Waste-Bench: A Comprehensive Benchmark for Evaluating VLLMs in Cluttered Environments

Anonymous ACL submission

Abstract

Recent advancements in Large Language Models (LLMs) have paved the way for Vision Large Language Models (VLLMs) capable of performing a wide range of visual understanding tasks. While LLMs have demonstrated impressive performance on standard natural images, their capabilities have not been thoroughly explored in cluttered datasets where there is complex environment having deformed shaped objects. In this work, we introduce a novel dataset specifically designed for waste classification in real-world scenarios, characterized by complex environments and deformed shaped objects. Along with this dataset, we present an in-depth evaluation approach to rigorously assess the robustness and accuracy of VLLMs. The introduced dataset and comprehensive analysis provide valuable insights into the performance of VLLMs under challenging conditions. Our findings highlight the critical need for further advancements in VLLM's robustness to perform better in complex environments. The dataset and code for our experiments will be made publicly available.

1 Introduction

In recent years, Large Language Models (LLMs) (Chung et al., 2024; Achiam et al., 2023; Touvron et al., 2023) have demonstrated remarkable capabilities in understanding, reasoning, and generating text for a diverse range of open-ended tasks. Models such as PaLM 2 (Anil et al., 2023) and Falcon (Penedo et al., 2023) have showcased exceptional performance in commonsense reasoning, multilingual applications, and various Natural Language Processing (NLP) tasks. Building on their success, Vision-Language Large Models (VLLMs) (Fang et al., 2023; Touvron et al., 2023; Zheng et al., 2023) have emerged, extending these capabilities to multimodal domains by integrating visual and textual data. Notable examples, including multimodal GPT-4 and open-source models like LLaVA (Achiam et al., 2023; Liu et al., 2023, 2024), excel in a variety of multimodal tasks, demonstrating their versatility in real-world applications (Hu et al., 2023; Vinyals et al., 2015; Chou et al., 2020).

Despite advancements in Vision-Language Models (VLLMs), their application in complex, cluttered environments remains underexplored. Traditional object detectors, such as Faster R-CNN (Ren, 2015) and YOLO (Redmon, 2016), are effective for visual localization and classification tasks. However, these models are inherently limited to predefined categories and lack the multimodal reasoning capabilities needed to address open-ended queries or context-aware tasks. In contrast, VLLMs leverage the ability to align visual and textual modalities, enabling them to process complex queries like, "Which items in the scene are recyclable under current lighting conditions?" or "How many soft plastic items are overlapping with metal objects?" These reasoning tasks are critical in domains such as waste classification, where cluttered scenes and diverse object categories introduce significant challenges.

To address these challenges, we propose Waste-Bench, a benchmark designed to evaluate the robustness and reasoning capabilities of VLLMs in the context of waste classification. Unlike existing benchmarks, such as SEED-Bench (Li et al., 2023) and MV-Bench (Li et al., 2024), which focus primarily on general visual comprehension, Waste-Bench targets the unique complexities of real-world waste management scenarios, including cluttered scenes, deformed objects, and ambiguous visual cues. By systematically evaluating pre-trained VLLMs, Waste-Bench highlights their baseline capabilities and limitations, offering actionable insights to guide the improvement of future VLLMs.

Furthermore, Waste-Bench is intended to complement existing datasets, enriching them with challenging scenarios that encourage greater robustness and adaptability in models. By incorporating diverse data distributions into training pipelines, models can achieve better trade-offs between task-specific robustness and generalization. This approach aligns with robust learning paradigms, which suggest that exposure to diverse, challenging data distributions can enhance model generalization while minimizing the risks of performance degradation on simpler tasks (Havrilla et al., 2024). To improve VLLMs in such environments, techniques like domain adaptation and adversarial training can be employed to expose the models to more realistic, noisy, and cluttered data. Additionally, incorporating multi-modal learning, including multispectral data, and using data augmentation strategies during training can help VLLMs better adapt to complex, cluttered environments. Fine-tuning models on Waste-Bench's diverse and complex scenarios ensures that they become more robust to variations in visual cues, allowing them to handle the unique challenges of waste classification tasks effectively.

Models trained on simpler datasets often experience a performance drop when evaluated in cluttered environments, primarily due to insufficient exposure to noise, occlusions, and ambiguities during training. To address this challenge, Waste-Bench exposes models to more complex and realistic waste classification scenarios. By training models on these challenging conditions, Waste-Bench helps to reduce the performance gap between regular and cluttered environments, improving model generalization without sacrificing accuracy. Although the performance discrepancy between regular and cluttered environments has not been extensively studied in VLLMs, this issue is well-known in traditional vision tasks. In literature, various waste classification methods have been proposed (Xia et al., 2024; Mao et al., 2021; Feng et al., 2022; Meng et al., 2022), they pose limitations in the presence of complex scenarios where there exists an unclear boundary information. Waste-Bench aims to mitigate this gap by training models on more challenging, real-world data, making them more adaptable and robust. Our contributions are as follows:

- A Waste-Bench designed to evaluate the robustness and reasoning capabilities of VLLMs in waste classification, addressing the complexities of real-world applications.
- We evaluate VLLMs, uncovering significant challenges, especially in reasoning within cluttered scenes with deformed objects.
- We identify that VLLMs struggle with various tasks on Waste-Bench, guiding future waste management improvements.

2 Related Work

Vision Large Language Models (VLLMs) (Zhu et al., 2024; Shao et al., 2023) have demonstrated remarkable capabilities in engaging with visual content, offering a wide range of potential applications. Notable models in this domain include Qwen (Bai et al., 2023), which has consistently demonstrated superior performance across various downstream tasks. Gemini-Pro and GPT-4o (Reid et al., 2024; OpenAI, 2024) exemplifies state-of-the-art performance with its advanced reasoning and interaction capabilities, paving the way for the development of versatile multimodal conversational assistants. All these models perform extremely well on wide range of image understanding tasks like caption generation, visual question answering and so on. These models accept both visual and textual inputs and generate textual responses. From an architectural perspective, VLLMs typically combine pre-trained vision backbones (Fang et al., 2023) with large language models (Touvron et al., 2023; Zheng et al., 2023) using connector modules such as MLP adapters, Q-former (Dai et al., 2024), and gated attention (Alayrac et al., 2022).

