*Journal of Harbin Engineering University*
*ISSN: 1006-7043*

Vol 47 No. 05
May 2026

# EcoVision: Multimodal Waste Material Recognition using Vision Language Models (VLMs)

**Manoah Edwin Paul[1], B. Amutha[2]**

[1] Department of Computer Intelligence, SRM Institute of Science and Technology, Kattankulathu, Tamil Nadu, India
[2] CTECH, SRM Institute of Science and Technology, Kattankulathu, Tamil Nadu, India

**Abstract**

Effective waste segregation is important for increasing recycling efficiency and reducing environmental impact. Because conventional management systems depend heavily on mechanical techniques that categorize waste by size rather than material composition, they often give poor recovery rates for recyclable resources. To resolve these issues, we present Ecovision, a multimodal waste recognition framework combining open-vocabulary object detection with vision language classification.

The system combines Grounding DINO for object detection with a CLIP based model for material classification. We achieve zero-shot classification via prompt based inference, enabling the system to identify material categories without task specific training. To improve performance in complex, real world conditions, we apply Low-Rank Adaptation (LoRA) to the visual encoder using a small dataset of cluttered field images. This architecture is specifically designed to combine with existing mechanical segregation pipelines at landfill and recycling facilities.

We tested the model using the TrashNet and TACO datasets and real-world images. By giving few fewshot adaptation improves classification accuracy while maintaining zero shot generalization. Ultimately, EcoVision is a system that can easily grow and be used to identify wastes automatically, making recycling and waste management smarter.

**Keywords**: CLIP, Grounding DINO, Open vocabulary object detection, Vision language models, Waste material recognition, Waste segregation, Zero shot learning.

## 1. Introduction

Due to rapid development, the consumption of goods by people is increasing at a very high rate, this consumption rate is causing a quick increase in the amount of waste produced globally. India is the most populated country in the world and we produce approximately 62 million tonnes of waste every year. This is lot of waste for it to be properly processed manually thus it affects landfill structures and recycling systems.

The current system in India segregates wastes using mechanical techniques such as trommel cylinders and chemicals to remove the moisture and then pass the wastes through screens. Since these systems segregate wastes based on size, it is not that efficient. Recycling wastes based on the composition of the material would produce a much better result in terms of recycling efficiency. Deep Learning model solutions though accurate often require huge amounts of data to train initially and a lot of time is required to retrain if new labels have to be introduced, this is not practical in real world scenarios. Infrastructures in developing cities have grown to use IoT based systems for real time monitoring and resource optimization and thus it is also observed that use of smart systems and digitilization of waste collection improves efficiency [12].

The key contributions of this work are:

- EcoVision a pipeline that is designed as a system that combines Grounding DINO for detecting objects and CLIP for classifying them.
- A zero shot classification approach using material based prompts instead of training data.
- Improving the model using Low-Rank Adaptation (LoRA) which makes small, efficient changes with only a few real-world images.
- The model was evaluated across clean, semi structured and complex real world waste scenarios.

## 2. Related Work

### CNN Based Waste Classification

Previous researches on waste classification is dependent on Convolutional Neural Networks. TrashNet [10], is an widely used dataset in this field, it categorizes waste into classes like plastic, glass, metal, paper, and cardboard and models trained on such datasets, using models like ResNet, VGG, and MobileNet, demonstrated good classification accuracy under clean conditions [4], [11]. However, these methods need a lot of labelled data, which makes them tough to adapt to real world pipelines. CNN models often learn too much from clean data and then struggle with real world conditions like damaged items, blocked views mixed waste, or changing lighting. Most importantly, supervised CNNs aren't flexible in adding a new type of material means collecting new data and retraining the whole model. This increases computational cost and makes it difficult to use in real-world waste processing environments, where new types of waste frequently appear.

**Real World Waste Datasets**

Rather than using datasets with controlled environments, datasets like TACO (Trash Annotations in Context) [9] were introduced to get a natural real world data. TACO provides labelled images of wastes in dirty and cluttered places like streets, parks and beaches with messy backgrounds and objects that overlap and are hidden. Although training models on this dataset helps researchers evaluate supervised models, they are often expensive and time consuming to manually annotate.

