F1. Across models, English normalized performance exceeds Bangla by 3.78 points on average, with the largest observed gap for environmental reasoning. Removing the image reduces normalized performance by 66.29 points for recognition, 52.89 for description, and 11.42 for environmental reasoning. The smaller environmental-reasoning ablation gap indicates that, in this benchmark, many environmental questions remain partly answerable from linguistic and domain priors. These results show why answer quality alone is insufficient for multilingual multimodal evaluation and motivate reporting cross-lingual consistency and sensitivity to visual evidence alongside conventional metrics.

1 Introduction

Plastic waste produces persistent environmental burdens through litter accumulation, drainage blockage, and movement into waterways. Vision-language models (VLMs) offer a flexible interface for recognizing visible waste, describing image-supported properties, and answering environmental questions, but a correct answer does not necessarily imply that the model used the image. This concern is well established in visual question answering (VQA), where models can exploit linguistic priors instead of visual evidence [Antol et al., 2015, Goyal et al., 2017]. It is especially relevant for environmental questions, because generic facts about plastic pollution may remain plausible even when the pictured object is unknown.

Multilingual evaluation adds a second challenge. Bangla remains comparatively underrepresented in multimodal benchmarks, despite recent resources such as ChitroJera, BanglaProtha, and BanglaMedVQA [Barua et al., 2024, Fahim et al., 2026, Ahmed et al., 2026]. We therefore introduce **VisText-VLM2Plastic**, a controlled English–Bangla plastic-waste VQA benchmark with 402 image records derived from 136 base images and 2,412 paired question–answer instances. The benchmark covers five waste categories and three task types: recognition, visual description, and environmental reasoning.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We study three questions: how VLM performance changes across the three task types; how large the English–Bangla gap is for semantically aligned questions; and how strongly performance changes when the image is removed. We evaluate GPT-4.1 mini, Gemini 2.5 Flash, Qwen2.5-VL-7B-Instruct, and Gemma-3-4B zero-shot. In addition to conventional answer-quality metrics, we report bilingual consistency and an image-ablation diagnostic. Full related work, dataset construction, quality control, prompts, metric definitions, statistical procedures, and additional analyses are provided in the Appendix.

2 Benchmark and Evaluation

Each image record is paired with six questions: English and Bangla versions of recognition, visual-description, and environmental-reasoning tasks. This yields 804 QA pairs per task type and 1,206 instances per language. Augmented variants preserve a `source_base_id` and remain grouped during splitting to prevent near-duplicate leakage. The benchmark is class-imbalanced, with flexible wrappers dominant; therefore recognition is reported with both accuracy and macro-F1. Dataset composition, split statistics, construction, and quality-control details are in Appendix C.

All models are evaluated zero-shot. Recognition outputs are mapped deterministically to the five target categories. Open-ended responses are evaluated primarily with BERTScore-F1, with BLEU-4 and ROUGE-L as complementary lexical metrics [Zhang et al., 2020, Papineni et al., 2002, Lin, 2004]. English–Bangla performance is compared on a normalized 0–100 primary-score scale: accuracy for recognition and BERTScore-F1 for the two open-ended tasks. Bilingual consistency additionally compares paired English and Bangla outputs. To test image sensitivity, each question is evaluated both with the image and without it; the reported visual-grounding gain (VGG) is the Image+Q score minus the Q-only score. We treat VGG as benchmark-conditioned sensitivity to image removal, not as a standalone causal measure of grounding. Full implementation details are in Appendix D.The complete benchmark and evaluation flow is summarized in Figure 1.