Benchmarking VLLMs With the growing number of VLLMs emerging in the research community, several benchmarks have been proposed to evaluate and quantify these models for benchmarking and analysis purposes. Notable benchmarks in this domain include SEED-Bench (Li et al., 2023), which evaluates the visual capabilities of both image and video LMMs across multiple dimensions, and MV-Bench (Li et al., 2024), which curates challenging tasks to evaluate the spatial and temporal understanding of VLLMs. While these benchmarks provide effective insights into model performance, they primarily focus on general visual comprehension metrics. However, none of them specifically target complex cluttered environments and deformed shaped objects. In contrast, Waste-Bench is a comprehensive benchmark designed to assess the robustness and reasoning capabilities of VLLMs in waste classification.

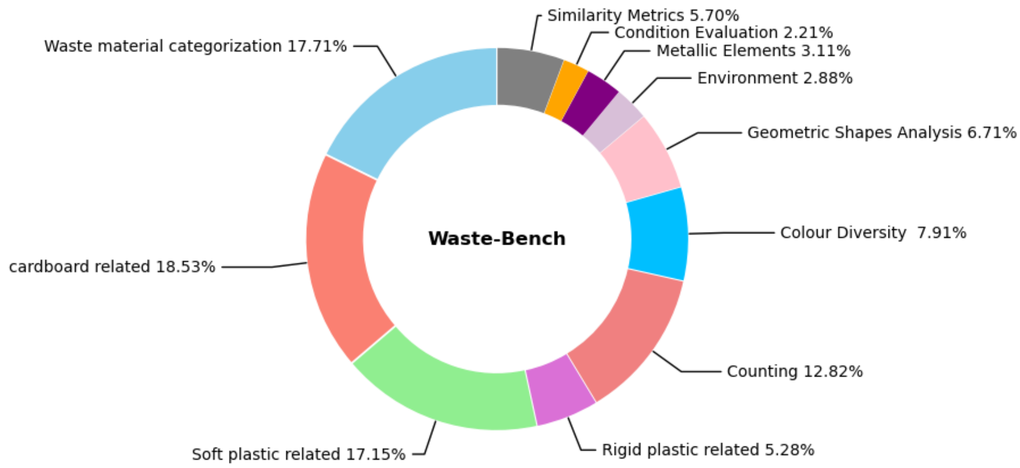


Figure 1: Waste-Bench comprises of 11 diverse complex question categories encompassing a variety of waste images context.

3 Waste-Bench

In this work, our objective is to develop a comprehensive benchmark to evaluate the robustness and reasoning capabilities of VLLMs in various complex and cluttered visual environments, spanning diverse scenarios. To achieve this, we introduce Waste-Bench. Initially, we offer a holistic overview of Waste-Bench and outline the diversity of questions it contains. Following this, we detail the creation process of Waste-Bench in Section 3.2. Performance evaluation including experiments and results are given in Section 4 and 5 respectively.

3.1 Waste-Bench Dataset

Waste-Bench encompasses 11 different question categories and 9,520 high-quality open-ended question-answer (QA) pairs, spanning 952 high-quality images with an average of 10 questions per image. These questions cover diverse categories related to real-world waste classification scenarios, including individual classification of waste classes, multi-class classification, shapes of objects, and colors. This comprehensive dataset is designed to rigorously test the capabilities of VLLMs in handling complex and cluttered visual environments. The question types and word cloud of frequent keywords is given in Appendix A.2.

3.1.1 Waste-Bench Different Question Types

To assess the robustness and reasoning capabilities of VLLMs in the Waste-Bench benchmark, we ensure it contains various question types to encompass a wide range of real-world complex and cluttered visual environments within each image. Below, we provide a detailed definition of the Waste-Bench as given in Figure 1.

- Single Class Classification (Cardboard, Metal, Soft Plastic, Rigid Plastic): This category includes questions that require the model to classify individual waste items into one of the specified single classes. The questions aim to determine whether the model can accurately identify and distinguish between different types of materials commonly found in waste.

- Multiclass Categorization: In this category, the models are challenged with images containing multiple deformed waste items that need to be classified into more than one category. The goal is to assess the model's ability to handle complex scenes where multiple waste types are present and need to be accurately categorized.

- Counting: This category involves tasks where the model must count the number of specific items or categories within an image. For example, counting the number of cardboard pieces or the number of recyclable items in a cluttered environment. The questions are designed to evaluate the model's precision in quantifying objects in a scene.

- Color Diversity: This question type tests the model's ability to distinguish and identify items based on color. Tasks in this category include identifying objects of a specific color or categorizing items by color diversity. It assesses the model's capability to utilize color as a key feature in classification.

- Geometric Shape Analysis: This category of questions focuses on the model's ability to recognize and categorize objects based on their geometric shapes. Questions involve identifying items with specific shapes, such as cylindrical, circular or rectangular objects, which are common in waste sorting processes.

- Complex and Cluttered Environment: This category includes questions to evaluate the model's performance in recognizing and reasoning about the environment in which waste is found. Model evaluates whether waste is in an indoor or outdoor setting. It includes questions that require comprehensive image analysis.

- Condition Evaluation: In this category, the model must evaluate the condition of waste items. This includes assessing whether items are intact, twisted, clean or dirty. The questions are designed to test the model's ability to make nuanced judgments about the state of objects.

- Similarity Metric: These questions require the model to compare and determine the similarity between different waste items. For example, identifying items that belong to the same category or have similar features. It assesses the model's ability to draw comparisons and make associations based on visual features, robustness in recognizing objects in challenging settings, and adaptability to varying conditions.

