Proceedings of the 11th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'25)
Paris, France - August, 2025
Paper No. MVML 113
DOI: 10.11159/mvml25.113

A Comparative Evaluation of Vision Language Models for Waste Classification in Few-Shot Settings

Jonas Funk[1], Paul Bäcker[1], Lukas Roming[1], Jerardh Josekutty[1], Georg Maier[1], Thomas Längle[1]
[1]Fraunhofer IOSB, Karlsruhe, Germany, Institute of Optronics, System Technologies and Image Exploitation
Fraunhoferstr. 1, 76131 Karlsruhe, Germany
jonas.funk@iosb.fraunhofer.de; paul.baecker@iosb.fraunhofer.de

Abstract – Efficient waste classification is essential for sustainable waste management systems. Accurate sorting can significantly enhance recycling efforts and reduce pollution. However, traditional computer vision methods often require large, annotated datasets and extensive retraining, limiting their adaptability to varying waste types and challenging real-world conditions. In this study, we evaluate the potential of Multimodal Large Language Models (MLLMs) and Vision-Language Models (VLMs) for adaptive waste classification, focusing on zero-shot and few-shot learning scenarios. Using datasets such as TrashNet and our custom MultiWaste dataset, we test a method using a CLIP VLM for feature extraction and a simple Nearest Neighbour (VLM-NN) approach for classification. This showcases robust few-shot capabilities and excellent scalability, achieving an accuracy of 97.74% on TrashNet. While MLLMs exhibit strong zero-shot capabilities, their utility diminishes with increasing labelled samples due to high computational costs. In contrast, VLM-NN offers efficient performance but struggles with extremely limited training data. Our results show the potential of Large Pretrained Models for the task of waste classification while providing guidance on which model architectures to consider for different amounts of training data.

***Keywords*:** Waste Classification, Multimodal Large Language Models, Vision Language Models, Zero-Shot Learning, Few-Shot Learning, CLIP

1. Introduction

Waste management plays a crucial role in reducing pollution and recovering resources. Effective use of waste, however, requires separation of different waste types, such as organics, different plastics, metals, and glass. Most countries have developed waste management systems that include pre-sorting by the consumer as well as several sorting steps in industrial sorting plants. The optimal process depends on the specific capabilities of the waste management plant, fluctuations in market prices for output products and changes in waste volumes, e.g., due to seasonal differences. However, to reduce complexity on the consumer end for waste separation, there are usually few waste classes, that are uniform across larger areas, constant in time and easy to understand. Therefore, a lot of potential lies in smart waste-bins, that use computer vision methods to guide the consumer in how to separate and pre-sort their waste into a set of generically labelled bins [1].

There are existing models for waste classification based on RGB image data, mostly based on Convolutional Neural Network (CNN) or Transformer architecture. However, these methods are usually hard to adapt to different class distributions or changing input streams, requiring costly retraining and novel data, for the proposed use case.

In this work, we examine different approaches for highly variable multi-label classification models, focusing on zero- and few-shot methods. We show that Vision Language Models can be effectively used for adaptive waste classification in few-shot and zero-shot settings. Specifically, we demonstrate the utility of CLIP-based feature extraction and Multimodal Large Language Models such as GPT-4o and LLaVA-OneVision, highlighting their classification capabilities. We identify the optimal conditions for each approach and provide insights into the trade-offs between performance, scalability, and adaptability.

2. Related Work

Extensive research has focused on classical methods and CNNs for waste classification [2], [3]. While these methods achieve high performance, such as a benchmark accuracy of 98% on TrashNet using an extended dataset with CNN architectures [4], they often require large volumes of training data, limiting their adaptability. More recent approaches aim to address this limitation by leveraging models with fewer trainable parameters. For example, [5] combine a Swin

Transformer with an autoencoder trained on TrashNet, achieving high accuracy at low parameter count. Similarly, [6] achieve good results by fine-tuning a Vision Transformer on TrashNet, demonstrating low power consumption of their approach, making it feasible for use in smart waste bins. However, all these methods require large amounts of labelled data, which results in lacking flexibility.

