Vision-Powered Cleaning Mode Detection in Robot Vacuum Cleaners Using a Novel Smart Home Dataset

Sanket Gharatkar
Samsung Research Institute Bangalore
Email: s.gharatkar@samsung.com

Suhel Khan
Samsung Research Institute Bangalore
Email: suhel.khan@samsung.com

Nagaraj Yendeti
Samsung Research Institute Bangalore
Email: yn.raj@samsung.com

***Abstract*—Domestic robot vacuum cleaners are widely used for automating floor cleaning and typically support multiple cleaning modes such as vacuum cleaning, mopping, and both. Accurate selection of the cleaning mode is essential for efficient operation and improved user experience. In this paper, we introduce a novel smart home dataset, Smart Clean, containing images of different floor types contaminated with common household waste. Using this dataset, we fine-tune Large Vision Models (LVMs) to automatically classify the appropriate cleaning mode among vacuum cleaning, mopping, or both. We present a comparative analysis of LVMs and evaluate their performance against traditional deep learning models such as MobileNetV2, ResNet152V2, and VGG19, as well as classical image processing approaches where certain dirt regions may remain undetected. We further evaluate model robustness under challenging conditions including floor reflections, blurred images, dirt size variations, and complex floor patterns. Experimental results show that our best-performing model, Phi-3.5-vision-instruct, outperforms the strongest CNN baseline (MobileNetV2) by 40% on the Smart Clean dataset. Additionally, our model achieves 66% accuracy in cases where image processing methods fail to detect certain regions. The proposed framework demonstrates strong potential for enabling intelligent cleaning mode selection in robotic vacuum systems, improving cleaning efficiency and enhancing user experience.**

***Index Terms*—Large Vision Models, Robot Vacuum Cleaners, Cleaning Mode Classification, Smart Clean Dataset**

I. Introduction

Robotic vacuum cleaners have become an important component of smart homes by automating routine floor cleaning tasks and improving user convenience. Modern devices support multiple cleaning modes—vacuuming, mopping, and vacuuming & mopping—but typically rely on user-selected or pre-configured settings. This rigid design can lead to suboptimal cleaning, inefficient resource usage, and reduced user satisfaction when the selected mode does not match the floor condition. For example, using vacuum mode on spilled water may damage the robot and result in ineffective cleaning, while mopping dry debris wastes energy and leaves dirt uncollected. These inefficiencies reduce cleaning effectiveness and can increase maintenance costs and user frustration. Table I illustrates how mismatched cleaning modes lead to inefficient operations and negative user feedback, whereas selecting the appropriate mode ensures optimal cleaning and a better user experience. These scenarios highlight the need for intelligent, context-aware decision-making in robot vacuum systems. To address this challenge, we propose an automated vision-based cleaning mode classification system that enables robotic vacuum cleaners (RVCs) to select the appropriate cleaning mode based on real-time floor observations. Central to our approach is *Smart Clean*, a novel dataset containing over 6000 images of dirty floors captured in real-world home environments across different floor types (e.g., tile, marble) and dirt categories (dry and wet). The dataset includes challenging conditions such as mixed waste types, reflections, blurriness, and complex floor patterns, enabling models to generalize to diverse environments. Using the Smart Clean dataset, we fine-tune Large Vision Models (LVMs) and compare their performance with CNN-based deep learning models such as MobileNetV2 [1], ResNet152V2 [2], and VGG19 [3], as well as classical image processing techniques such as MSER [4]. Experimental results show that LVM-based approaches provide significantly improved accuracy and robustness across challenging scenarios. Our best-performing model, Phi-3.5-vision-instruct [5], achieves superior performance, particularly in complex and edge cases.