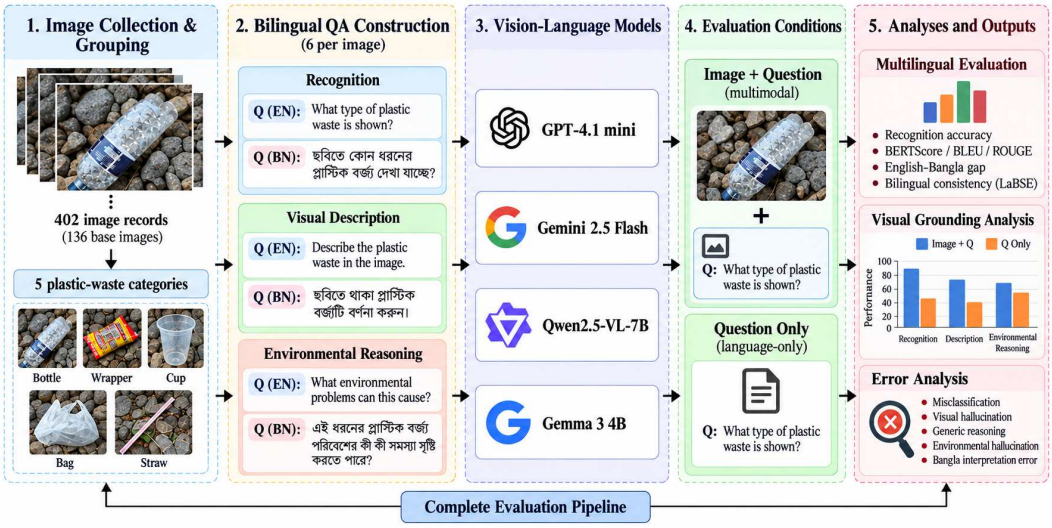


Figure 1: Benchmark and evaluation pipeline. Each grouped plastic-waste image produces aligned English and Bangla recognition, description, and environmental-reasoning questions. Four VLMs are evaluated under image+question and question-only conditions, followed by multilingual, grounding, and error analyses.

3 Results

3.1 Recognition and Open-Ended Answer Quality

Table 1 reports zero-shot recognition. Gemini 2.5 Flash has the highest point estimate at 93.5% accuracy and 91.3% macro-F1, followed by GPT-4.1 mini. Because the test split contains only 21

Table 1: Zero-shot recognition performance.

Model	Accuracy (%)	Macro-F1 (%)
GPT-4.1 mini	91.9	89.2
Gemini 2.5 Flash	**93.5**	**91.3**
Qwen2.5-VL-7B	88.7	85.6
Gemma 3 4B	85.5	81.4

Table 2: Open-ended response quality.

	Description			Environmental Reasoning		
Model	**BERT-F1**	**BLEU-4**	**ROUGE-L**	**BERT-F1**	**BLEU-4**	**ROUGE-L**
GPT-4.1 mini	0.868	0.431	0.619	0.840	0.347	0.566
Gemini 2.5 Flash	**0.881**	**0.452**	**0.641**	**0.855**	**0.369**	**0.588**
Qwen2.5-VL-7B	0.834	0.398	0.587	0.802	0.311	0.531
Gemma 3 4B	0.804	0.361	0.552	0.769	0.276	0.498

independent base images, small point-estimate differences should not be interpreted as a robust universal ranking; cluster-bootstrap uncertainty is discussed in Appendix D.4.

Open-ended results in Table 2 show higher answer–reference similarity for visual description than for environmental reasoning across all four models. BERTScore values are higher than BLEU-4 because semantically similar paraphrases can share relatively few exact four-grams. The environmental task also differs in answer length, answer space, linguistic demands, and required domain knowledge, so the lower scores should be interpreted as a task-type difference rather than an isolated causal measure of reasoning difficulty.

3.2 Language Gap and Image Sensitivity

Table 3 combines the two central diagnostics. All three task types show an English advantage, with the largest observed gap for environmental reasoning (4.15 points). However, the image-ablation pattern is much more uneven: recognition drops by 66.29 points without the image, description by 52.89, but environmental reasoning by only 11.42.

The English–Bangla pattern indicates that cross-lingual divergence is most visible in the most open-ended task, but the task types also differ in response format and generation demands. The ablation result is more diagnostic of benchmark behavior: many environmental questions remain partly answerable from linguistic and domain priors. Phrases such as "plastic waste" and "improper disposal" can elicit plausible statements about litter, drains, waterways, or wildlife even when the image is absent. Thus a high environmental answer-quality score does not by itself establish object-specific visual grounding, while a small VGG does not prove that a model generally ignores images; it can also reveal that a benchmark question is insufficiently discriminative without visual input.

Paired English–Bangla outputs show the same qualitative trend: consistency is highest for recognition and decreases for description and environmental reasoning. Full model-wise bilingual-consistency results, language-specific lexical metrics, category-level performance, qualitative errors, and visualizations are reported in Appendix E.

4 Discussion and Conclusion

The results support two main conclusions. First, multilingual answer quality cannot be summarized reliably by a single aggregate score. English performs better than Bangla across all three task types, and the gap is largest for environmental reasoning. Because lexical metrics are especially sensitive to surface-form variation, semantic and cross-lingual consistency measures are necessary complements.

Second, answer quality and visual dependence are distinct properties. Recognition and description are strongly affected by image removal, whereas environmental reasoning retains much of its score without visual input. For this benchmark, part of the environmental score therefore reflects pretrained

Table 3: Task-level language gap and visual-ablation diagnostics on the normalized primary-score scale.

Task	EN	BN	Gap	I+Q	Q	VGG
Recognition	91.53	87.90	3.63	89.71	23.42	**66.29**
Description	86.43	82.88	3.55	84.65	31.76	**52.89**
Environmental	83.68	79.53	**4.15**	81.60	70.18	**11.42**

environmental knowledge and linguistic priors rather than image-specific evidence. This motivates benchmark designs with counterfactual visual pairs and category-specific questions whose correct answer must change with the image.