- Combined Classification and Counting: This category merges classification and counting tasks, requiring the model to not only classify multiple items in a scene but also provide accurate counts for each category. This combined approach tests the model's capability to perform multiple reasoning tasks simultaneously.

These question types present in our dataset help to rigorously test the capabilities of VLLMs in handling the intricacies of waste classification in complex and cluttered environments.

3.2 Building Waste Bench Benchmark

The Waste-Bench benchmark is carefully constructed through a four-step process using a dataset of 952 images. Initially, 11,424 Question/Answer (Q/A) pairs are generated, capturing information from the images. With filtering process given in Stage 1, this number is reduced to 9,520, ensuring relevance and quality. A focused refinement filtered out 1,920 Q/A pairs, representing approximately 20% of the original set. Each step is presented in detail below, and can be visually explored in Figure 2.

Stage 1: Data Collection and Annotation We thoroughly reviewed various datasets and used ZeroWaste (Bashkirova et al., 2022) with waste images in cluttered environment. We pre-processed the metadata provided with the images to ensure accurate representation of the categories assigned to each image. Following the image collection process, the Gemini-Pro v1.5 was employed for generating descriptive captions, optimized specifically for this task, while Gemini-Pro v1.0 (49.45% accuracy) was used for classification tasks. Caption generation focused solely on producing accurate, context-rich descriptions, distinct from classification challenges. A rigorous human-in-the-loop process ensured caption quality and refined captions for relevance, clarity, and accuracy. Inter-annotator agreement confirmed consistency, and benchmarking demonstrated Gemini-Pro v1.5's captions were competitive with state-of-the-art models. Two human assistants reviewed, filtered and corrected the generated captions as well as generated question answers based on following verification criteria.

- Ensured the content's relevance to the images and accuracy. Both of the human assistants verified the content relevant to the image and if the content is more than 80% related we will keep it otherwise, discard it.

- Clear and coherent language.

- Accurate and relevant answer

The prompt used for generating captions is provided in Figure 2. These prompts included ground truth information (e.g., class names, categories, and

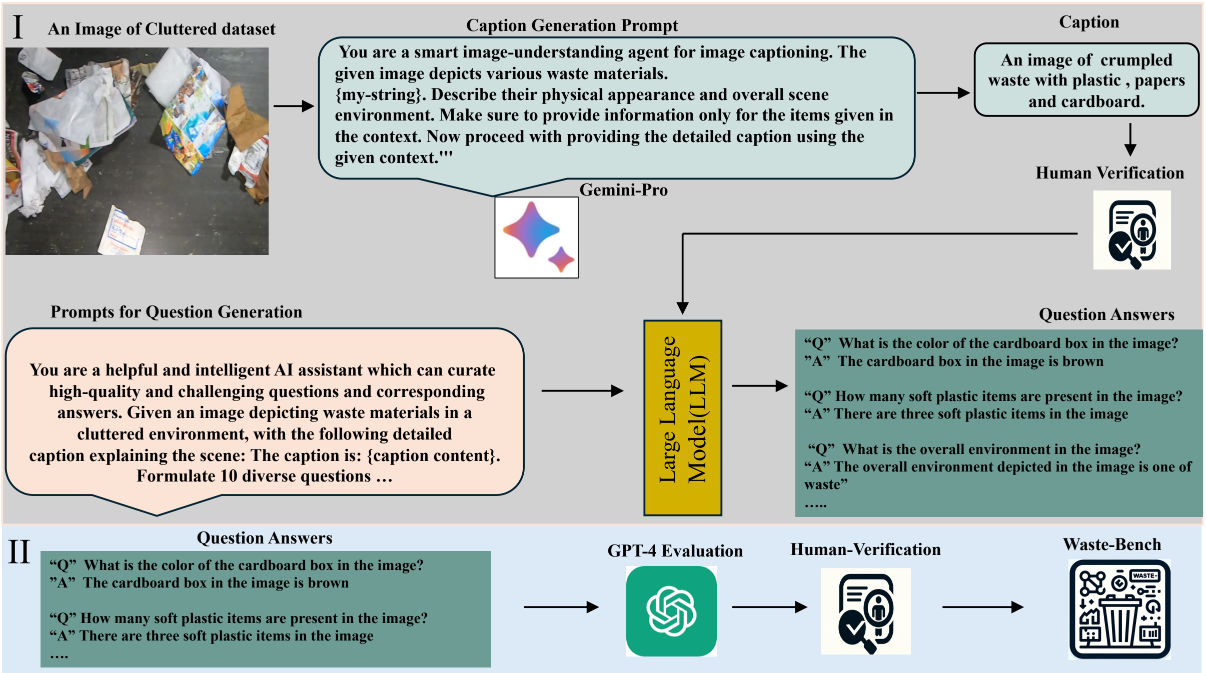
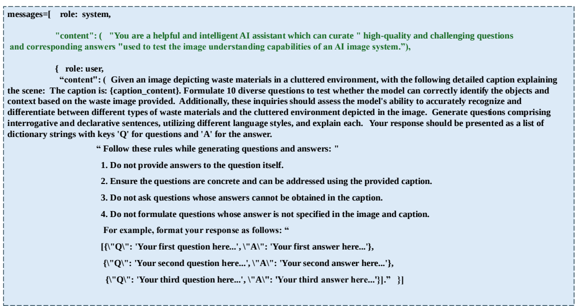


Figure 2: Step I: Gemini-Pro generates detailed waste image captions, verified by human annotators. Step II: Nearly 10k diverse questions are generated from these captions, evaluated by GPT-4, and verified by humans.

masks) from the dataset's JSON annotations to guide LLMs in producing contextually accurate outputs.