Few-Shot learning provides a promising solution by enabling models to generalize from small, labelled datasets. Techniques like the Self-Optimal Feature Transform [7] and Region Comparison Network [8] optimize feature representations for efficient classification and enhance model interpretability, allowing for few-shot approaches in the feature space.

Vision Language Models build on these advancements by seamlessly integrating textual and visual data. SgVA-CLIP [9], which fine-tunes pre-trained CLIP features using an adaptation layer, exemplifies the powerful combination of vision-language capabilities and few-shot learning. Leveraging extensive pre-training on diverse datasets, VLMs demonstrate remarkable versatility. For instance, Recycle-BERT [10] highlights the applicability of language models in waste management by extracting critical insights from literature on plastic recycling. Techniques such as Chain-of-Thought (CoT) prompting, as shown by Wei et al. [11], significantly enhance reasoning, enabling VLMs to address complex classification challenges effectively. Collectively, these innovations underscore the transformative potential of VLMs in overcoming the limitations of traditional trash classification methods.

While Multimodal Large Language Models and Vision Language Models have been applied across various fields, the most closely related work to ours is a paper that employs a CLIP architecture for the classification of construction waste [12]. To the best of our knowledge, we are the first to apply these technologies in the domain of household waste. Furthermore, our paper provides a comprehensive comparison of language-based models, emphasizing their key strengths in zero-shot and few-shot scenarios, as well as their scalability in contrast to more commonly used benchmark models.

3. Methods

This section outlines our approach to waste classification, beginning with the three datasets utilized and the challenges they present. We then explain how our models are structured to address these challenges through various architectures and prompting techniques.

3.1. Dataset

To evaluate our approaches, we selected three different datasets from the waste domain. The reason for the selection of the datasets is the following. We want to stay comparable, for that we use the TrashNet [13] dataset which has often been used in the literature on waste classification to assess the performance of a diverse range of models with varying architectures, making it a valuable benchmarking tool for new approaches.

The TrashNet dataset has a total of 2,527 images of class cardboard, glass, metal, paper, plastic, and trash, representative of household waste. To ensure comparability to other research done on TrashNet we split the dataset into training, validation, and test with a ratio of 70:13:17. One challenge in the dataset is the “trash” class, that refers to the U.S. regulation for trash and includes napkins, styrofoam, chip bags and candy wrappers among others. This is in contrast to the class separation of “residual waste” which is common in Europe.

To prove the capabilities of our approaches further in the waste domain, we utilize the RealWaste [14] dataset. This dataset contains 4808 samples spanning 6 classes taken from real landfill, differing strongly from the pristine images in TrashNet. Here again we use a 70:13:17, train, validation, and test split ratio.

To further test our approaches in multilabel, real-life settings we created our own, unpublished MultiWaste dataset, which includes 245 images of single waste objects. Waste objects were collected and provided by Lobbe RSW GmbH, a waste disposal company. A line-scan camera acquired the samples on a conveyor belt in our lab. 95, 50, and 100 images were used for training, validation, and test respectively. Each image was given one or more labels of plastic, metal, and paper/cardboard. A few example images of our dataset are shown in Fig. 1.