VisText-VLM2Plastic is intended as a controlled diagnostic benchmark rather than a comprehensive representation of plastic pollution or Bangla multimodal use. Its principal limitations are the modest number of independent base images, class imbalance, semi-automated construction, controlled language, and the imperfect cross-lingual comparability of automatic metrics. These limitations and additional threats to validity are detailed in Appendix G. Future work should increase independent visual diversity, strengthen native-speaker and expert annotation, report fixed model snapshots and metric configurations, and design environmental questions that require visual evidence more directly.

References

R. Ahmed, I. Tahmid, M. S. Hossain, T. H. Tomal, M. M. Jawad, A. B. Uddin, M. Fahim, and M. F. A. Bhuiyan. Evaluating large vision language models on bangla medical visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 37362–37378, 2026. doi: 10.18653/v1/2026.findings-acl.1862.

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 (ICCV)*, pages 2425–2433, 2015.

S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

D. D. Barua, M. S. U. R. Sourove, M. F. Ishmam, F. Haider, F. T. Shifat, M. Fahim, and M. F. Alam. Chitrojera: A regionally relevant visual question answering dataset for bangla. *arXiv preprint arXiv:2410.14991*, 2024.

M. Fahim, M. S. U. Rahman, A. M. Rahman, M. F. Ishmam, M. T. Rahman, F. T. Shifat, F. Haider, and M. F. A. Bhuiyan. Banglaprotha: Evaluating vision language models in underrepresented long-tail cultural contexts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1159–1169, 2026.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 878–891, 2022. doi: 10.18653/v1/2022.acl-long.62.

Gemma Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Google DeepMind. Gemini 2.5 flash model documentation and model card. Google, 2025.

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 (CVPR)*, pages 6904–6913, 2017.

K. Kataruka, H. Maurya, A. Vats, M. Mandal, K. Raja, and P. K. Chandaliya. Wasteassistant: Regulation-guided visual question answering framework for intelligent waste segregation and sustainable management. *arXiv preprint arXiv:2607.10610*, 2026.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

H. J. Malla, M. Bazli, and M. Arashpour. Enhancing waste recognition with vision-language models: A prompt engineering approach for a scalable solution. *Waste Management*, 204:114939, 2025. doi: 10.1016/j.wasman.2025.114939.

OpenAI. Introducing gpt-4.1 in the api. OpenAI, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002. doi: 10.3115/1073083.1073135.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*, 2020.

A Extended Related Work

A.1 Visual Question Answering and Language Priors

VQA requires models to jointly reason over images and natural-language questions [Antol et al., 2015]. However, Goyal et al. [2017] showed that models can exploit question regularities and language priors, producing strong scores without sufficiently interpreting the image. Our work follows this diagnostic perspective but directly removes the image to quantify modality dependence across task types.

A.2 Bangla Multimodal Evaluation

Bangla multimodal evaluation remains comparatively limited. ChitroJera introduced regionally relevant Bangla VQA data [Barua et al., 2024], BanglaProtha studied Bengali cultural understanding across VLMs [Fahim et al., 2026], and BanglaMedVQA evaluated domain-specific medical reasoning in Bangla [Ahmed et al., 2026]. Our benchmark differs by focusing on environmental reasoning and by using semantically aligned English–Bangla questions for the same images.

A.3 Vision-Language Models for Waste Understanding

Recent work has extended waste analysis beyond supervised image classification. Malla et al. [2025] evaluated VLMs for zero-shot waste recognition and reported substantial gains from prompt optimization. WasteAssistant further introduced waste-oriented VQA and regulation-aware reasoning across multiple waste categories [Kataruka et al., 2026]. In contrast, our benchmark focuses on multilingual robustness and image sensitivity, asking whether Bangla performance differs from English and how strongly environmental-answer quality changes when visual input is removed.

B Benchmark Overview Figures and Tables

Table 4: Benchmark composition by plastic-waste category.

Category	Image Records	Fraction
Plastic wrapper / flexible packaging	224	55.7%
Plastic bottle	70	17.4%
Plastic cup	48	11.9%
Plastic straw	36	9.0%
Plastic bag / plastic sheet	24	6.0%
Total	**402**	**100%**

C Dataset Benchmark

The benchmark contains 402 image records derived from 136 underlying base images. Multiple records correspond to augmented variants of the same source image; therefore, a `source_base_id` is preserved for every record, and all augmentations originating from the same source image are treated as a single group during dataset splitting. The final benchmark contains five dominant plastic-waste categories, whose record-level distribution is summarized in Table 4. The dataset is imbalanced, with flexible wrappers comprising more than half of the image records, motivating the use of macro-averaged metrics in addition to micro accuracy. The dataset is split into 282/58/62 train, validation, and test image records, corresponding to 95/20/21 base images and 1,692/348/372 QA pairs, respectively, and no `source_base_id` appears in more than one partition. This grouping prevents visually near-identical augmented samples from appearing across development and evaluation sets, which could otherwise inflate performance. Because the 62 test records correspond to only 21 independent base images, uncertainty is estimated at the base-image-group level rather than by treating augmented records as independent. The complete benchmark and evaluation flow is summarized in Figure 1.