Even if that were not a blocking point, most existing datasets prioritize object level classification, thus categorizing items by product type though it may be useful for object detection, industrial recycling pipelines rely on material level classification (e.g., distinguishing plastic, metal, paper, or glass) because material composition directly determines how waste is separated and processed.

**Vision Language Models for Waste Recognition**

Vision language models (VLMs) provide a flexible alternative for waste classification. Models like CLIP (Contrastive Language Image Pretraining) [6] learn to connect images and text, so they can classify images by comparing them with text descriptions. This helps zero shot learning, by allowing the model to recognize new categories without additional training data. New studies have successfully adapted CLIP for waste classification via prompt engineering [1], thus demonstrating that multimodal AI models can outperform traditional CNNs in few shot settings [3].

Even with these advantages, zero shot VLMs face challenges in cluttered, real world environments like landfills. When the objects that the model has to classify are damages or partly hidden, VLM may get biased and focus on visual features instead of material. To improve the field based performance and robustness of ecovision, we use zero shot along with few shot and parameter efficient LoRA into the CLIP model [5]. LoRA helps fine tune the model using a small set of labelled image set while freezing the pretrained parameters. This helps the model to adapt to specific wastes environments.

**3. Problem Statement**

Accurate material level classification is a crucial for efficient, automated waste management. While traditional approaches utilizing CNN models trained on datasets like TrashNet [10] achieve high accuracy under clean conditions, the models' performance degrades significantly in real world environments where objects are cluttered, deformed and are under different lighting conditions.

Vision language models (VLMs) like CLIP (Contrastive Language Image Pretraining) [6] provide a reliable alternative. By learning a shared semantic representation between visual and textual embeddings, CLIP enables classification through similarity matching between image features and natural language descriptions(prompts). This helps zero shot inference, allowing the model to recognize new categories without having to retrain. Recent studies show that applying prompt engineering to adapt CLIP for waste recognition have improves performance in few shot settings over traditional convolutional architectures [1], [3].

However, zero shot VLMs still face significant struggle in complex landfill environments. As waste items frequently appear damaged, occluded, or mixed, semantic ambiguity arises during visual and textual matching, this causes the predictions to skew toward dominant visual features rather than actual material composition. Furthermore, since real world facilities must process multiple objects within cluttered scenes, localization is just as important as classification. EcoVision addresses these issues by implementing Grounding DINO [7], an open vocabulary object

detection model that locates waste items using natural language prompts.

When, given an input image captured from a waste processing conveyor belt I ∈ ℝ^(H×W×3), the object detection model produces a set of probable bounding boxes B = {$b_1$, $b_2$, ..., $b_n$}, where each $b_i$ corresponds to a detected waste region.

For each detected region $b_i$, the vision encoder of the VLM generates a visual embedding $v_i = f_{img}(b_i)$, where $f_{img}$ represents the image encoder. Similarly, material categories are defined using textual prompts T = {$t_1$, $t_2$, ..., $t_m$}, which the text encoder converts into semantic embeddings $e_j = f_{text}(t_j)$.

Classification is then performed using cosine similarity between the visual and textual embeddings:

$$S(v_i, e_j) = (v_i \cdot e_j) / (||v_i|| \; ||e_j||) \quad (1)$$

The predicted material class is determined as:

$$\hat{y}_i = \text{argmax}_j \, S(v_i, e_j) \quad (2)$$

While this zero-shot formulation enables open-vocabulary classification, the domain mismatch between pretrained visual features and actual landfill environments can destabilize predictions. To improve the robustness of the model under such conditions, ecovison incorporates Low Rank Adaptation (LoRA) into the CLIP vision encoder [5]. This helps retrain just the top layer of a model by adjusting weights rather than having to retrain the whole model.

$$W' = W + BA \quad (3)$$

here W represents the pretrained weight matrix, and A and B are low rank adaptation matrices. This approach drastically reduces the number of trainable parameters while facilitating highly efficient domain adaptation using only a small subset of labeled waste images. By fusing open-vocabulary object detection with parameter-efficient VLM adaptation, EcoVision delivers a robust material recognition framework for real-world waste segregation.

