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 exceptional abilities to comprehend, reason, and generate text across a wide range of open-ended tasks. Notably, PaLM 2 (Anil et al., 2023) excels in commonsense reasoning, multilingual capabilities, and advanced coding, while Falcon (Penedo et al., 2023) shows excellent performance in multiple Natural Language Processing(NLP) tasks. This success of LLMs is attributed to their superior performance on various tasks (Qin et al., 2023; Devlin et al., 2019).

Building on the advancements of LLMs, Vision-Language Models (VLLMs) have emerged, leveraging aligned image-text data from web imagery and manual annotations to facilitate effective self-supervised vision-language modeling, including caption generation (Vinyals et al., 2015; Chou et al., 2020). This progress is exemplified by models like multimodal GPT-4 (Achiam et al., 2023; Liu et al., 2023) and open-source initiatives such as LLaVA (Liu et al., 2024). These VLLMs, developed through generative pretraining and instruction-tuning, excel in zero-shot task completion across a variety of user-oriented multimodal tasks. Their advanced capabilities are paving the way for the development of versatile multimodal conversational assistants with extensive applications in real-world scenarios (Hu et al., 2023). Vision Large Language Models (VLLMs) (Zhu et al., 2024; Shao et al., 2023; Yu et al., 2023) have demonstrated remarkable capabilities in engaging with visual content, offering a wide range of potential applications. While several benchmarks have been suggested to evaluate these capabilities, there are still challenges and opportunities for further development in this field (Yu et al., 2023; Shi et al., 2023). Notably, domains such as waste classification and segregation for improved recycling, reducing material generation, and minimizing environmental impact present significant opportunities. These advancements can lead to a substantial positive impact on environmental sustainability.

Motivated by the wide-scale applications of Vision Large Language Models (VLLMs) and the lack of comprehensive benchmarking efforts for complex visual environments, especially for waste, we present a new benchmark, Waste-Bench, to thoroughly assess the performance of VLLMs. As shown in Figure 3, Waste-Bench evaluates VLLMs on key aspects of single and multi-class recognition, robustness, and reasoning in visual tasks. It encompasses scenarios that closely mimic real-world conditions, including cluttered waste images with deformed objects. Waste-Bench is an open-ended visual QA and classification benchmark focusing

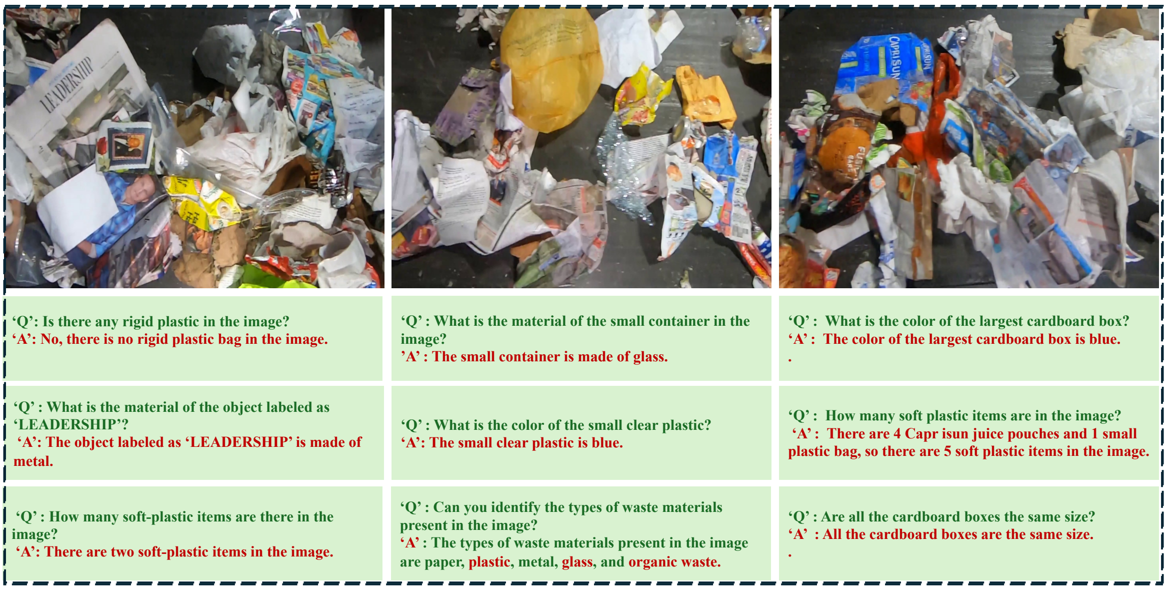

Figure 1: Examples illustrating the challenges faced by models in interpreting cluttered scenes. The model struggles with recognizing shapes, counting objects, comparing material sizes, and identifying deformed and unrecognized objects. The cluttered environment and deformed shapes significantly impact the model's accuracy across different scenarios, as revealed by the specific questions accompanying each image.