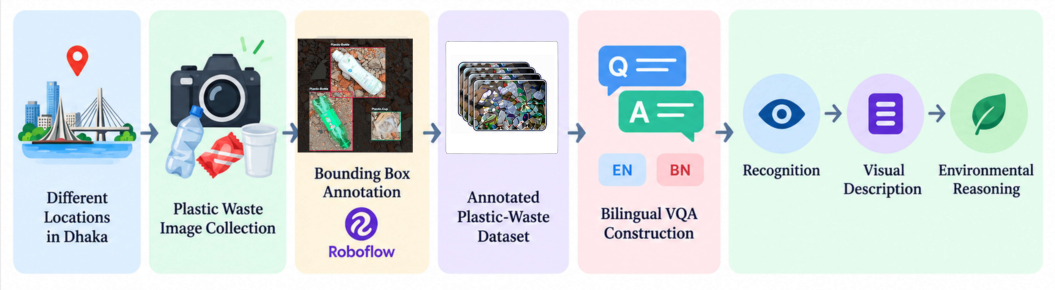


Figure 2: Data collection and bilingual VQA benchmark construction pipeline. Plastic-waste images are collected from different locations in Dhaka, annotated using bounding boxes, organized into an annotated dataset, and converted into English–Bangla VQA pairs spanning recognition, visual description, and environmental reasoning.

Each image record is associated with six QA pairs: English recognition, Bangla recognition, English visual description, Bangla visual description, English environmental reasoning, and Bangla environmental reasoning. This produces 2,412 QA pairs, evenly divided into 1,206 English and 1,206 Bangla instances. Question construction follows a controlled template-based procedure in which surface wording varies while task semantics are preserved. All 2,412 questions have unique surface strings in the current release. Reference answers are generated from a constrained, category-conditioned answer bank constructed from the collected bilingual responses. The design therefore favors controlled paired comparison over naturalistic linguistic diversity. Repeated answer semantics can also make some questions, especially generic environmental questions, partly recoverable from language and domain priors; we account for this limitation when interpreting the image-ablation results. The overall data-collection and construction workflow is illustrated in Appendix C, Figure 2, and the trade-offs are discussed in Appendix G.

The overall dataset collection and construction process is illustrated in Figure 2

C.1 Quality Control

The benchmark records image-quality and generation-status metadata. Of the 402 images, 385 are labeled `clear` and 17 `moderate`; none are retained as fully ambiguous. Seventeen images carry a low-image-quality warning, while one source has an empty label file and one has a missing label file.

Our construction is intentionally conservative but semi-automated. Category mappings and QA generation are not equivalent to independent expert annotation of every image. This distinction is essential: the benchmark should be interpreted as a controlled diagnostic dataset rather than a naturalistic human-authored VQA corpus.

C.2 Task Types

The benchmark contains exactly 804 QA pairs for each task type.

Recognition (Closed-form). Questions ask which plastic-waste category is visually dominant. Answers are intentionally concise and category-centered.

Visual Description (Open-ended visual). Questions request observable properties such as general shape, flexibility, visible structure, openness, or deformation. Answers are designed to avoid claims that cannot be reliably inferred from visual evidence.

Environmental Reasoning (Open-ended domain). Questions ask why improper disposal may be environmentally problematic or what environmental concern is associated with the observed waste item. Reference answers use conservative concepts such as persistent litter, land and water pollution, movement into drains or waterways, wildlife exposure, and waste-management challenges.

Importantly, environmental answers do not claim exact degradation times, chemical toxicity, carbon footprint, contamination status, or direct medical consequences unless such evidence is externally established.

D Experimental Setup

Experiments were conducted in Google Colab using an NVIDIA Tesla T4 GPU, with the pipeline implemented in Python using the required deep-learning and vision-language libraries. The T4 GPU was used for dataset preprocessing, local inference with the open-source VLMs, and metric computation, while the proprietary models were accessed through their respective APIs. We evaluate four representative VLMs spanning proprietary and openly available model families: GPT-4.1 mini, using the fixed API snapshot `gpt-4.1-mini-2025-04-14` where available, which supports image inputs and improved multimodal understanding and instruction following [OpenAI, 2025]; Gemini 2.5 Flash, which serves as a proprietary multimodal baseline optimized for efficient reasoning [Google DeepMind, 2025]; Qwen2.5-VL-7B-Instruct, which supports dynamic-resolution visual processing and multilingual vision-language interaction [Bai et al., 2025]; and the multimodal Gemma-3-4B Instruction-Tuned model, which provides vision understanding and expanded multilingual capabilities [Gemma Team et al., 2025]. The comparison is intended to represent different deployment regimes, model scales, and accessibility conditions rather than to establish a universal ranking, since model families and API implementations may evolve over time .