## 4. Proposed Methodology

Multimodal AI pipelines use visual, textual and IoT sensor data to enhance decision making in complex enviroments[14]. EcoVision combines such a multimodal vision language recognition pipeline with mechanical waste sorting that is designed to work in real world recycling settings. A rotating trommel cylinder first sorts mixed municipal waste by size before sending it to a conveyor belt. This conveyor moves at a steady speed of about 0.5 m/s, which keeps the image stable and reduces motion blur.

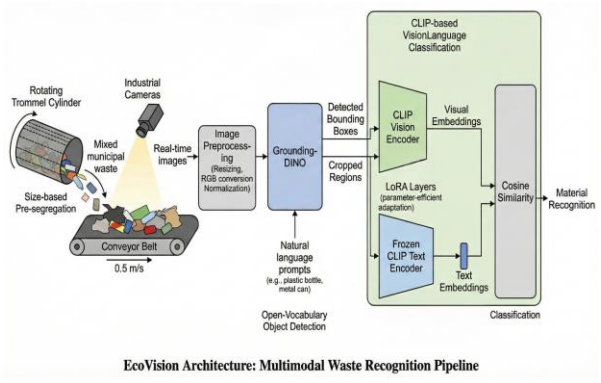


**Fig. 1. Overall Architecture of the EcoVision Multimodal Waste Segregation and Vision Language Classification Pipeline.**

EcoVision's integration with existing trommel-based pipelines was directly based on what was seen in the field at the Perungudi landfill processing facility. Putting the visual recognition module downstream of the mechanical size separation cuts down on the amount of waste the system has to deal with while still working with the current infrastructure in cities.

The overall design of EcoVision (shown in Fig. 1) uses mechanical separation, open-vocabulary object detection, and vision-language classification to make it possible to recognise materials on a large scale.

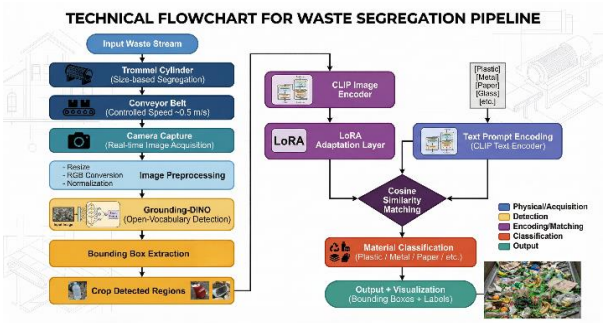


**Fig. 2. Algorithm/ Flow behind ecovision**

**Alogorithm:**

**Input:**

Image I that is captured via camera

Detection prompts

Classification prompt

**Output:**

Labels for detected material

**Steps:**

1. Fetch input Image
2. Preprocess the image:
a. Resize, nomalize and convert image to RGB
3. Image and detections prompts are inputted into Grounding DINO
4. Obtain the bounding boxes
5. For each bounding box in the image, extract the region.
6. If the area of the crop is lesser than threshold then process continues.
7. Get the embedding for the image
8. Apply the LoRA adaptations
9. Normalize the embedding
10. For each class in the prompt, compute the text embedding
11. Compare image and text embedding and get the cosine similarity
12. Select the predicted class (class with highest probability)
13. Output for the bounding box and repeat for every box

**Image Acquisition and Preprocessing**

The waste stream is captured in real time by industrial cameras mounted above the conveyor. The captured frame is represented as I ∈ ℝ^(H×W×3), where H and W stand for the height and width of the image, and the three channels stand for the RGB colour space.

Before inference, the images are preprocessed:

- Changing the size to what the vision encoder needs for input resolution.
- Change to the standard RGB format.
- Normalising the values of the pixels.

This operation is formally defined as:

$I' = P(I)$ (4),

where $P(\cdot)$ is the preprocessing function that is used on the raw input image. EcoVision is different from traditional computer vision pipelines that classify whole images. Instead, it uses a CLIP-based architecture [6] to prepare images for multimodal embedding extraction, making sure that the model's visual and textual encoders work perfectly together.

**Object Detection Using Grounding Dino**

Grounding Dino is used in ecovision for object detection since landfill wastes have overlapping objects. Grounding-DINO is different from regular detectors that only work with fixed label sets. It can also work with natural language prompts like "plastic bottle," "metal can," "cardboard box," and "paper waste."