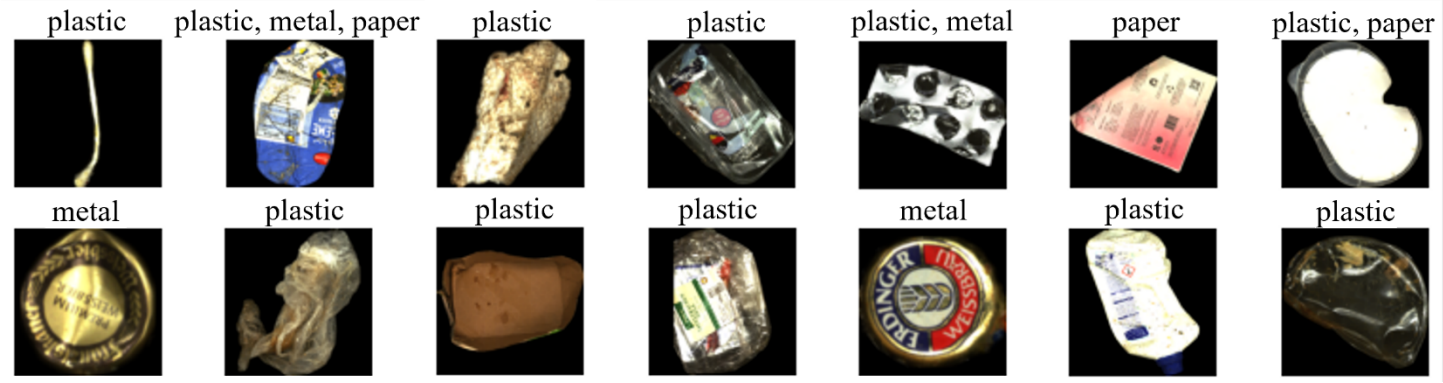

Fig. 1: Example images from our MultiWaste dataset.

3.2. Used Models

In this work, we use the term *Multimodal Large Language Models* (MLLM) to describe models that have been trained mostly on text data but also feature image processing capabilities, namely GPT-4o [15] and LLaVA-OneVision [16]. Both MLLMs generate their output using next-token prediction based on the Transformer architecture.

We distinctively use the term *Vision Language Models* (VLM) when referring to models that have been trained mainly on image understanding tasks (for instance, CLIP [17]).

Our first MLLM model is the LLaVA-OneVision 7B model, specifically the lmms-lab/llava-onevision-qwen2-7b-ov for generating outputs tailored to image captioning and understanding. This model is a compact and efficient multimodal language model designed to process and interpret visual inputs, making it well-suited for resource-constrained environments. For image processing, it combines a modified CLIP framework, SigLIP [18], as the image encoder with a 2-layer MLP projection layer to embed images into its token space. It's trained on multi-image prompts, which makes it suitable for our task.

As our second tested MLLM, we used the GPT-4o-2024-08-06 model via OpenAI's API. This model is a state-of-the-art large multimodal language model optimized for processing and interpreting complex visual and textual inputs.

As our primary VLM, we use the EVA-CLIP 18B model as a pure feature extractor [19].

3.3. Compared Classification Architectures

In our MLLM architecture, shown in Fig. 2 on the right, we perform classification in two steps. First, we combine a CoT prompt that explains the classification task with the image to be classified. The model returns a response, that includes CoT-style "reasoning" along with a prediction of the output label. In the second step, we use the CoT response obtained from the first step alongside a labelling prompt to extract an isolated class label. Both steps can be executed using either GPT-4o or LLaVA-OneVision. In the first step, we may employ either a standard CoT prompt or an enhanced CoT prompt that includes either images or textual descriptions of the images. The descriptions are generated by inputting training images along with a creation prompt, as detailed in the supplements, into the LLaVA-OneVision model. We ensured that the generated descriptions comply with the maximum sequence length of the CLIP model used in the VLM approach, which is 77 tokens.

In our VLM architecture, we classify test images by embedding them using the EVA-CLIP-18B model and assigning the label of the Nearest Neighbour (NN) based on cosine similarity. For the VLM + Image method, the test split is compared against the image embeddings of the train split. For the VLM + Description method, we transform the images in the training set to short image description text using the LLaVA-OneVision model. We then embed these textual outputs into the same feature space, using the text embedding module of the CLIP model.