D.1 Zero-Shot Evaluation and Metrics

All models are evaluated in a zero-shot setting without fine-tuning on the training or validation split. For the primary multimodal condition, each model receives the image together with the question in its original language and a short instruction encouraging direct, visually grounded, and scientifically conservative responses. A representative prompt is: "Answer the question using the image. Use only visually supported information for recognition and description. For environmental questions, provide a concise scientifically conservative answer. Do not invent brands, exact chemical properties, locations, toxicity, or hidden objects." Temperature is set to 0 where supported, output length is capped to avoid unnecessary elaboration, and Bangla questions are processed directly without translation into English. For recognition, model outputs are normalized to the five target categories using a deterministic synonym mapping; for example, "bottle," "plastic bottle," and "bottle-like plastic container" are mapped to `plastic_bottle`. Recognition performance is measured using micro accuracy,

$$\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\hat{y}_i = y_i\right], \tag{1}$$

where y_i and $\hat{y}_i$ denote the reference and predicted categories, respectively, and macro-F1,

$$F1_{\mathrm{macro}} = \frac{1}{C}\sum_{c=1}^{C} F1_c, \tag{2}$$

where C is the number of waste categories. Macro-F1 is reported because the benchmark is class-imbalanced and flexible-wrapper examples are dominant. For open-ended visual-description and environmental-reasoning responses, exact match is insufficient because semantically valid answers may differ in wording. We therefore combine semantic and lexical metrics. BERTScore-F1 is used as the primary automatic metric because it compares candidate and reference answers using contextual embeddings and is more tolerant of paraphrasing [Zhang et al., 2020]. BLEU-4 measures modified n-gram precision [Papineni et al., 2002], while ROUGE-L captures longest-common-subsequence overlap [Lin, 2004]; both are reported as complementary lexical measures. Because English and Bangla differ in morphology, segmentation, and surface realization, lexical scores are interpreted cautiously and are not treated as perfectly language-neutral. The BERTScore encoder and scoring configuration can also affect cross-lingual comparability and must therefore be reported explicitly. For reproducibility, the experiments record the exact model or API snapshot, inference date, decoding parameters, package versions, and model configuration.

D.2 Bilingual Consistency

Independent English and Bangla scores do not establish whether a model gives semantically equivalent answers to equivalent questions. We therefore compute a Bilingual Consistency Score (BCS). For recognition, consistency is defined as category agreement across paired English and Bangla outputs. For open-ended answers, paired responses are represented in a multilingual sentence-embedding space using LaBSE, which was developed to provide language-agnostic sentence representations across more than 100 languages [Feng et al., 2022]. Given paired English and Bangla outputs e_i and b_i, respectively, we compute

$$s_i = \cos\left(\mathrm{LaBSE}(e_i), \mathrm{LaBSE}(b_i)\right). \tag{3}$$

After validating a semantic threshold τ on a manually reviewed validation subset, BCS is defined as

$$BCS = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[s_i \geq \tau\right]. \tag{4}$$

The threshold τ is selected on a manually reviewed validation subset and then held fixed during test evaluation.

D.3 Visual-Grounding Ablation

Our central diagnostic removes the image while leaving the question unchanged.

The multimodal condition is

$$\mathcal{M} : (I, Q) \rightarrow A, \tag{5}$$

whereas the language-only condition is

$$\mathcal{L} : Q \rightarrow A. \tag{6}$$

For an evaluation metric S, we define Visual Grounding Gain (VGG) as

$$VGG = S(\mathcal{M}) - S(\mathcal{L}). \tag{7}$$

Within a fixed benchmark and metric, a larger positive value indicates greater performance sensitivity to image removal. A small value can arise because the model relies weakly on the image, because the question can be answered from linguistic or domain priors, or both. VGG is therefore a diagnostic of benchmark-conditioned image sensitivity rather than a standalone causal measure of visual grounding.

This distinction is particularly important for environmental reasoning because general facts concerning plastic pollution may remain recoverable even when the model does not know whether the image contains a bottle, wrapper, cup, bag, or straw.