TABLE I: Examples of Efficient and Inefficient cleaning operation scenarios

Waste Type	Cleaning Mode	Efficient Operation	User Feedback
Water Spilled	Mopping	Yes	Positive
Water Spilled	Vacuum Cleaning	No	Negative
Small paper pieces	Vacuum Cleaning	Yes	Positive
Small paper pieces	Mopping	No	Negative
Small paper pieces and water both	Vacuum Cleaning and Mopping	Yes	Positive
Small paper pieces and water both	Vacuum cleaning and Mopping	No	Negative

II. Related Work

Robotic vacuum cleaners (RVCs) have transformed domestic cleaning by automating floor maintenance tasks, with many offering cleaning modes such as vacuuming, mopping, or a hybrid of both. However, most commercial RVCs rely on user-selected or manually configured cleaning modes, which lack contextual intelligence and often result in suboptimal cleaning and user dissatisfaction. Early approaches in adaptive cleaning explored sensor-based heuristics for floor mapping and obstacle detection [6], but these techniques were limited to spatial navigation rather than understanding floor conditions. Traditional image processing methods such as MSER [4] (Maximally Stable Extremal Regions) and edge-based segmentation were explored for dirt detection, but these

methods often failed in complex household environments with variable lighting, floor reflections, different dirt sizes and mixed waste types. In the domain of computer vision, convolutional neural networks (CNNs) like MobileNetV2 [1] and ResNet [7] have been applied to image classification tasks with reasonable success in controlled environments. However, their generalization fails when deployed in real homes, especially in scenarios involving small or partially visible dirt, reflections, and floor texture ambiguity. Our work aims to bridge this gap by introducing a novel, real-world dataset tailored for smart cleaning and leveraging LVMs to classify floor conditions and infer the correct cleaning mode autonomously. Unlike traditional methods that focus on dirt detection alone, our approach integrates floor type and dirt characteristics to ensure optimal cleaning efficiency. Additionally, we benchmark our system against both CNN-based baselines (e.g., MobileNetV2 [1], ResNet152V2 [2], VGG19 [3]) and classical image processing techniques (e.g., MSER [4]), demonstrating significant improvements in accuracy and robustness across diverse scenarios.

III. Proposed Method

We propose an architecture that enables robotic vacuum cleaners (RVCs) to autonomously identify the most suitable cleaning mode (vacuum, mop, or vacuum & mop) using real-time visual input of the floor. This vision-based system addresses the limitations of existing RVCs that lack contextual awareness of floor or dirt type and rely on static cleaning configurations. The architecture integrates Large Vision Models (LVMs) directly into the robotic vacuum's processing pipeline, enabling intelligent, context-aware decision-making for optimal cleaning efficiency.

A. System Architecture

1. *Camera Control Unit (CCU):* The robot's onboard camera captures real-time floor images during movement. This unit ensures high-resolution, low-latency image acquisition, which is critical for accurate cleaning mode classification.

2. *Advanced Cleaning Detection Module (ACDM):* This module acts as the core vision intelligence, where the captured image is buffered and analyzed using fine-tuned LVMs. These models classify the scene into one of the three cleaning modes based on the detected floor type and dirt characteristics. For example, if the image shows water spilled on a tiled floor, the model predicts "Mopping" as the appropriate cleaning mode.

3. *Control Unit Integration:* Once a prediction is made, the robot's control unit translates the classification result into actionable cleaning instructions, such as turning on the mop module, vacuum fan, or both. This integration ensures seamless operation and minimizes delays between detection and action.

This modular architecture is designed to be adaptable to newer environments and compatible with future LVMs or sensing mechanisms, making it a scalable solution for smart cleaning applications.