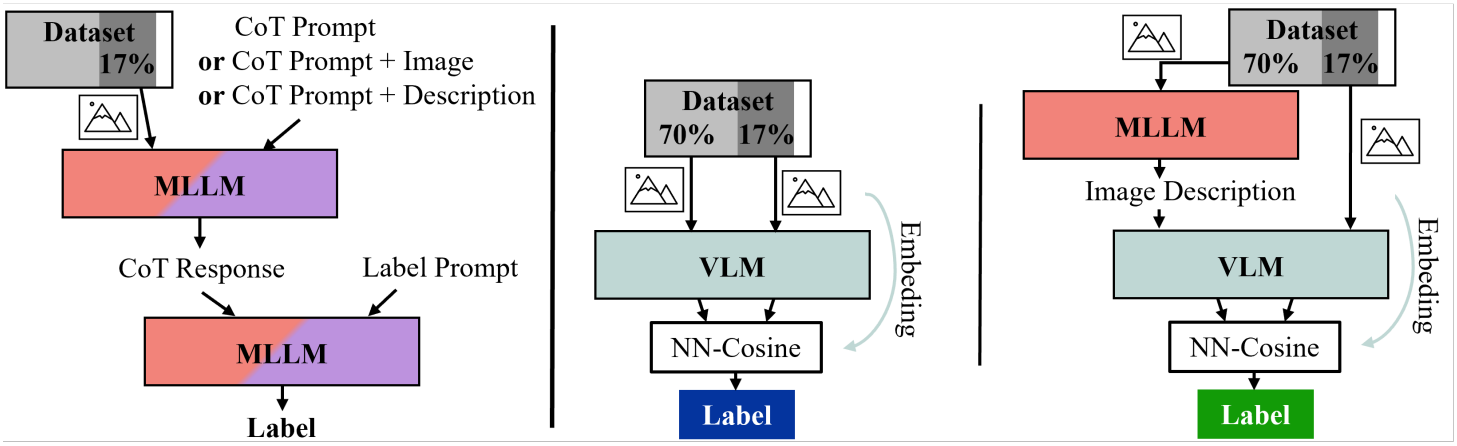

Fig. 2: Classification Architectures. Left: MLLM. Middle: VLM + Image. Right: VLM + Description.

The different embeddings of our two VLM approaches, shown in Fig. 3, provide insights into NN classification. Clustering of image and description embeddings belonging to the same class can be observed in both cases. However, the separation between classes and class purity is more prominent when using images directly. This, along with the sub-clustering especially prominent in the metal, plastic, and trash classes, can be traced back to the high information content when using images rather than solely descriptions. In contrast, when using description embeddings, the problem of the trash class becomes clearly visible as it merges with other classes.