D.4 Statistical Analysis

Augmented variants originating from the same base image are statistically correlated. Treating all 62 test records as independent observations would therefore underestimate uncertainty. We use a cluster-bootstrap procedure based on `source_base_id`, rather than an ordinary QA-level bootstrap. For uncertainty analyses, base-image groups are sampled with replacement using bootstrap resampling. All augmented records and associated QA pairs belonging to a sampled base image are retained together. Figures report percentile-based 95% confidence intervals. Pairwise model comparisons use the same bootstrap samples across models, producing paired bootstrap estimates of performance differences. Small differences in point estimates should not be treated as robust model ordering unless the paired uncertainty on the difference supports that interpretation.

E Extended Results and Discussion

E.1 Category-Level Recognition

The difference between recognition accuracy and macro-F1 is consistent with the class imbalance in Appendix C: performance on the dominant flexible-wrapper class has greater influence on micro accuracy. Category-level results reveal greater sensitivity for bag and straw examples, which have fewer samples and can visually overlap with flexible packaging.

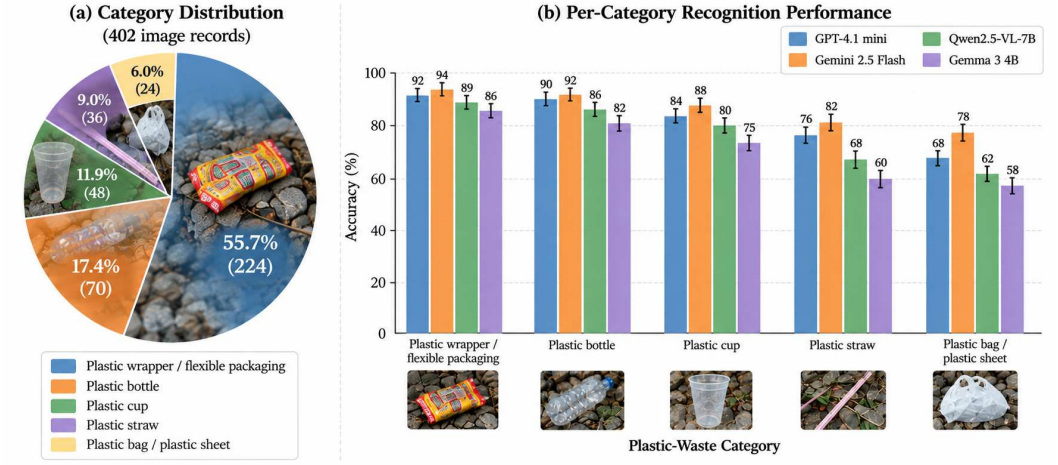


Figure 3: Category distribution and per-category recognition performance. Error bars represent 95% cluster-bootstrap confidence intervals over source base images.

E.2 Detailed English–Bangla Results

Table 5: Language-level performance and bilingual consistency. Normalized performance is on a 0–100 scale and BCS on a 0–1 scale.

	Language Performance (0–100)			Bilingual Consistency (0–1)		
Model	**English**	**Bangla**	**Δ EN–BN**	**Rec.**	**Desc.**	**Env.**
GPT-4.1-mini	89.23	85.87	3.36	0.94	0.89	0.84
Gemini 2.5 Flash	**90.60**	**87.43**	**3.17**	**0.96**	**0.91**	**0.87**
Qwen2.5-VL-7B	85.90	82.23	3.67	0.91	0.85	0.79
Gemma-3-4B	83.10	78.20	4.90	0.87	0.81	0.73
Mean	–	–	3.78	–	–	–

Table 6: Language-specific open-ended evaluation.

Language	**BERT-F1**	**BLEU-4**	**ROUGE-L**
English	0.842	0.382	0.574
Bangla	0.806	0.329	0.531

The BLEU gap is proportionally larger than the BERTScore gap, suggesting that some Bangla responses may remain semantically appropriate despite greater surface-form variation. BLEU should therefore not be used alone to infer weaker Bangla reasoning.

E.3 Detailed Image-Ablation Interpretation

Recognition behaves as expected: removing the image produces a large performance collapse. Description shows a similarly large reduction, indicating that reference-aligned answers to these