on waste recycling. The performance of VLLMs on the Waste-Bench benchmark reveals that these models struggle to accurately comprehend complex visual environments and identify objects, particularly in cluttered scenes and when dealing with deformed shapes, counting tasks, and other challenging aspects as given in Figure 1. Extensive quantitative and qualitative analyses using the Waste-Bench benchmark provide important insights into these VLLMs based on their failure cases and individual performances across diverse visual scenarios. As illustrated in Figure 1, these shortcomings highlight the need for improved robustness and reasoning capabilities in VLLMs to better handle the intricacies of real-world environments. Our main contributions can be highlighted as below:

- We present Waste-Bench, a comprehensive benchmark designed to assess the robustness and reasoning capabilities of Vision Large Language Models (VLLMs) in waste classification, reflecting the complexities of real-world applications.

- We comprehensively evaluate a range of VLLMs, including both open-source and closed-source models. Our evaluation reveals that most models exhibit significant performance challenges, highlighting their limited reasoning capabilities in cluttered scenes with deformed shaped objects.

- We extensively analyze VLLMs on the Waste-Bench benchmark, focusing on scenarios where models struggle, such as identifying deformed shapes, navigating cluttered scenes, and performing counting tasks. Our findings provide insights to enhance future human-centric AI systems' robustness and reasoning for waste classification and management.

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. LLaVA (Liu et al., 2024) and CogVLM (Wang et al., 2023) have shown robust capabilities in integrating vision and language, enabling them to excel in multimodal tasks. MiniGPT-4 and InstructBLIP (Zhu et al., 2024; Dai et al., 2024) further enhance these capabilities by leveraging generative pretraining and instruction-tuning to achieve strong zero-shot task completion. Additionally, Gemini-Pro (Reid et al., 2024) exemplifies

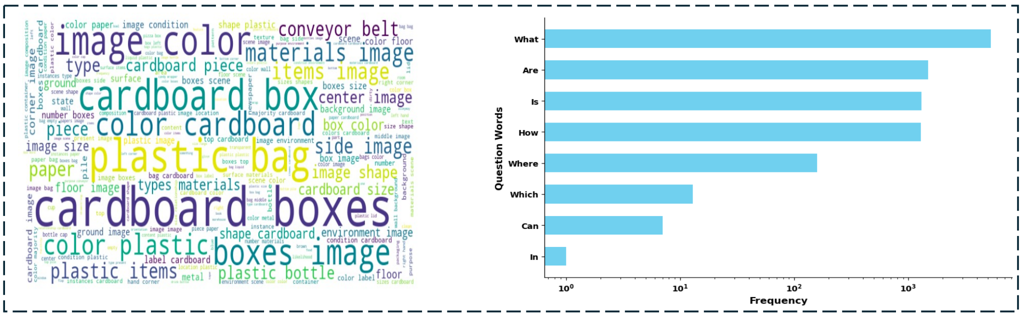

Figure 2: Waste-Bench Overview. Left: Illustration of the most frequent keywords in the answer set of the Waste-Bench benchmark. Right: Frequency distribution of question types.

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 answring and so on. These models accept both visual and textual inputs and generate textual responses. From an architectural perspective, Vision Large Language Models (VLLMs) typically combine pretrained 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., 2023b), which evaluates the visual capabilities of both image and video LMMs across multiple dimensions, and MV-Bench (Li et al., 2023a), 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.

Additionally, LVLM-eHub (Xu et al., 2023) offers an interactive model comparison platform through image-related queries, allowing for a more dynamic evaluation of VLLMs. OwlEval (Zhou et al., 2023) and MM-Vet (Zhang et al., 2024) further underscore comprehensive Vision-Language(VL) skills by introducing evaluation metrics that transcend mere model hierarchies. MME (Chen et al., 2022) also stands out by providing a multi-modal evaluation framework that assesses the integration of vision and language capabilities. These benchmarks contribute to a more holistic understanding of VLLM performance in various complex and realistic scenarios.