The detector predicts a set of candidate bounding boxes based on the preprocessed input image I' and a set of textual detection prompts

$T_d = \{t_1, t_2, ..., t_k\}$:

$B = \{b_1, b_2, ..., b_n\}$ (5)

where each $b_i = (x_1, y_1, x_2, y_2)$ shows the top-left and bottom-right corners of a detected waste area.

Then, each region is cut out of the original image: $C_i = \text{Crop}(I', b_i)$ (6),

and sent on for material classification. Separating localisation from classification makes the system more modular, so new object types can be added quickly without having to retrain the detector.

**Vision–Language Classification with CLIP**

Using descriptive text prompts and the CLIP text encoder [6], we define material categories like plastic, metal, paper, glass, cardboard, organic, and textile. The vision encoder $f_{img}(\cdot)$ makes a visual embedding for each cropped waste region $C_i$:

$f_{img}(C_i) = v_i$ (7)

The text encoder $f_{text}(\cdot)$ also turns each material prompt into a textual embedding:

$e_j = f_{text}(t_j)$ (8).

Based on the cosine similarity the object is classified:

$\text{Sim}(v_i, e_j) = (v_i . e_j)/(||v_i|| \; ||e_j||)$ (9)

The predicted material class is:

$y_i = \text{argmax}_j \; \text{Sim}(v_i, e_j)$(10)

The robustness of few shot in ecovision is improved by taking average of few prompts per class as this will make the model less sensitive to the wording of the prompt.

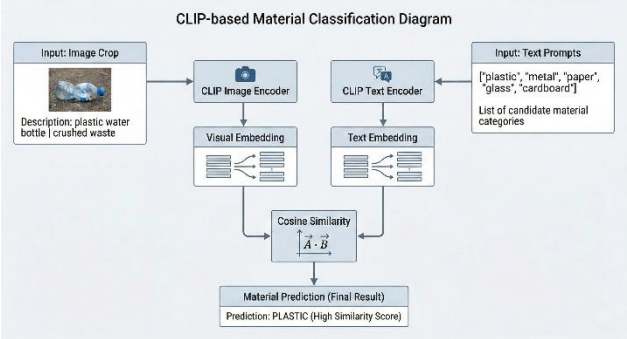


**Fig. 3. Vision-Language Similarity Matching**

**LoRA**

The zero shot classification of CLIP is strong but the features that it was trained with will not understand the context of landfills. It has been proven that inserting LoRA adapters into CLIP resulted in good improvements in accuracy[17]. LoRA is used in ecovision in the projection layers of CLIP to train CLIP and help it understand landfill context. [5] Low rank matrices are used by LoRA to change original weight matrix as this will help avoid changing the whole network.

$W' = W + BA$ (11)

here W is the frozen weight matrix and A and B are low-rank matrices that can be trained. Doing so will help reduce the number of trainable parameters but keeps the knowledge of the base model intact. Using small field collected dataset to train these parameters will help make the model work efficiently based on how landfills objects are cluttered and this also helps the model generalize better.

**Inference Pipeline and Deployment**

While classifying a detected object, the model does the four following steps to process the image,

- Grounding DINO uses prompts to identify possible objects in an image.
- The region of the detected objects of cropped with bounding boxes.
- CLIP model that has been retrained using LoRA is applied to each bounding box.
- Each bounding box is labelled.

This pipeline will work well on low end hardwares as well and one can add new categoried without training the whole model again.

**5. Applications**

EcoVision improves the efficiency of the existing mechanical waste processing processes employed in waste processing plants by using a material level recognition model. Eco vision proves to be better than traditional systems where size based sorting is performed to classify wastes and it picks up where the traditional size based sorting systems fail as the model supports a wide range of waste management and environmental monitoring by combining open vocabulary object detection with multi modal material based classification and since ecovision is implemented right at the conveyor belt, it reduces latency and computational overhead[15].

**Municipal Waste Processing Facilities**

In most of the municipal treatment plants they use mechanical equipment like trommel cylinders to sort the waste by size instead of composition, therefore the recyclable objects often get stuck with organic or inert objects.

EcoVision is used right after this type of manual segregation in the downstream to classify items on the conveyor belt based on their material. Grounding DINO is used to find single objects, and CLIP based multimodal embeddings predict the kind of material the detected objects are.