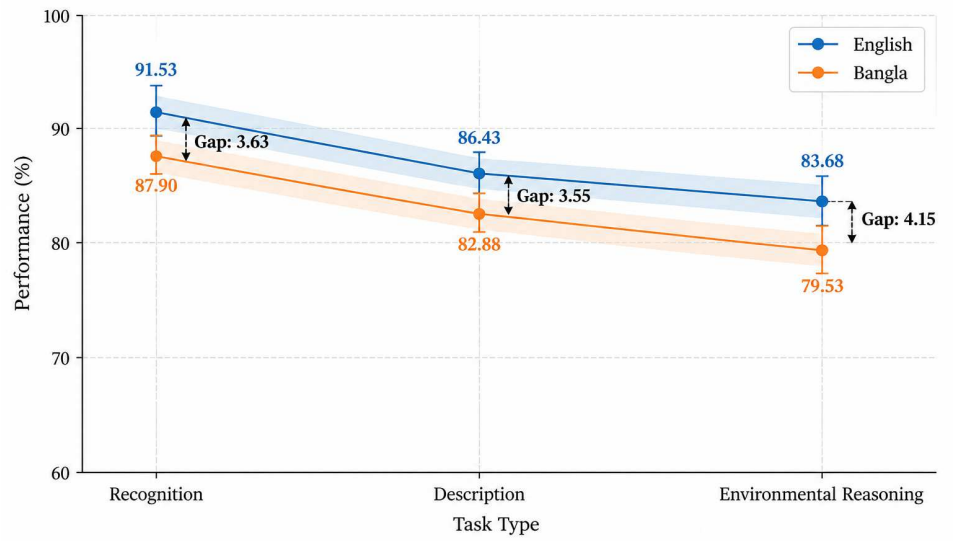


Figure 4: English–Bangla performance gap across recognition, description, and environmental reasoning. Error bars represent 95% cluster-bootstrap confidence intervals over source base images.

benchmark questions are strongly sensitive to visual evidence. Environmental reasoning behaves differently: 70.18 normalized performance remains without the image, compared with 81.60 under multimodal input. This indicates that the current environmental-question templates and reference semantics permit many broadly acceptable answers to be recovered from linguistic and domain priors. Consequently, VGG should be interpreted as a benchmark-conditioned sensitivity diagnostic rather than a universal measure of grounding.

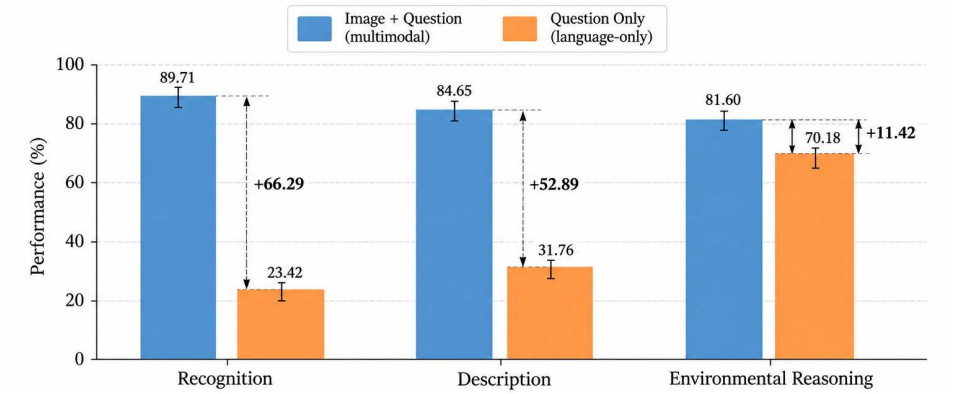


Figure 5: Visual-ablation analysis. Recognition and description exhibit substantially larger multimodal gains than environmental reasoning.

E.4 Cross-Lingual Consistency

A model may obtain respectable aggregate scores in both languages while producing contradictory answers to paired questions. The observed BCS pattern declines from recognition to description to environmental reasoning. Recognition is stable because outputs often reduce to the same short category; open-ended tasks create more opportunities for differences in specificity, factual content, and grounding. Because task format and response length change simultaneously, this trend should not be interpreted as an isolated effect of reasoning complexity.

Table 7: Qualitative error distribution within the manually audited subset.

Error Type	English	Bangla
Object misclassification	6.5%	9.7%
Visual hallucination	5.4%	8.1%
Generic environmental reasoning	15.1%	**21.5%**
Environmental hallucination	4.3%	6.5%
Bangla interpretation error	–	7.5%

E.5 Extended Discussion

Recognition accuracy alone is insufficient to characterize environmental multimodal capability. The English–Bangla gap is largest for environmental reasoning, but the comparison also reflects differences in response length, answer space, and linguistic generation demands. A methodological advantage of the benchmark is that English and Bangla questions are paired for the same image and semantic intent, reducing visual-content confounds.

The image-ablation experiment exposes a benchmark-design issue as well as a model-behavior issue. If an environmental question can be answered nearly as well without the image, the score partly reflects environmental language knowledge and question priors. Future benchmark extensions should reduce such shortcuts through counterfactual visual pairs, where similar question structures are paired with different waste categories and the reference answer requires category-specific consequences.

Automatic metrics capture different aspects of response quality. BLEU rewards local n-gram precision [Papineni et al., 2002], ROUGE-L measures longest-common-subsequence overlap [Lin, 2004], and BERTScore is more tolerant of semantic paraphrases [Zhang et al., 2020]. None directly establishes visual grounding or scientific correctness; the benchmark therefore combines them with modality ablation and qualitative audit. The proprietary models have higher point estimates than the selected open checkpoints, but this comparison spans different scales, training corpora, architectures, and deployment constraints and should not be interpreted as evidence that proprietary VLMs are inherently superior.