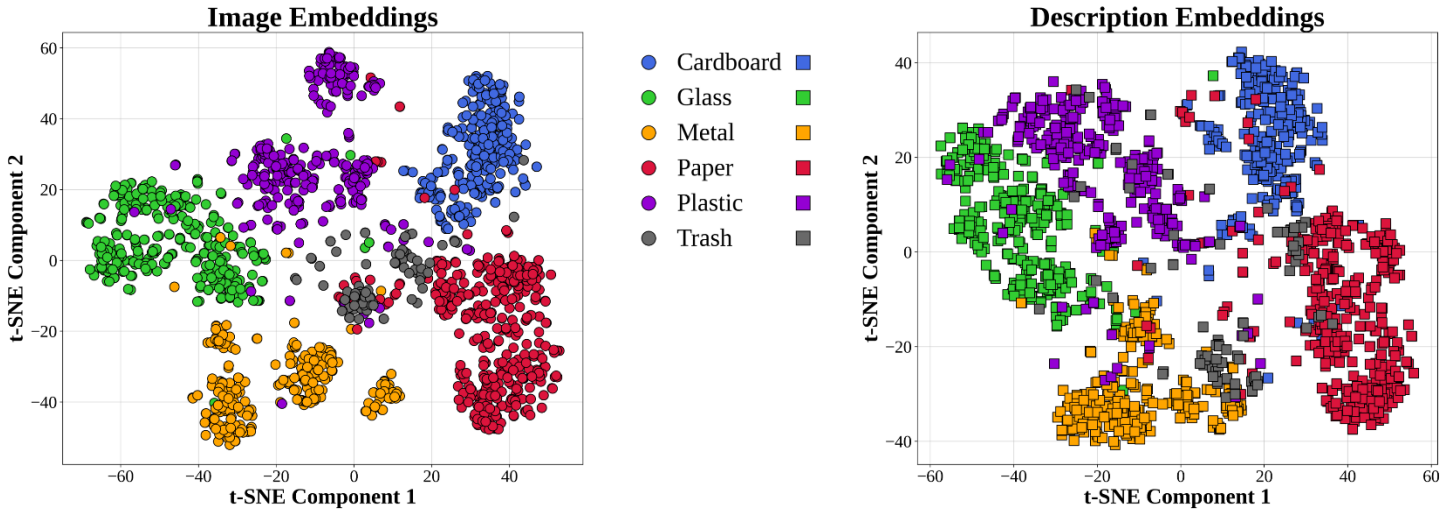

Fig. 3: Visualization of image/description EVA-CLIP-18B embeddings of a single TrashNet train split done with t-SNE.

3.4. Prompts

We used a CoT prompting approach, doing the recognition and reasoning process stepwise, similar to [20]. We noticed that the results are sensitive to small changes in the prompt. One observation was that GPT-4o improved when including a more detailed description of the "trash" class, whereas the performance of the LLaVA-OneVision model was reduced. A possible reason for that is difficulties in the smaller model handling large contexts (for instance due to quadratic scaling).

For that reason, we used the plain CoT prompt for the LLaVA-OneVision model and the CoT prompt with the trash clue for GPT-4o. We have included the detailed prompts in the supplementary material.

4. Results

We conducted two experiments to evaluate the effectiveness of the different approaches. The first experiment focused on assessing how the models perform with varying numbers of training samples, providing insights into their adaptability in limited data scenarios. The second experiment extended the evaluation to other datasets, including our own MultiWaste dataset, to validate our findings and explore the models' performance in more diverse and challenging real-world settings.

4.1 Sweeps on TrashNet

In this section, we evaluate our different methods by sweeping across the number of training samples for images and descriptions. This includes the models presented in Section 3.3. As a baseline, we use DINOv2 [21] (ViT-S/14 distilled and ViT-L/14 distilled) as a feature extractor with NN and EfficientNet-B0 [22] as a trained CNN for classification.

For the MLLMs we used a custom procedure for determining the samples to use as examples. We first perform a zero-shot classification on the training dataset. Then, for testing n-shot classification, we randomly select n images or descriptions of the images that have been falsely classified by the zero-shot model in the training set. This is analogous to a real testing scenario, where one would start off with a zero-shot model and then test on some test samples. If something was falsely classified, it is provided to the model with the appropriate label. However, it must be noted that for n-shot classification in this setting, a much larger number of labelled samples is required, since only the wrong-classified samples in the training set are used. For instance, a n=10 shot classification at a zero-shot accuracy of 90% would require 100 labelled samples.
For each model, we run a set of different input images or descriptions. We repeat the random selection of the input multiple times (5 times for the MLLMs and 100 times for the VLM-NN). The results are shown in Fig. 4 with the curve hull indicating the standard deviation of the results across a number of different randomly picked samples from the training set. Fig. 4 shows the result for our first train/test split.