In contrast, Waste-Bench is a comprehensive benchmark designed to assess the robustness and reasoning capabilities of VLLMs in waste classification. The Waste-Bench benchmark includes scenarios with cluttered images and deformed objects to simulate real-world conditions. It aims to thoroughly evaluate the performance of VLLMs in challenging visual environments, providing a more rigorous assessment than existing benchmarks.

3 Waste-Bench

In this work, our objective is to develop a comprehensive benchmark to evaluate the robustness and reasoning capabilities of Vision Large Language Models (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.

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

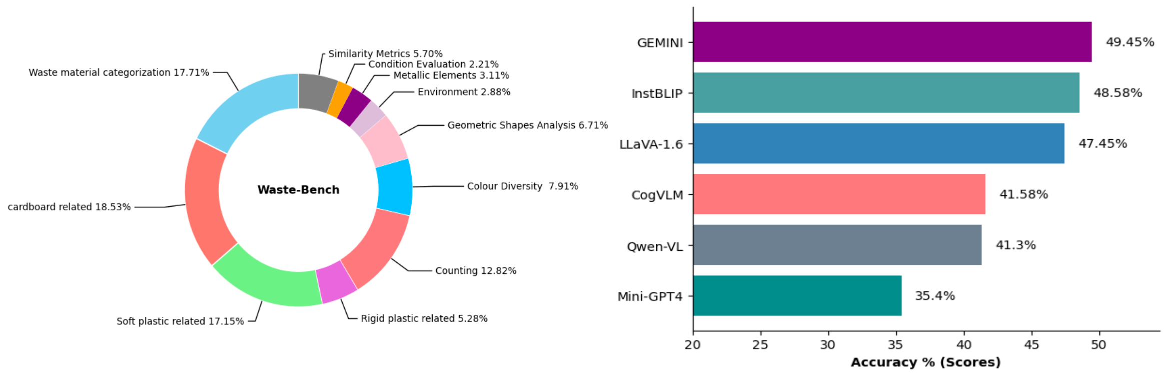

Figure 3: Left: Waste-Bench comprises of 11 diverse complex quesiton categories encompassing a variety of waste images context. Right: Overall performance of VLMMs across the images.

colors. This comprehensive dataset is designed to rigorously test the capabilities of Vision Large Language Models (VLLMs) in handling complex and cluttered visual environments.In Figure 2 (right), we present the distribution of different question types in Waste-Bench, aimed at evaluating model robustness related to classification categories. Figure 2 (left) shows a word cloud of frequent keywords in the answer set, emphasizing objects and attributes relevant to waste classification.

3.1.1 Waste-Bench Different Question Types

To assess the robustness and reasoning capabilities of Vision Large Language Models (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 3.

- 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

After defining the waste dataset question categories, we now proceed to building the Waste-Bench benchmark, which consists of four steps. Each step is presented in detail below, and can be visually explored in Figure 4.

Stage 1: Data Collection and Annotation. We thoroughly reviewed various datasets to find those that represent waste images within cluttered environments. We meticulously pre-processed the metadata provided with the images to ensure accurate representation of the categories assigned to each image. The test dataset contains 952 images. Following the image collection process, we utilized the Gemini-Vision model to generate high-quality captions for these images. These captions were subsequently verified by experienced human annotators. We adhered to stringent annotation and verification instructions to ensure a robust and reliable set of captions. The prompt used for generating captions is provided in Figure 4. Personalized annotation guidelines were used for each image category to ensure accuracy.

Stage 2: Question-Answer Generation The first challenge is to select an evaluation setting to assess VLLMs. Inspired 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. With VLLMs being increasingly integrated into waste management systems, it's crucial to validate their ability to accurately analyze and respond to questions about waste objects in cluttered environments. In evaluating the capabilities of VLLMs, our goal is to determine whether these models can understand the input image not only by analyzing spatial content and recognizing classes but also by comprehending the underlying rationale behind the depicted waste objects and their relationships with the surrounding context. This involves creating questions that go beyond simple image comprehension and require the model to engage in complex logical inference and contextual understanding. Specifically, we create question types that test the model's ability to classify objects based on recognition, color, shape, single class, multiclass, condition, and other relevant aspects in complex, cluttered settings. It was particularly challenging to ensure that the models not only correctly analyzed the images but also responded accurately and appropriately to the questions posed. Example prompts used as instructions to LLMs for curating QA pairs are provided in Figure 4.