EcoVision gives an output for each detected object i:

$O_i = (x_i, y_i, w_i, h_i, c_i)$ (12)

here $x_i$ and $y_i$ define the location of the object, $w_i$ and $h_i$ define the size of the bounding box, and $c_i$ defines the predicted material label. These coordinates can help robotic arms, air jets, or conveyor diverters send recyclables to the right processing streams.

**Material Recovery Facilities (MRFs)**

Although MRFs sort through huge amounts of mixed waste to find plastics, metals, and paper, the damaged packaging, food residue, and mixed materials pose a contamination issue. EcoVision tackles this issue by classifying objects based on their materials instead of objects themselves. Instead of classifying things into groups based only on their shape, the system predicts what the material is made out of.

| **Object** | **Conventional Detection** | **EcoVision Output** |
|---|---|---|
| Plastic bottle | Bottle | Plastic |

| Object | Conventional Detection | EcoVision Output |
|---|---|---|
| Crushed aluminium can | Can | Metal |
| Torn cardboard packaging | Box | Cardboard |

This kind of differentiation helps ensure that recycling systems correctly route materials even when objects are heavily deformed.

**Smart Waste Collection Systems**

The ecovision model can also be deployed and be used with smart trash cans thus helping improve the collection of wastes right from the neighbourhood even before they reach the landfill. With ecovision inside collection bins, can keep an eye on:

- Waste composition
- Recycling compliance
- Contamination levels

Aggregating these classification outputs across multiple area allows EcoVision to generate actionable, data driven insights into city wide waste patterns.

**Environmental Monitoring and Waste Analytics**

EcoVision is an efficient tool that can not only be used in factories but also for environmental monitoring programs to keep an eye on how trash is spread in public places. Some important uses are:

- Roadside litter monitoring
- Landfill composition analysis
- Coastal waste tracking

By processing the captured images across different area and calculating the frequency of specific material type, EcoVision can compute the relative proportion of waste materials:

$P_c = N_c/N_{total}$ (13)

here $N_c$ is the number of detected objects belonging to class c, and $N_{total}$ is the overall count of detected waste objects. These metrics can be used for planning, targeted waste reduction strategies, and recycling infrastructure optimization.

Newly developing smart cities use a lot of IoT architectures that can handle real time data, hence a specific waste management system that can handle real time waste data in required [13].

**6. Experimental Results and Discussion**

Ecovision has been tested on a few datasets to determine its performance with no training (zero shot), and to check how much it can improve when it is trained with few samples, then finally how it performs with landfill images. The model was evaluated using TrashNet dataset[10], TACO dataset [9] and custom field dataset. This test helps determine how well the model performs on a cluttered scene.

**Table 1: Datasets used for Ecovision Training and Evaluation**

| Dataset | Description | Classes | No. of Images | Characteristics | Usage |
|---|---|---|---|---|---|
| TrashNet [10] | Curated waste dataset | 6 | ~2,500 | Clean, centered objects | Zero shot baseline |
| TACO [9] | Real world litter dataset | 28 (mapped) | ~1,500 | Heavy clutter, multiple objects | Robustness evaluation |
| Custom Cluttered Subset | Real world cluttered samples | 8 | ~100 | Occlusion, mixed waste | Robustness adaptation |

For compatibility across datasets, TACO's **28 object categories were mapped to the seven material classes used in EcoVision**, namely cardboard, glass, metal, organic, paper, plastic, and textile.

**Experimental Setup**

All experiments utilized a CLIP-based vision–language model [6] integrated with Grounding-DINO for open-vocabulary object detection [7]. The evaluation pipeline operates in three sequential stages:

- Open-vocabulary detection using Grounding-DINO.
- Bounding box cropping and preprocessing.
- Material classification utilizing CLIP embeddings.

During inference, Grounding-DINO initially processes an image I to produce a set of bounding boxes:

$B = \{b_1, b_2, ..., b_n\}$ (17)

where each bounding box $b_i = (x_i, y_i, w_i, h_i)$ represents the location and dimensions of a detected waste object. Each region is then cropped and forwarded to the CLIP encoder to extract visual embeddings $v_i$.

Material classification is performed by computing the cosine similarity between the visual embeddings and the textual prompt embeddings:

$Sim(v, t) = (v \cdot t) / (||v|| \ ||t||)$ (18)

where v is the visual embedding and t is the textual embedding of a material prompt. The class with the highest similarity score is outputted as the material label.