Stage 2: Question-Answer Generation Inspired


```
messages=[    role:  system,

        "content": (   "You are a helpful and intelligent AI assistant which can curate " high-quality and challenging questions
 and corresponding answers "used to test the image understanding capabilities of an AI image system."),

        {   role: user,
         "content": (  Given an image depicting waste materials in a cluttered environment, with the following detailed caption explaining
the scene:  The caption is: {caption_content}. Formulate 10 diverse questions to test whether the model can correctly identify the objects and
context based on the waste image provided.  Additionally, these inquiries should assess the model's ability to accurately recognize and
differentiate between different types of waste materials and the cluttered environment depicted in the image.  Generate questions comprising
interrogative and declarative sentences, utilizing different language styles, and explain each.   Your response should be presented as a list of
dictionary strings with keys 'Q' for questions and 'A' for the answer.
                  " Follow these rules while generating questions and answers: "
                   1. Do not provide answers to the question itself.
                   2. Ensure the questions are concrete and can be addressed using the provided caption.
                   3. Do not ask questions whose answers cannot be obtained in the caption.
                   4. Do not formulate questions whose answer is not specified in the image and caption.
                    For example, format your response as follows: "
                   [{\"Q\": 'Your first question here...', \"A\": 'Your first answer here...'},
                    {\"Q\": 'Your second question here...', \"A\": 'Your second answer here...'},
                     {\"Q\": 'Your third question here...', \"A\": 'Your third answer here...'}]."   }]
```

Figure 3: Prompts Used for Generating Question-Answer Pairs.

by human interaction in day-to-day life, we aim to simulate a similar style of interaction with VLLMs by curating open-ended QA pairs to evaluate these models for robustness and reasoning. We feed detailed ground-truth image captions to GPT-3.5, which are utilized to generate open-ended questions covering both reasoning and robustness aspects.

The questions designed go beyond basic image comprehension, requiring complex logical inference and contextual understanding. These questions test the model's ability to classify objects by recognition, color, shape, and other relevant aspects in complex settings, ensuring accurate and appropriate responses. Prompt used for curating QA pairs is mentioned in Figure 3.

Stage 3: QA Pairs Filtration

After generating QA pairs, a human-in-the-loop review involving two human assistants identified approximately 20% of the pairs as noisy. These noisy pairs included irrelevant, unanswerable, or repetitive questions, such as those with answers embedded within the questions. To address these issues, an exhaustive filtering process was conducted, ensuring that the QA pairs met the relevance and alignment criteria based on the image evaluation.

For the review process, we applied similar rules as those used for caption generation. Two human assistants reviewed the question-answer pairs based on the following criteria:

- The QA pairs needed to be related to the verified captions, with both assistants agreeing that the content was over 80% relevant to the image.
- The language was checked for clarity.
- Answers were verified for accuracy and relevance.

This process ensured that only relevant, accurate, and clear question-answer pairs were retained, resulting in a curated set of 9,552 high-quality QA

"You are an intelligent chatbot designed for evaluating the correctness of AI assistant predictions for question-answer pairs. Your task is to compare the predicted answer with the ground-truth answer and determine if the predicted answer is correct or not. Here's how you can accomplish the task: "------",

INSTRUCTIONS:
- Focus on the correctness and accuracy of the predicted answer with the ground truth.
- Consider predictions with less specific details as correct evaluation, unless such details are explicitly asked in the question.

Please evaluate the following question-answer pair:
Question: {question}, Ground truth correct Answer: {ground truth}, Predicted Answer: {predicted}
Provide your evaluation as a correct/incorrect prediction along with the score where the score is an integer value between 0 (fully wrong) and 5 (fully correct). The middle score provides the percentage of correctness.
Please generate the response in the form of a Python dictionary string with keys 'predicted', 'score', and 'reason', where the value of 'predicted' is a string of 'correct' or 'incorrect', the value of 'score' is in INTEGER, not STRING, and value of 'reason' should provide the reason behind the decision.

Only provide the Python dictionary string. For example, your response should look like this: {'predicted': 'correct', 'score' : 4.8, 'reason': reason}."

Figure 4: Evaluation prompt used.

pairs. These pairs provide a robust foundation for the **Waste-Bench** benchmark. Appendix A.1 provides a quantitative overview of the results.

Stage 4: Evaluation Procedure Previous methods like MM-VET(Yu et al., 2023) and SEED-BENCH (Li et al., 2023) have used LLMs as judges for open-ended QA benchmarks. We follow a similar approach, employing GPT-4 to evaluate the correctness of VLLM predictions against ground-truth answers. VLLMs generate predictions based on image-question pairs, which are then assessed by GPT-4 through binary judgments, with reasoning provided for each decision. The evaluation prompt as given in Figure 4, used in our study was designed to guide the LLMs in assessing the accuracy and quality of the responses generated by VLLMs on the Waste-Bench dataset. This prompt provided the LLMs with specific instructions to compare the model-generated answers with ground-truth answers, make binary correctness judgments. The prompt also emphasized the importance of providing reasoning for each evaluation, ensuring that the judgments were not only accurate but also interpretable and consistent. To ensure accuracy, two assistants reviewed the evaluation results. To validate the performance across all models, we observed a high consistency between GPT-4 and human evaluations, as given in Table 1 below.

	GPT		Human	
Model	**CogVLM**	**InstructBLIP**	**InstructBLIP**	**CogVLM**
Performance	45%	59%	63%	46%

Table 1: Comparison of model performance between GPT and Human evaluations across different models.

4 Performance Evaluation on Waste-Bench

Both open-source and closed-source models were explored and selected for evaluation. In total, seven models were evaluated. Among the open-source models, five recent VLLMs were included: InstructBLIP, LLaVA-1.6, CogVLM, Qwen-VL, and MiniGPT-4. For closed-source models, GPT-4o and Gemini-Pro were used. Our work focuses on evaluating existing VLLMs to highlight their limitations in cluttered environments. While VLLMs are costly to train, our evaluation reveals key challenges, and future work will address issues like hallucination and robustness for better performance in complex tasks.

4.1 Main Experiments on Waste-Bench

All models were used in their pre-trained state to ensure a fair comparison across different architectures, detail given in in Appendix Table 6. Given the diversity of the models employed, specific hyperparameter tuning was not performed for individual models; instead, the focus was on evaluating their inherent capabilities. Each model was assessed under consistent conditions, using a single NVIDIA 24GB GPU to run the experiments, ensuring uniformity in computational resources across the tasks.