Stage 3: QA Pairs Filtration After generating QA pairs, a manual filtration step is employed, with human assistance to verify each generated QA pair. Therefore, an exhaustive filtering process is conducted which involves QA rectification and removing those samples which are not relevant to the image or evaluation type. This process results in a final set of 9552 high-quality QA pairs for the Waste-Bench benchmark.

Stage 4: Evaluation Procedure Previous methods in the literature have explored using LLM models as judges for quantifying results in open-ended QA benchmarks. We adopt a similar approach and instruct LLMs to act as evaluators to assess the correctness of predicted responses from VLLMs compared to ground-truth answers. We generate open-ended predictions from VLLMs by providing image-question pairs as inputs and then present the model predictions and their corresponding ground-truth responses to the LLM Judge alongside the

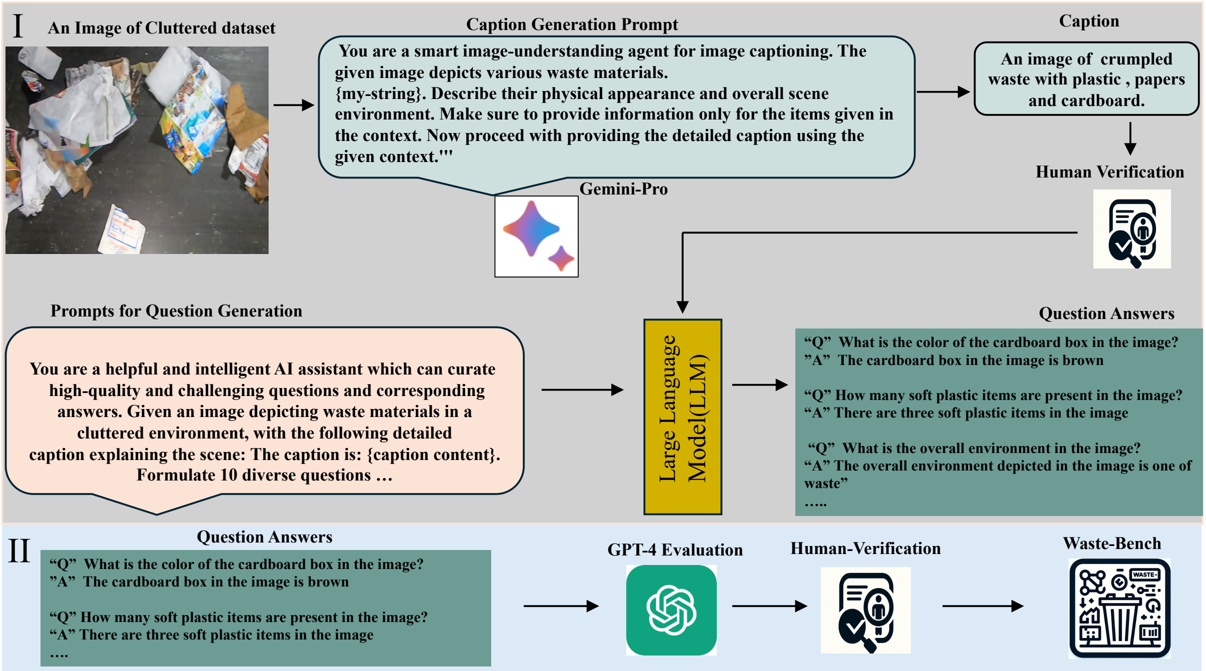

Figure 4: Step I: Gemini-Pro generates detailed captions for images of waste, which are then verified by human annotators. Step II: Nearly 10k diverse questions are generated from these captions, evaluated by GPT-4, and verified by humans.