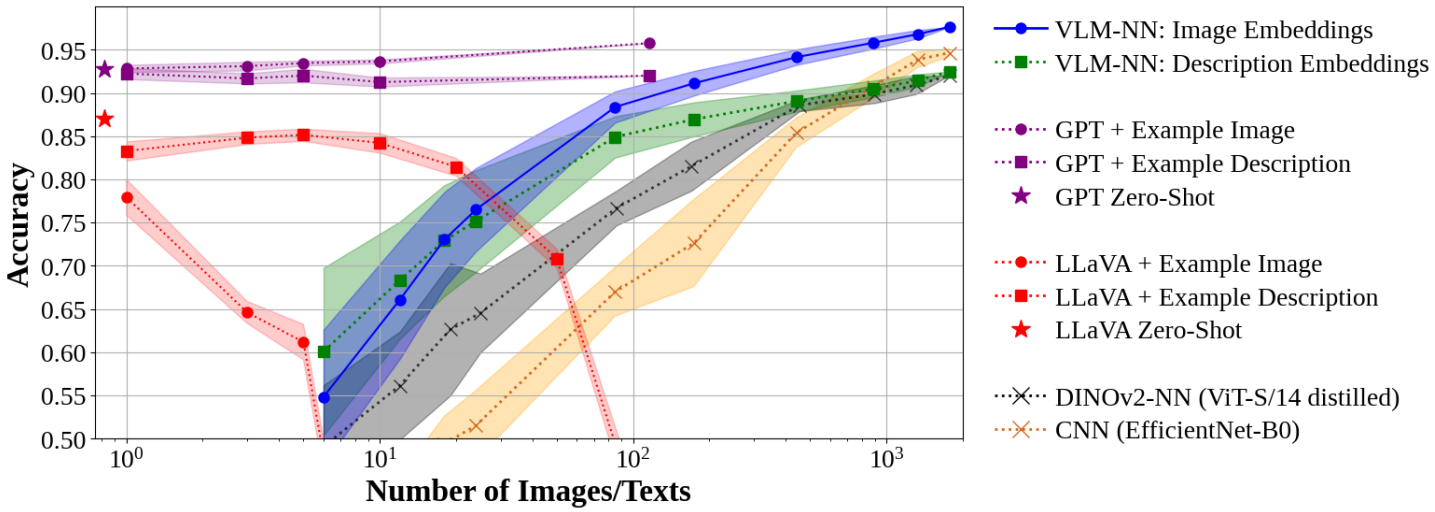

Fig. 4: Sweeping curve of our compared methods, on single TrashNet split.

MLLMs perform well in the zero-shot approach. When adding additional samples, the small MLLM deteriorates for both included images or descriptions, whereas GPT-4o is able to slightly improve as number of included images increases. A reason for this could be issues with long contexts. Larger number of samples have a significant cost penalty, though as the self-attention mechanism is inherently quadratic. Although methods like Flash-Attention and Caching can mitigate this quadratic scaling, the cost of running inference on a single input image at n=116 came in at around $0.08 when using the OpenAI API at the time of writing.

For larger training sets, utilizing a VLM-NN led to good results, comparable with the current state-of-the-art. This also has the benefit of fast inference, as only a single encoding and a NN search must be performed. Transforming the train images

into text descriptions before embedding them improves on average the performance when dealing with a low number of samples. For larger training datasets, however, the performance decreases. This is likely due to higher generalization of the image description.

We have further conducted a 5-fold cross validation study on different train/test splits, that confirm the results, shown in Table 1.

Table 1: Results of classification experiments.

Methods	TrashNet	RealWaste	MultiWaste
	accuracy	accuracy	accuracy
Vision Transformer [6]	0.9698	-	-
DenseNet121 [2],[13]	0.9500	0.8919	-
DINOv2[1,2] ViT-S/14 distilled	0.9160 ± 0.0092	-	0.5400
DINOv2[1,2] ViT-L/14 distilled	0.9583 ± 0.0106	-	0.5700
EfficientNet-B0[1,2]	0.9428 ± 0.0089	-	-
LLaVA-OneVision Zero-Shot[1]	0.8925 ± 0.0136	0.7659 ± 0.0121	0.5600
GPT-4o Zero-Shot[1]	0.9277 ± 0.0054	0.8390 ± 0.0054	0.7400
EVA-CLIP-8B[1,2]	0.9736 ± 0.0110	0.9525 ± 0.0039	0.7200
EVA-CLIP-18B[1,2]	0.9774 ± 0.0070	0.9580 ± 0.0016	0.7500

1: For TrashNet and RealWaste: Our accuracy was calculated as mean of 5 different splits. (Always 70/17/13).
2: All images used from the train split.