In Table 2, we present the evaluation results of diverse range of models including five open source, two closed source and human upper bound to provide comprehensive benchmark . All evaluations were conducted according to the settings specified officially as discussed in Appendix A.3 and Table 6. VLLMs find it challenging to perform well and thus show inferior performance when evaluated on the Waste-Bench dataset, particularly in cluttered scenes with deformed shaped objects. Interestingly, the performance of models like LLaVA-1.6, and InstructBLIP is relatively higher compared to models such as Qwen-VL and MiniGPT-4. For instance, Gemini achieves an accuracy of 49.45% , however MiniGPT-4 suffers severely with these particularly challenging conditions and thus under perform. The Gpt-4o model surpasses the performance of all models and achieves high gains compared to other models. However, it still remains at the lower end of performance for this type of dataset, with an accuracy around 57%. GPT-4o handles cluttered scenes with deformed shaped objects, better than others, indicating a more sophisticated understanding of complex visual contents. The Table 3 compares the performance of various VLLMs across different waste classification tasks. GPT-4 performs well in most categories, especially in Counting (60.00) and Condition Evaluation (60.00), while MiniGPT-4 shows weaker results, particularly in Single Class Classification (22.00). Models like Gemini and LLAVA exhibit moderate perfor-

Model	**Version**	**LLM**	**Accuracy (%)**
GPT-4	GPT-4o	Proprietary LLM	57.52
Gemini	Gemini-1.0 Pro	Proprietary LLM	49.45
InstructBLIP	BLIP-2_Vicuna_Instruct	Vicuna-7B	48.58
LLaVA	LLaVA-1.6	Vicuna-7B	47.45
Qwen-VL	Qwen-VL-Chat	Qwen-7B	41.30
CogVLM	CogVLM-chat-v1.1	Vicuna-7B	41.58
MiniGPT-4	MiniGPT-4	Vicuna-7B	36.40
Human Upper Bound	N/A	N/A	**81.20**

Table 2: Evaluation results VLLMs highlighting open-source and closed-source models.

Question Category	**GPT-4**	**Gemini**	**InstructBLIP**	**LLAVA**	**Qwen-VL**	**CogVLM**	**MiniGPT-4**
Single Class Classification	49.00	38.00	46.00	35.00	28.50	36.50	22.00
Multiclass Categorization	54.00	44.00	36.50	37.00	34.00	30.50	32.00
Counting	60.00	52.00	50.00	45.50	43.00	40.50	31.00
Color Diversity	42.00	35.00	39.00	48.00	38.00	27.50	30.00
Geometric Shape Analysis	55.00	49.00	44.00	41.50	45.50	39.00	36.50
Complex and Cluttered Environment	38.00	42.00	52.00	58.00	51.00	47.00	39.00
Condition Evaluation	60.00	57.00	48.50	49.50	38.00	33.00	35.00
Similarity Metric	53.50	47.00	38.50	56.00	44.50	50.50	29.00
Combined Classification and Counting	44.00	48.00	53.00	44.50	39.00	41.00	36.00

Table 3: Comparison of different models across question categories, highlighting the performance of open-source and closed-source models.

mance, with LLAVA excelling in Condition Evaluation (58.00). The values are rounded to whole numbers for simplicity and clarity.

5 Key Highlights and Qualitative Results

The evaluation of VLLMs on the Waste-Bench benchmark reveals critical insights valuable for future model development, focusing on model performance under various conditions and highlighting strengths and areas for improvement.

Real-World Waste Classification Challenges: Models that perform well on simplified environments often struggle with the complexities of Waste-Bench, particularly when it comes to counting irregularly shaped objects or accurately identifying colors in cluttered scenes. For instance, as illustrated in Figure 9, Q2, a model incorrectly predicted the color of a plastic bag due to a colored paper beneath it, highlighting challenges of real-world waste classification, where objects are frequently stacked or partially obscured to make it difficult to predict. Models often struggle with correctly identifying colors in cluttered scenes due to the lack of real-world complexity in their training data. Enhancing training with diverse and realistic samples could help improve their accuracy and robustness in complex environment.

Challenges in Rare Class Recognition: Models often struggle to accurately recognize and classify less frequent categories in cluttered scenes, particularly when objects are deformed. As seen in Q3, models mislocate or miss the metal, highlighting the need for improved training on diverse variety of deformed object shapes in cluttered environment which are often encountered in real world streams.

Weak Classification in Cluttered Environments: The responses in Question 1 highlight key challenges in accurate material identification, particularly in scenes where objects are partially obscured. For example, while some models like GPT-4o correctly identify a range of materials, others like LLaVA and Qwen-VL struggled, with differentiating between visually similar objects, leading to incomplete or incorrect classifications. This inconsistency underscores the need for further refinement of VLLMs to improve their robustness in real-world applications, such as automated waste management, where precise identification is critical. Further insights are given in Appendix A.4.

Potential Data Leakage: This is dataset which is maintained by independent research group and cannot be obtained by using web crawling techniques which VLLMS use to curate their datasets.

Figure 5: Qualitative results illustrating models struggling with identifying shapes, colors, and recognizing rare classes within cluttered scenes, indicating areas for further investigation and improvement.

Model	MM-VET	MV-Bench	SEED-Bench	Waste-Bench
GPT-4	-	-	-	57.5
Gemini	-	-	-	49.4
InstructBLIP	69.9	51.0	61.7	48.6
LLaVA	46.6	53.0	66.7	47.4
Qwen-VL	-	73.0	54.8	41.3
CogVLM	-	-	-	41.6
MiniGPT-4	47.9	29.5	49.2	36.4
Human Upper Bound	-	-	-	81.2

Table 4: Comparison of VLLM recognition performance across different benchmarks in terms of accuracy. Note: In this table, – values indicate results not provided.