evaluation prompt. The Judge determines whether the prediction is correct or incorrect through a binary judgment, assigns a score from 1 to 5 representing the quality of the prediction, and provides reasoning to explain its decision. The evaluation prompt used in our case is shown in Figure 5.

```
"role": "system",
"content": "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.\n"
"- Consider predictions with less specific details as correct evaluation, unless such details are explicitly asked in the question.\n"
},
{
"role": "user",
"content": please evaluate the following question-answer pair:\n\n"
Question: {question}\n
Ground truth correct Answer: {ground truth}\n"
Predicted Answer: {predicted}\n\n"
"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 5: The prompt is designed to enable the Language Model to act as an evaluation judge, assessing and scoring the performance of VLLMs. It categorizes responses as accurate or not and assigns a score from 1 to 5 based on the correctness and quality of the prediction. Additionally, it also provides the reasoning.

4 Performance Evaluation on Waste-Bench

Both open-soruce and closed-source models are explored and selected for the evaluation. We evaluate six models in total where among the open-source models, we evaluate five recent VLLMs, including InstructBLIP, LLaVA-1.6, CogVLM, Qwen-VL, and MiniGPT-4. For evaluating closed-source models, we use Gemini-Pro.

4.1 Main Experiments on Waste-Bench

In Table 1, we present the evaluation results of various Vision Large Language Models (VLLMs) across accuracy metrics on the Waste-Bench dataset. We analyse these results and present several key findigs.

Open Source VLLMs Struggle on Waste Bench having cluttered environment : All open-source VLLMs find it challenging to perform well and thus show inferiror performance when evaluated on the Waste-Bench dataset, particularly in cluttered scenes. Additionally cluttered scenes filled with deformed shaped objects make the task more competitive. 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 ac-

Model	Versions	LLM	Accuracy	Score
Gemini	Gemini-1.0 Pro	Proprietary LLM	49.45	3.09
LLaVA	LLaVA-1.6	Vicuna-7B	47.45	3.06
Qwen-VL	Qwen-VL-Chat	Qwen-7B	41.30	2.60
MiniGPT-4	MiniGPT-4	Vicuna-7B	36.40	2.53
CogVLM	Cogvlm-chat-v1.1	Vicuna-7B	41.58	2.81
InstructBLIP	BLIP-2_Vicuna_Instruct	Vicuna-7B	48.58	3.03

Table 1: Evaluation results of various VLLMs across different accuracy metrics. We present results for both open-source and closed-source models, providing a comprehensive assessment of their performance.

curacy of 49.45% with a score of 3.09, however MiniGPT-4 suffers severely with these particularly challenging conditions and thus under perform. Table 1 results show Accuracy of the response and the score of the models where total score is 5.

Closed Source Model Perform Competitively on Waste-Bench:
As shown in Table 1, the Gemini model surpasses the performance of open-source 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 below 50%. GEMINI handles cluttered scenes with deformed shaped objects, better than others, indicating a more sophisticated understanding of complex visual contents. In handling cluttered conditions with mixed and deshpaed objects, Gemini maintains a performance with an accuracy of 49.45% and a score of 3.09.

Comparison Across Models: As evident from Table 1, among the models evaluated, Gemini consistently outperforms others with the highest accuracy of 49.45%. This is followed closely by InstBLIP with an accuracy of 48.58% and 42.29%, respectively. On the other hand, models like MiniGPT-4 and Qwen-VL show lower performance metrics, with MiniGPT-4 having the lowest scores 36.40%.

4.2 Key Highlights and Qualitative Results

Based on the evaluation of Vision Large Language Models (VLLMs) on the Waste-Bench benchmark, several key insights have emerged that provide valuable guidance for future development. This analysis focuses on the models' performance under different conditions, highlighting their strengths and areas needing improvement. Models show weak reasoning capabilities, often failing to accurately identify objects and understand contexts in cluttered environments. For instance, cluttered scenes lead to frequent classifications, such as confusing different types of plastics or failing to recognize partially obscured objects, thus ignoring the presence of the objects in the image. Few samples are shown in the Figure 6 for reference.

Issues in Real-World Waste Classification: Models which shows super performance on organized exhibit less promising results on Waste-Bench especially in counting irregular shaped objects and and many a times predicting the colour wrong because of clutter and presence of one object on top of the other. Figure 6 second row shows the question asked about color of the plastic bag and as transparent plastic is present on top of cardboard it seems as pink so model predicts it as pink. Most models are trained on datasets that lack the complexity of real-world waste scenarios. This training bias results in poor generalization to the diverse conditions of Waste-Bench. Enhanced training strategies, including diverse and realistic samples, are needed to improve robustness.