4.2 RealWaste and MultiWaste

Our findings are also confirmed on the RealWaste dataset, as shown in Table 1. While the zero-shot MLLM approach results in a slightly lower accuracy, stemming from the more challenging 9-way classification task on images of degraded waste. The VLM-NN approach not only holds up but outperforms the published accuracy of Single et al. [13].

The MultiWaste dataset is the hardest to classify. Not only does the possibility of multiple labels make the classification more difficult but also the images are way harder to interpret, stemming from a real waste sorting application and not manually taken. While the smaller MLLM shows a significant decrease in accuracy on this challenging task, the zero-shot approach using GPT-4o slightly outperforms EVA-CLIP-8B and significantly outperforms DINOv2. This highlights the strength of the method in settings with limited training data.

5. Conclusion

Our findings demonstrate significant potential for leveraging both Multimodal Large Language Models (MLLMs) and Vision-Language Models (VLMs) in waste classification tasks, particularly under constrained training scenarios. MLLMs excel in zero-shot classification. While larger MLLMs demonstrate strong few-shot capabilities when including additional images, the performance of smaller models tends to decline when additional examples are provided, likely due to limitations in context window length and increased computational costs. This cost-effectiveness concern becomes even more pronounced in scenarios requiring longer context windows, making MLLMs less practical for applications needing extensive fine-tuning or large-scale inference.

VLMs employing Nearest Neighbour (NN) matching in embedding space demonstrate strong performance, particularly in few-shot settings, with low computational requirements. Notably, our VLM-NN approach showed efficiency and scalability, making it well-suited for real-world applications. However, its performance declines when the number of training samples is very low, highlighting a drawback in scenarios with minimal labelled data availability.

Our findings are validated on RealWaste, with our VLM-NN method being 6 percentage points higher than published state-of-the-art on this dataset. The lower overall performance on MultiWaste can be attributed to the much more challenging task, as the MLLM outperforms the DINOv2 benchmark.

While the potential exposure of MLLMs and VLMs to datasets like TrashNet during pretraining could raise questions about generalization, we estimate the effect to be negligible compared to the billions of images included in their training data. Applying the models to our own MultiWaste dataset also addresses this concern.

Future research should explore embedding space optimization by employing automatically generated example descriptions, such as those derived from class labels, to reduce reliance on extensive labelled datasets. Investigating more advanced methods, such as SgVA-CLIP, instead of simple NN-based approaches, can further improve accuracy, especially in the range between 10 and 100 training samples. Additionally, expanding benchmarks to include datasets with greater class diversity and real-world noise will provide further insights. Testing models on unconventional class distributions, such as mixing arbitrary classes, could reveal their adaptability and generalization capabilities.

By addressing these avenues, future studies can further enhance the applicability of vision-language technologies in adaptive and efficient waste classification systems, paving the way for smarter, more sustainable waste management solutions.

Acknowledgements

We thank Dr. Almuth Müller for her valuable feedback and insights.

You can find our supplementary material here:
https://drive.google.com/drive/u/1/folders/1cTRwXnQ5EZLvv2MaqarIPHTKV9N9DLgF

References

[1] S. Chowdhury, M. A. N. Bary, A. Abrar, A. Islam, A. Islam, A.M. Nakib and J.H. Emon, "Sustainable Waste Management Using Deep Learning and Smart Bins," *Br. J. Environ. Sci.*, vol. 12, no. 6, 2024.

[2] W. Lu and J. Chen, "Computer vision for solid waste sorting: A critical review of academic research," in *Waste Management*, vol. 142, pp. 29–43, doi:10.1016/j.wasman.2022.02.009, 2022.