6 Validation and Comparison Across Other BenchMarks

The Table 4 compares the accuracy of various VLLMs across various benchmarks. Notably, the table illustrates the diverse challenges posed by each benchmark, with Waste-Bench offering a unique set of difficulties due to its focus on cluttered scenes with deformed objects. The performance of models such as LLaVA, InstructBLIP, and Qwen-VL shows a noticeable drop in accuracy on Waste-Bench compared to SEED-Bench and MV-Bench. This highlights the increased complexity and difficulty in real-world waste classification scenarios and need to optimize current models for the unique challenges of waste classification.

7 Conclusion

In this paper, we evaluated various VLLMs in complex environments with deformed objects, revealing significant weaknesses in the identification of shapes, colors, and locations. We introduced the Waste-Bench benchmark, which features multiple categories to enable a comprehensive validation of these models. The Waste-Bench benchmark provides a robust framework for assessing VLLMs in challenging conditions, aiding in the development of more resilient and accurate models for real-world applications like waste segregation and autonomous waste management.

Limitations Our study, though comprehensive, has some limitations. The scope of our evaluation was limited to a specific set of cluttered environments, which may not fully represent the variety of real-world scenarios. In addition, the models were tested under controlled conditions and their performance in more dynamic and unpredictable settings remains to be explored. We tested models on a variety of questions to ensure robust testing for our evaluation purposes, accuracy and score were calculated and seemed sufficient, showcasing the robustness of our approach. Incorporating additional evaluation methods in future work could provide a more complete understanding. Despite these limitations, our findings offer valuable insight and a strong foundation to advance research in this area.

Ethics Statement We constructed this dataset based on images given in the zwaste-f dataset (Bashkirova et al., 2022). We constructed this data set based on images provided in the Zerowaste-F dataset (Bashkirova et al., 2022). This data set includes various images of waste in cluttered environments to simulate real-world conditions. Some images contain identifiable objects, but we ensured that no personal identification details are included. When used properly, our image and annotation dataset provides significant value for evaluating waste classification models.

References

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Dina Bashkirova, Mohamed Abdelfattah, Ziliang Zhu, James Akl, Fadi Alladkani, Ping Hu, Vitaly Ablavsky, Berk Calli, Sarah Adel Bargal, and Kate Saenko. 2022. Zerowaste dataset: Towards deformable object segmentation in cluttered scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21147–21157.

Shih-Han Chou, Wei-Lun Chao, Wei-Sheng Lai, Min Sun, and Ming-Hsuan Yang. 2020. Visual question answering on 360° images. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1596–1605.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*.

Zhicheng Feng, Jie Yang, Lifang Chen, Zhichao Chen, and Linhong Li. 2022. An intelligent waste-sorting and recycling device based on improved efficientnet. *International Journal of Environmental Research and Public Health*, 19(23):15987.

Alex Havrilla, Andrew Dai, Laura O'Mahony, Koen Oostermeijer, Vera Zisler, Alon Albalak, Fabrizio Milo, Sharath Chandra Raparthy, Kanishk Gandhi, Baber Abbasi, et al. 2024. Surveying the effects of quality, diversity, and complexity in synthetic data from large language models. *arXiv preprint arXiv:2412.02980*.

Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, and Xiang Li. 2023. Rsgpt: A remote sensing vision language model and benchmark. *arXiv preprint arXiv:2307.15266*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Wei-Lung Mao, Wei-Chun Chen, Chien-Tsung Wang, and Yu-Hao Lin. 2021. Recycling waste classification using optimized convolutional neural network. *Resources, Conservation and Recycling*, 164:105132.

Jing Meng, Ping Jiang, Jianmin Wang, and Kai Wang. 2022. A mobilenet-ssd model with fpn for waste detection. *Journal of Electrical Engineering & Technology*, 17(2):1425–1431.

OpenAI. 2024. Hello gpt-4o. Accessed: 2024-05-26.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

J Redmon. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

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

Shaoqing Ren. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*.

Wenqi Shao, Yutao Hu, Peng Gao, Meng Lei, Kaipeng Zhang, Fanqing Meng, Peng Xu, Siyuan Huang, Hongsheng Li, Yu Qiao, et al. 2023. Tiny lvlm-ehub: Early multimodal experiments with bard. *arXiv preprint arXiv:2308.03729*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Zhongyi Xia, Houkui Zhou, Huimin Yu, Haoji Hu, Guangqun Zhang, Junguo Hu, and Tao He. 2024. Yolo-mtg: a lightweight yolo model for multi-target garbage detection. *Signal, Image and Video Processing*, pages 1–16.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhou, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

A Appendix

A.1 Data Filtration

Table 5 presents an overview of the dataset statistics, including the total number of images and question-answer (Q/A) pairs. The dataset initially contains 952 images and 11,424 Q/A pairs. However, approximately 20% of the Q/A pairs (1,904 pairs) were filtered out, leaving a total of 9,520 updated Q/A pairs for further analysis. This filtration process ensures that the data used for evaluation is of higher quality and relevance to the task at hand

Images	Q/A	Filtered	Updated
952	11424	~20% [1904]	9520

Table 5: Dataset Statistics: Overview of Total and Filtered Question-Answer Pairs

A.2 WasteBench Insights

Figure A.1 provides two visualizations related to the answers in the study. On the left, a word cloud is displayed, representing the most common keywords found in the responses. This visualization highlights the frequency and prominence of key terms, offering insights into the main themes and concepts discussed in the answers. On the right, a bar chart shows the distribution of question types, providing an overview of the variety and balance of questions posed during the study. Together, these