Recognition and Counting Challenge: Models generally struggle with recognizing and classifying objects across all classes in cluttered environments. They often face significant challenges with soft plastics, which exhibit a wider range of shapes, sizes, and levels of transparency, complicating object enumeration. As illustrated in Figure 6, questions related to the shape and color of soft plastics are frequently answered incorrectly by the models. This discrepancy highlights the difficulties models encounter in accurately identifying and classifying objects in cluttered environments. Additionally, models often struggle with partially occluded objects or objects that are very small, sometimes failing to recognize them entirely. However, models perform slightly better on cardboard due to its distinctive features. Cardboard typically exhibits consistent visual features such as texture, color, and edges, which facilitate easier recognition and classification. These features are less susceptible to

Figure 6: Qualitative results illustrating models struggling with identifying shapes, colors, and recognizing rare classes within cluttered scenes, indicating areas for further investigation and improvement.

the effects of clutter compared to the more varied appearances of soft plastics. Cardboard items are often larger and more easily distinguishable, simplifying the counting process. Thus, while there are overall challenges, the models show a relatively better performance with cardboard due to its distinct and consistent visual features.

Classifying Visually Similar Objects: The models often struggle with accurately predicting similar objects due to the complexity and clutter in the scenes. For instance, in the case of identifying hard plastic, the models frequently confuse it with soft plastic. In cluttered scenes, soft and hard plastics may overlap or be partially obscured, further complicating the classification task. Even small amounts of noise in the images can distort the visual features that the models rely on to differentiate between soft and hard plastics. This added complexity degrades the model's performance and increases the likelihood of misclassification. As illustrated in Figure 1, the image in the center shows an example where the model confused soft plastic and hard plastic, classifying both as plastic. This response is highlighted by the model's answer to Question 3.

Challenges in Rare Class Recognition: Models often struggle with accurately recognizing and classifying less frequent categories within cluttered scenes, especially when these objects are deformed. This difficulty is particularly evident in the case of metals, a class with a small number of instances in the images. As illustrated in fig:evaluation (bottom row), the models frequently miss minor details or fail to identify metals, indicating a need for better handling.

Challenges with Noise and Enhanced Lighting: While not the main focus of our paper, we observed that introducing noise or enhanced lighting conditions in images exacerbates performance issues in some models. For instance, some models suffer a significant drop in accuracy with noise, highlighting their vulnerability, whereas others demonstrate better noise-handling capabilities. These findings suggest the importance of considering environmental factors in future evaluations.

5 Conclusion

In this paper, we evaluated various VLLMs in complex environments with deformed objects, revealing significant weaknesses in identifying shapes, colors, and locations. We introduced the Waste-Bench benchmark, featuring multiple categories to enable 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. Additionally, 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 an even more comprehensive understanding. Despite these limitations, our findings offer valuable insights and a strong foundation for advancing research in this area.

Ethics Statement We constrcuted this dataset based on images given in zwaste-f dataset (Bashkirova et al., 2022). We constructed this dataset based on images provided in the Zerowaste-F dataset (Bashkirova et al., 2022). This dataset 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.

Author Chen et al. 2022. Mme: Multi-modal evaluation for vision-language models. *Conference on Multimodal Benchmarking*, 10:345–356.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

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*.

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*.

Author Li et al. 2023a. Mv-bench: Evaluating spatial and temporal understanding of vllms. *Journal of Multimodal Research*, 56:789–800.

Author Li et al. 2023b. Seed-bench: Evaluating visual capabilities of image and video lmms. *Proceedings* *of the Conference on Vision and Language Integration*, 34:123–134.

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.

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*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384.

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*.

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*.

Zhelun Shi, Zhipin Wang, Hongxing Fan, Zhenfei Yin, Lu Sheng, Yu Qiao, and Jing Shao. 2023. Chef: A comprehensive evaluation framework for standardized assessment of multimodal large language models. *arXiv preprint arXiv:2311.02692*.

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.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.

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*.

Author Zhang et al. 2024. Mm-vet: A benchmark for evaluating vision-language skills. *Journal of Multimodal Research*, 23:789–800.

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*.

Author Zhou et al. 2023. Owleval: A comprehensive benchmark for vision-language models. *International Conference on Vision-Language Models*, 12:567–578.

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*.