[3] R. A. Aral, Ş. R. Keskin, M. Kaya and M. Hacıömeroğlu, "Classification of TrashNet Dataset Based on Deep Learning Models," in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 2058–2062, doi:10.1109/BigData.2018.8622212, 2018.

[4] A. Masand, S. Chauhan, M. Jangid, R. Kumar and S. Roy, "ScrapNet: An Efficient Approach to Trash Classification," in *IEEE Access*, vol. 9, pp. 130947-130958, doi:10.1109/ACCESS.2021.3111230, 2021.

[5] N. Islam, H. Noor and M. Ahmed, "Enhancing Garbage Classification with Swin Transformer and Attention-Based Autoencoder: An Efficient Approach for Waste Management," in *Proceedings of International Conference on Information Technology and Applications*, Singapore: Springer, vol 839, pp. 423-433, doi:10.1007/978-981-99-8324-7_36, 2024.

[6] K. Huang, H. Lei, Z. Jiao and Z. Zhong, "Recycling Waste Classification Using Vision Transformer on Portable Device," in *Sustainability*, vol. 13, no. 21, doi:10.3390/su132111572, 2021.

[7] D. Shalam and S. Korman, "The Self-Optimal-Transport Feature Transform," arXiv:2204.03065, 2022.

[8] Z. Xue, L. Duan, W. Li, L. Chen and J. Luo, "Region Comparison Network for Interpretable Few-shot Image Classification," arXiv:2009.03558, 2020.

[9] F. Peng, X. Yang, L. Xiao, Y. Wang and C. Xu, "SgVA-CLIP: Semantic-guided Visual Adapting of Vision-Language Models for Few-shot Image Classification," arXiv:2211.16191, 2023.

[10] A. Kumar, B. R. Bakshi, M. Ramteke and H. Kodamana, "Recycle-BERT: Extracting Knowledge about Plastic Waste Recycling by Natural Language Processing," in *ACS Sustainable Chemistry & Engineering*, vol. 11, no. 32, pp. 12123–12134, doi:10.1021/acssuschemeng.3c03162, 2023.

[11] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," arXiv:2201.11903, 2023.

[12] Y. Sun, Z. Gu and S. B. Yang, “Probing vision and language models for construction waste material recognition”, in *Automation in Construction*, vol. 166, doi:10.1016/j.autcon.2024.105629, 2024.
[13] G. Thung and M. Yang, “Classification of Trash for Recyclability Status,” Stanford University, 2016.
[14] S. Single, S. Iranmanesh and R. Raad. “RealWaste: A Novel Real-Life Data Set for Landfill Waste Classification Using Deep Learning,” in *Information*, vol. *14*(12), 633, doi:10.3390/info14120633, 2023.
[15] OpenAI, (2023). GPT-4o System Card [Online]. Available: https://cdn.openai.com/gpt-4o-system-card.pdf.
[16] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu and C. Li, “LLaVA-OneVision: Easy Visual Task Transfer,” arXiv:2408.03326, 2024.
[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” arXiv:2103.00020, 2021.
[18] X. Zhai, B. Mustafa, A. Kolesnikov and Lucas Beyer, “Sigmoid loss for language image pre-training,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975-11986, 2023.
[19] Q. Sun, J. Wang, Q. Yu, Y. Cui, F. Zhang, X. Zhang and X. Wang, “EVA-CLIP-18B: Scaling CLIP to 18 Billion Parameters,” arXiv:2402.04252, 2023.
[20] G. Zheng, B. Yang, J. Tang, H. Y. Zhou and S. Yang, “DDCoT: Duty-Distinct Chain-of-Thought Prompting for Multimodal Reasoning in Language Models,” arXiv:2310.16436, 2023.
[21] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin and P. Bojanowski, “DINOv2: Learning Robust Visual Features without Supervision,” arXiv:2304.07193, 2024.
[22] M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in *Proceedings of the 36th International Conference on Machine Learning, PMLR*, vol. 97, pp. 6105-6114, 2019.