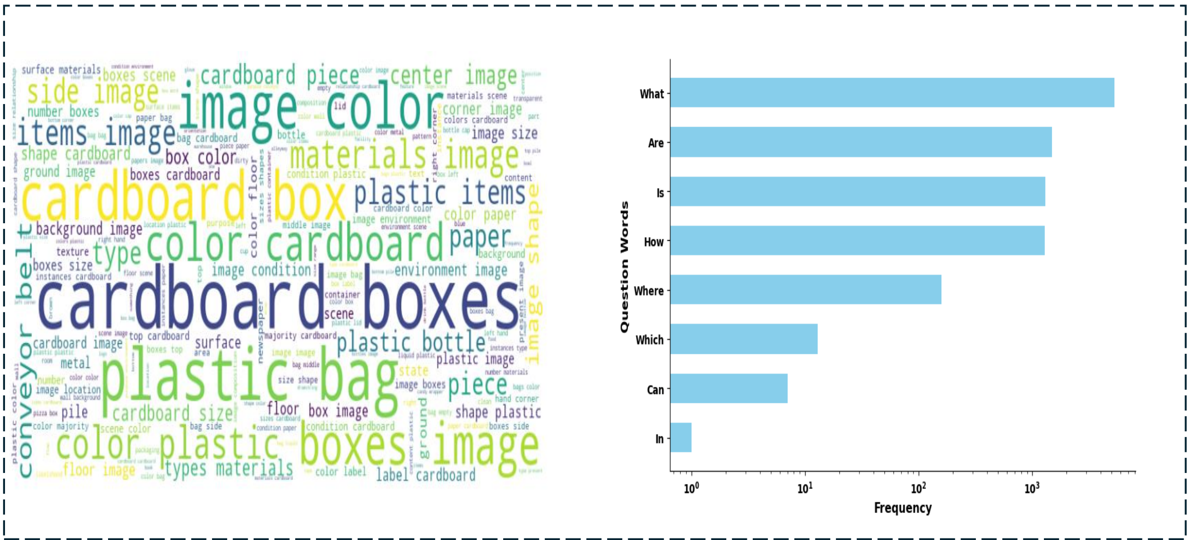


Figure 6: Waste-Bench Overview. Left: Most frequent keywords in the answer set, Right: Frequency distribution of question types.

Figure 7: Q/A generation from Caption

figures help to further understand the characteristics of the responses and the types of questions that were most prevalent in the dataset

A.3 Experimental Settings

As given in Table 6, all models were used in their pre-trained state to ensure a fair comparison across different architectures. Given the diversity of the models employed, specific hyperparameter tuning was not performed for individual models; instead, the focus was on evaluating their inherent capabilities. Each model was assessed under consistent conditions, using a single NVIDIA 24GB GPU to run the experiments, ensuring uniformity in computational resources across the tasks.

A.4 Insights

Recognition and Counting Challenge: Models generally struggle with recognizing and classifying objects across all classes in cluttered environments. As illustrated in Figure 8 , the models face significant challenges when dealing with complex and cluttered environments, as shown by the incorrect answers highlighted in red. However, we included a case where the models performed better, such as accurately identifying the dominant color in the image, with few models providing the correct answer. This contrast highlights that while models can handle simpler tasks, like recognizing a dominant color in scenarios with clear and singular visual cues, they continue to struggle with more complex tasks that require understanding spatial relationships and object classification in cluttered environments. Including this case emphasizes that while there are areas where models show reasonable performance, significant gaps remain in more challenging real-world scenarios

However, the models struggle significantly when dealing with more complex tasks, like identifying the shape and size of objects or differentiating between similar materials in cluttered environments. Despite clear instructions regarding the presence of only one rigid plastic item, the responses varied widely, highlighting ongoing challenges in

Model	Architecture	Context Length	Evaluation Mode
GPT-4o	closed-source	2,048 tokens	zeroshot, pre-trained wts
GeminiPro1.5	closed-source	2,048 tokens	Caption, QA tasks
GeminiPro1.0	Proprietary closed-source	2,048 tokens	zeroshot, pre-trained wts
InstructBLIP	BLIP-2_Vicuna_Instruct (Vicuna-7B)	2,048 tokens	zeroshot, pre-trained wts
LLaVA	LLaVA-1.6 (Vicuna-7B)	2,048 tokens	zeroshot, pre-trained wts
Qwen-VL	Qwen-VL-Chat (Qwen-7B)	2,048 tokens	zeroshot, pre-trained wts
CogVLM	CogVLM-chat-v1.1 (Vicuna-7B)	2,048 tokens	zeroshot, pre-trained wts
MiniGPT-4	MiniGPT-4 (Vicuna-7B)	2,048 tokens	zeroshot, pre-trained wts

Evaluation Process	Details
Evaluation Method	Models were evaluated on Waste-Bench tasks, including classification, counting, color recognition, and other categories. GPT-4 evaluated model predictions.
Human Verification	Two human evaluators verified model predictions, showing high consistency with GPT-4 evaluations.
Error Handling	Default safety mechanisms were employed to prevent out-of-memory errors and ensure stable performance.

Table 6: Experimental Setup and Model Specifications.

spatial reasoning and object recognition. These inconsistencies emphasize that while models can handle basic visual tasks, they falter when faced with more intricate aspects of real-world scenes, such as understanding object relationships or accurately assessing size and material properties

A.5 Challenges with Noise, Enhanced Lighting and Shaded Degradations

:

While not the main focus of our paper, we further extended our evaluation to assess the models' performance across various degradations. Our experiments revealed that introducing noise, shading, and enhanced lighting conditions in the images exacerbates performance issues in the models, as shown in Table 7. For instance, some models experience a significant drop in accuracy when noise is introduced, highlighting their vulnerability, while others exhibit better noise-handling capabilities. These findings underscore the importance of incorporating environmental factors into future model evaluations. To ensure consistency in our experiments, we applied fixed levels of degradation. Specifically, we used a gradient mask for shading with an initial intensity of 0.7, a Gaussian noise with a sigma value of 7, and a brightness factor of 1.2 for enhanced lighting in the HSV color space. Evaluating these natural degradations is crucial for understanding the robustness of models in real-world scenarios, where ideal conditions are seldom guaranteed. By testing models under these challenging conditions, we are able to identify vulnerabilities and areas for improvement, ensuring that models are better equipped to handle diverse and unpredictable environments. This is also important in considering the performance measure of VLLMs in applications other than waste such as surveillance, autonomous driving, and environmental monitoring, where models need to be resilient to a wide range of environmental factors and disruptions.