F Qualitative Error Analysis

F.1 Error Analysis

We manually audited a stratified subset of predictions covering models, languages, and question types. The annotation protocol distinguishes five failure categories. Because this analysis is conducted on a subset rather than the full test set, the resulting percentages are interpreted descriptively rather than as population estimates.

Object misclassification. The dominant plastic item is assigned to an incorrect category.

Visual hallucination. The model invents a visible property, object, label, color, cap, opening, or condition that is not sufficiently supported by the image.

Generic environmental reasoning. The answer is environmentally plausible but nearly independent of the specific image or waste category.

Environmental hallucination. The model makes unsupported claims concerning toxicity, chemical composition, health consequences, or other scientific properties.

Bangla interpretation error. The model misinterprets the intent of the Bangla question or responds with semantically inappropriate content.

The observed qualitative error distribution is shown in Table 7.

The observed error distribution identifies generic environmental reasoning as the dominant failure mode, particularly for Bangla. This behavior aligns with the visual-ablation analysis: models may

possess broadly correct knowledge about plastic pollution while failing to condition that knowledge strongly on the precise visual evidence.

G Limitations and Threats to Validity

Dataset size, diversity, and statistical power. The benchmark contains 136 underlying base images, and the test split contains only 21 independent base images. The 402 image records include augmented variants, so the effective visual diversity and independent test support are substantially smaller than the raw record count suggests. Cluster bootstrap addresses dependence among augmentations, but it cannot create additional independent visual diversity. The benchmark should therefore be interpreted as a focused diagnostic resource rather than a comprehensive representation of global plastic pollution, and small model-to-model differences should be treated cautiously.

Category imbalance and category overlap. Flexible wrappers account for 55.7% of image records, whereas plastic bags represent only 6.0%. Micro accuracy can therefore overstate performance on the dominant class. Macro-F1 and per-category evaluation reduce, but do not eliminate, this limitation. In addition, categories such as "plastic wrapper / flexible packaging" and "plastic bag / plastic sheet" can overlap visually. Future releases should publish explicit category decision rules and human agreement on ambiguous examples, together with per-category counts at the independent base-image level.

Controlled rather than naturalistic language. Questions are generated through controlled templates and reference answers are drawn from a constrained answer bank. This improves consistency and enables paired bilingual comparisons but yields less linguistic diversity than human-authored VQA. The current benchmark contains 2,412 unique question surface forms but only 100 distinct reference answer strings, so models may benefit from repeated answer semantics. This design can also make generic environmental questions partly answerable without an image, which means the small environmental VGG is partly a property of benchmark construction as well as model behavior. Future versions should incorporate independently authored references, greater paraphrastic diversity, and counterfactual or category-specific questions whose correct answer changes with the visual evidence.

Annotation and bilingual validity. Benchmark construction is semi-automated rather than independently expert-annotated at the full-dataset level. This creates a potential confound between model multilingual weakness and unnatural or imperfect Bangla wording. Before release as a high-stakes public benchmark, the full test set or a clearly specified stratified subset should be reviewed by native Bangla annotators for question naturalness and semantic equivalence, and by annotators for object category, visual grounding, and scientific correctness.

Cross-task comparability. Recognition, visual description, and environmental reasoning differ not only in semantic demands but also in answer length, answer space, evaluation metric, and required domain knowledge. Differences across these task types therefore should not be interpreted as a controlled causal effect of "reasoning complexity." The normalized score is a reporting convenience that places the primary task-specific metrics on a common numerical scale; it does not make the metrics psychometrically equivalent.

Environmental scope. Environmental answers intentionally remain general and scientifically conservative. The benchmark does not evaluate polymer chemistry, life-cycle assessment, detailed ecological toxicity, local waste regulations, or contamination-specific risks. Models should therefore not be interpreted as environmental-science experts on the basis of this benchmark.

Model drift. Closed VLM APIs may change over time. Reproducibility requires fixed snapshots where possible and detailed recording of model identifiers, inference dates, prompts, decoding parameters, and provider-side changes.

Metric validity. BERTScore, BLEU, and ROUGE compare generated responses with textual references; none guarantees factual correctness. Their behavior can also differ across English and Bangla, making the exact BERTScore encoder and configuration important for interpreting language

gaps. Similarly, LaBSE similarity may classify two fluent but factually incorrect responses as cross-lingually consistent, and BCS depends on the selected semantic threshold. Human review remains necessary for high-confidence conclusions.

Dependence among augmented samples. Augmented images derived from the same base image are statistically dependent. Confidence intervals based on individual image records or QA pairs would therefore be artificially narrow. Our source-group cluster bootstrap explicitly addresses this dependence.