**Zero Shot and Few Shot Evaluation**

Zero shot test aims to check the ability of CLIP model to classify waste materials with no field based training and by only using the prompts that is given. The CLIP model's text encoder takes a prompt for a class c and turns it into a semantic embedding $t_c$. Based on the cosine similarity the class is then chosen as label.

$c = \text{argmax}_c \, Sim(v, t_c)$ (19)

While testing zero shot CLIP model with TrashNet dataset, it got an accuracy score of 62%, this shows that CLIP model can classify most items without any retraining required.

Few shot prompts were used to make the model more stable. Multiple prompts per class were used in order to take into account the variations that an object can have while on a conveyor belt in a landfill, it might be dirty, wet or cluttered and it might also be crushed and different in shape. The final text embedding for each class has been achieved by taking the average of these embeddings.

$t_c = (1/k) \, Sum(f_{text}(pi))$ (20)

here k is the number of prompts for the class c, and the summation starts from i=1 till k. This multi prompting method per class improved the accuracy of the model upto 68%. This proves that multi layering prompt to describe to address all the various ways an object can look in a landfill improves the alignment of visual and text embeddings, which has also been stated in another research on the essential function of the prompt engineering in VLM [1]. It has also been observed that using LLM refined prompts improve the zero shot performance on top of this, a pseudo labelling scheme can even exceed a supervised detector [18].

**LoRA Based Adaptation**

Low-Rank Adaptation (LoRA) was added to the CLIP vision encoder to make performance even better in real-world situations. LoRA doesn't fine-tune the whole model; instead, it adds trainable low-rank matrices to the projection layers while keeping the pretrained backbone frozen. The modified projection layer is defined as:

$W' = W + BA$ (21)

W is the weight matrix that has been trained before, and B and A are the low-rank matrices that can be trained. This formulation, which is very efficient in terms of parameters, greatly lowers training costs while making domain adaptation very specific.

A subset of TACO images and 81 custom field images with cluttered backgrounds, mixed materials, and partially damaged waste were used to find the best LoRA parameters. Adding LoRA improved classification accuracy from 67% to 70.9%, which shows that lightweight parameter adaptation can successfully handle domain shift.

**Multi-Object Detection Performance**

Operational waste sorting necessitates the concurrent identification of various objects within a singular frame. To assess this, EcoVision underwent testing with Grounding-DINO detection in conjunction with CLIP-LoRA classification. If at least one detected object matched the ground truth material class, the image was correctly classified.

If in an image there are m detected objects,

$P = \{p_1, p_2, p_3, ..., p_m\}$

then the prediction is true when there is a $p_i$ in P such that $p_i = c_{true}$ (22)

here $c_{true}$ is the real label. Based on the prompt and the detection thresholds, multi object evaluation had an accuracy rate of 79–84%. This improvement in the accuracy of the model in classifying objects when compared to single object classification shows that the integration of the open vocabulary detection with multimodal classification improves the reliability of the model in a chaotic environment.

**Fig. 4. Multi object classification result from ecovision**

**Discussion**

The initial results with zero shot testing yielded results that shows the CLIP model with zero shot learning exihibits very reliable generalization, thus improving the recognition of waste waste materials without depending on a huge annotated datasets like existing CNN approaches.

Additionally, using hierarchical prompt engineering clearly improves classification accuracy by introducing important contextual information into textual embeddings. This method allows the model to effectively differentiate visually similar materials, such as cardboard, paper, and organic waste.

On top of prompt engineering, using LoRA to adapt the weights of a pretrained model based on a few field specific images in a dataset helps improve the performance of the model even further without actually impacting the base zero shot abilities of the model. This is because the LoRA optimizes only lightweight adapter parameters, the system adapts to landfill specific context using a minimal dataset. Hence, EcoVision requires considerably lesser annotated data than traditional supervised CNN models [4], but still retaining highly competitive performance in cluttered scenes.

But, there is a limitation in the current visual embedding space as the spectral similarity between soiled cardboard and organic waste, especially in a compost like environment. These visual features alone sometimes fail to capture these little material deviations. This can study can be further explored to investigate advanced hierarchical prompting strategies alongside multimodal sensor fusion such as near infrared (NIR) spectroscopy to enhance discrimination between visually overlapping materials. Optimization based approaches are being explored in smart city systems to improve efficiency and sustainability[16] hence this methodology can surely further optimized in the future.