Model	Normal	Noisy	Enhanced	Shaded
Gpt-4o	57.52	57.04	57.40	56.90
GEMINI	49.45	48.48	48.65	48.20
I.BLIP	48.58	46.29	47.20	46.25
LLaVA	47.45	47.03	46.90	46.16
CogVLM	41.58	40.15	40.50	39.73
Qwen-VL	41.30	39.40	40.58	37.09
MiniGPT4	36.40	36.21	36.90	35.20

Table 7: Evaluation results of various Vision Large Language Models (VLLMs) across different degradation scenarios and accuracy metrics.

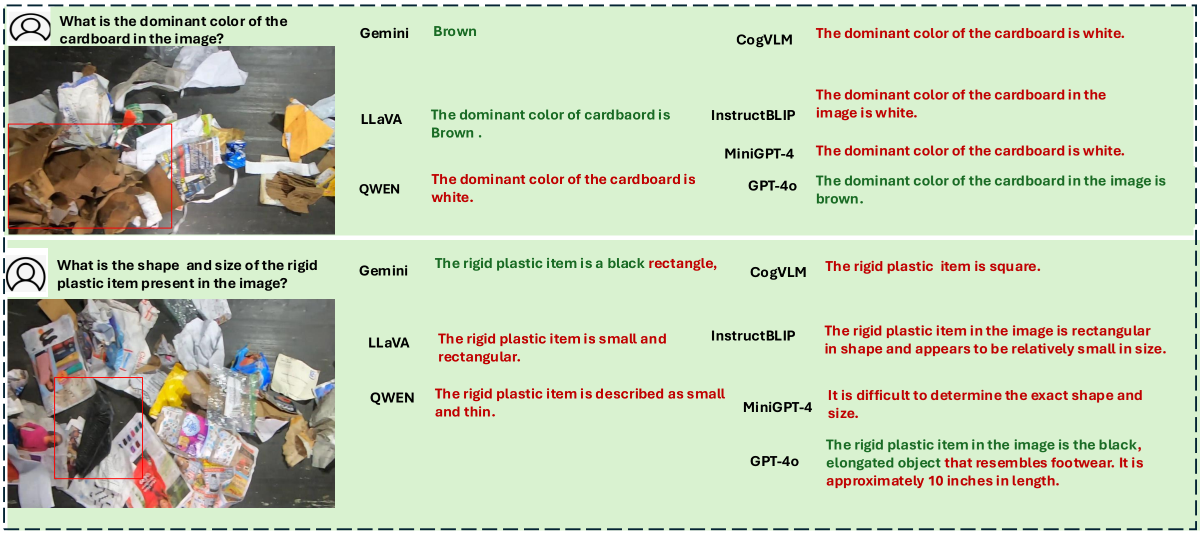


Figure 8: Qualitative results illustrating models struggling with identifying shapes, colors, and recognizing rare classes within cluttered scenes, indicating areas for further investigation and improvement.

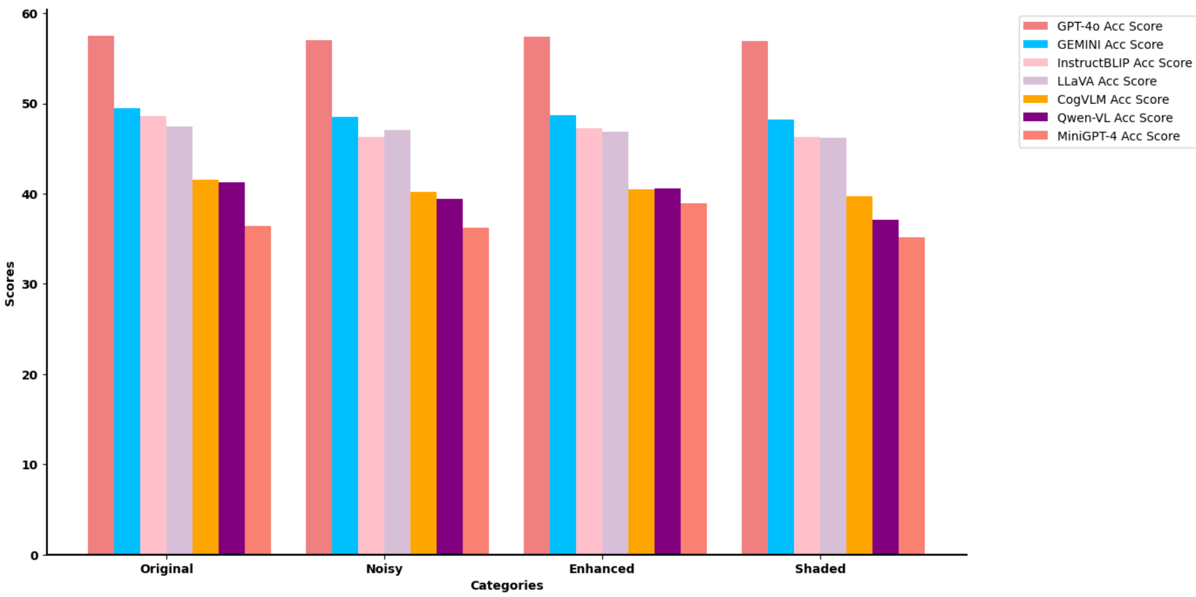


Figure 9: Performance comparison of various Vision Large Language Models (VLLMs) under different degradation scenarios. The chart illustrates how models like GPT-4, GEMINI, InstructBLIP, and others struggle with tasks involving shape recognition, color identification, and classification of rare classes within cluttered scenes, particularly under conditions of noise, enhanced lighting, and shading. This highlights the challenges VLLMs face in maintaining accuracy and robustness when subjected to real-world visual distortions.