**Acknowledgment**

I would like to sincerely thank the project supervisor Dr. B. Amutha and the Department of Artificial Intelligence and Data Science, SRM Institute of Science and Technology, for their guidance, support, and valuable feedback throughout the course of this work.

**References**

[1] H. J. Malla, M. Bazli, and M. Arashpour, “Enhancing waste recognition with vision-language models: A prompt engineering approach for a scalable solution,” Waste Management, vol. 204, p. 114939, 2025.

[2] T. Trang, H. V. Pham, S. D. Vu, T. M. Le, H. M. Tran, and S. V. T. Dao, “TrashVLM: Lightweight and efficiently fine-tuned vision-language models for waste classification,” in Proc. Int. Conf. Advanced Technologies for Communications (ATC), 2025, doi: 10.1109/ATC67618.2025.11268574.

[3] J. Funk, P. Bäcker, L. Roming, J. Josekutty, G. Maier, and T. Längle, “A comparative evaluation of vision–language models for waste classification in few-shot settings,” Fraunhofer IOSB, Karlsruhe, Germany, Tech. Rep., 2024.

[4] Md. Nahiduzzaman, Md. Faysal Ahamed, Mansura Naznine, Md. Jawadul Karim, Hafsa Binte Kibria, Mohamed Arselene Ayari, Amith Khandakar, Azad Ashraf, Mominul Ahsan, Julfikar Haider, "An automated waste classification system using deep learning techniques: Toward efficient waste recycling and environmental sustainability”, Knowledge-Based Systems, Volume 310, 2025, 113028, ISSN 0950-7051

[5] E. J. Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models", arXiv preprint arXiv:2106.09685, 2021.

[6] A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. Int. Conf. Machine Learning (ICML), 2021, pp. 8748–8763.

[7] S. Liu et al., “Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection,” arXiv preprint arXiv:2303.05499, 2023.

[8] J. Carion et al., “End-to-End object detection with transformers,” in Proc. European Conf. Computer Vision (ECCV), 2020, pp. 213–229.

[9] P. Proença and S. Simões, "TACO: Trash annotations in context for litter detection," arXiv preprint arXiv:2003.06975, 2020.

[10] G. Yang and Y. Thung, "TrashNet dataset," Kaggle Dataset, 2016. [Online]. Available: https://github.com/garythung/trashnet

[11] M. Minderer et al., "Revisiting the calibration of modern neural networks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 34, 2021, pp. 15682–15694.

[12] S. Delsi Robinsha and B. Amutha, "IoT architecture for energy management in smart cities," International Journal of Services Operations and Informatics, vol. 12, no. 4, pp. 325–343, 2023.

[13] Robinsha, S.D., Amutha, B. Velocious: A Resilient IoT Architecture for 6G Based Intelligent Transportation System with Expeditious Movement Mechanism. *Wireless Pers Commun* (2024). https://doi.org/10.1007/s11277-024-11072-9

[14] Delsi Robinsha, S., Amutha, B. (2025). Interfacing Multi-modal AI with IoT: Unlocking New Frontiers. In: Singh, A., Singh, K.K. (eds) Multimodal Generative AI. Springer, Singapore. https://doi.org/10.1007/978-981-96-2355-6_14

[15] Amutha, B. "Edge Computing for Energy-Efficient Internet of Things: Concepts, Technologies." Innovation and Sustainability in Electric and Autonomous Mobility (2025): 209.

[16] Ranjusha, K.P., Amutha, B. Sustainable development of E-mobility in urban areas using knowledge-based artificial network (KANM), behavioral learning theory (BLT) and distributed optimization algorithm (DOA). Int. j. inf. tecnol. (2026). https://doi.org/10.1007/s41870-025-03115-6

[17] M. Zanella and I. Ben Ayed (2024) – *"Low-Rank Few-Shot Adaptation of Vision-Language Models," in Proc. IEEE/CVPR Workshops, 2024, pp. 1593–1603.*

[18] H. Abid et al. (2025) – "Robust and Label-Efficient Deep Waste Detection," BMVC 2025 (arXiv: 2508.18799).