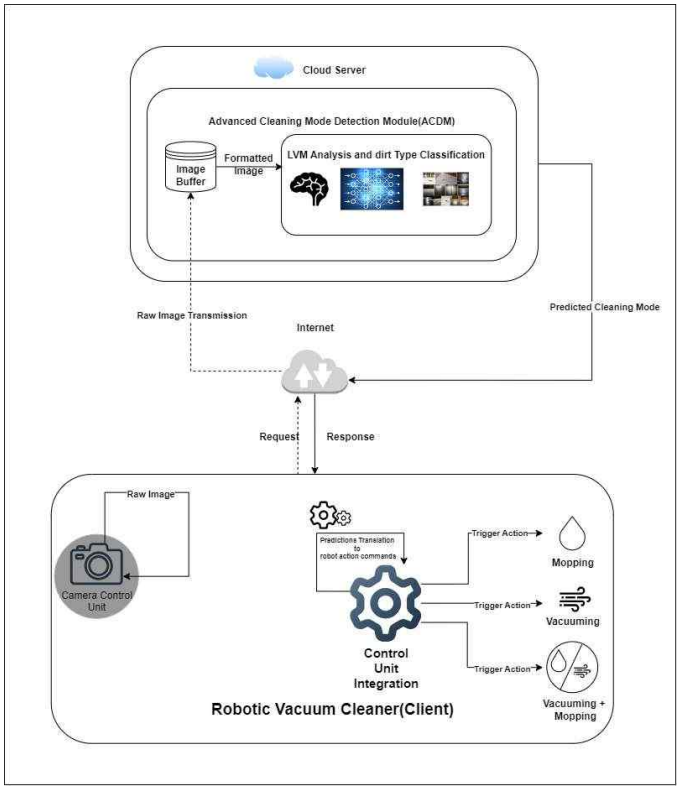

Fig. 1: Architecture diagram for Proposed Architecture.

B. Smart Clean Dataset

To support this architecture, we created Smart Clean [8][9][10], a novel dataset, comprises over 6,000 images in PIL format, as illustrated in Figure 2. These images were captured and refined using a combination of automated tools and human curation to ensure high quality and diversity. The dataset is systematically organized with images classified into three distinct classes: Vacuum Cleaning, Mopping, and Vacuum Cleaning & Mopping. This classification was achieved through a meticulous review process by reviewers. The dataset captures diverse real-world smart home environments with varying:

· Floor Types (tile, marble, wood)
· Dirt Types (wet: water, mud; dry: paper, dust)
· Dirt Combinations (e.g., mixed wet and dry patches)
· Floor reflections, zoom-level variation, and blurs due to robot motion

Table 3 provides an overview of the dataset distribution across training, validation, and test sets.

To enhance robustness, we applied a few augmentation techniques, including blur, invert, crop, zoom and noise addition, to simulate real-world distortions. This ensures that the models generalize well to challenging scenarios encountered in everyday use.

TABLE II: Smart Clean dataset class distribution

Dataset Split	Vacuum Cleaning	Mopping	Vacuum Cleaning & Mopping
Train	2511 images	1981 images	860 images
Validation	282 images	251 images	87 images
Test	72 images	69 images	71 images

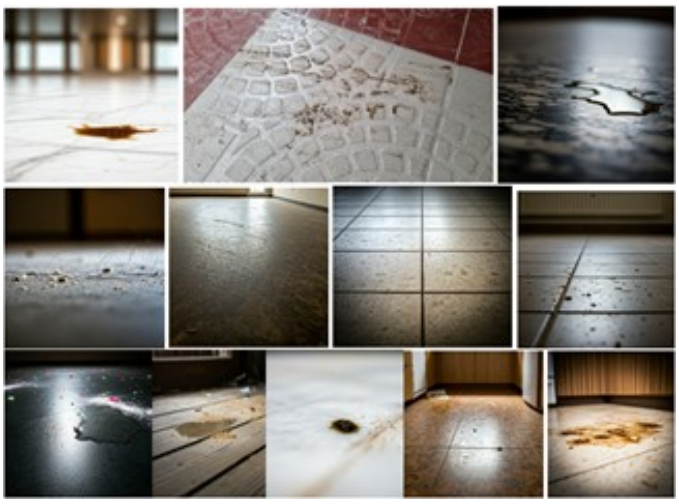

Fig. 2: Some images from Smart Clean dataset.

C. Data Augmentation

To enhance the robustness of the fine-tuned models and simulate real-world variations encountered by robotic vacuum cleaners, we applied a series of controlled image augmentations on the Smart Clean dataset. These augmentations were designed to expose the models to visual distortions typical of home environments, such as lighting inconsistencies, motion blur, and sensor noise. The following transformations were applied:

· **Gaussian Blur**:

Ø Simulating camera motion during robot movement: (radius = 0.5)

· **Brightness Adjustment**:

Ø Reduced Brightness (factor = 0.80)

Ø Increased Brightness (factor = 1.2)

· **Contrast Adjustment**:

Ø Enhancing or flattening texture features: (factor = 1.0)

· **Horizontal Flip**:

Ø Introducing spatial variation due to robot path symmetry.

· **Gaussian Noise Injection**:

Ø Modelling sensor or environment-induced image noise: (mean = 0.0, variance = 0.05)

Ø **Zoom**

D. Model Pipeline

We fine-tuned six state-of-the-art instruction-tuned LVMs, like Phi-3.5-vision-instruct , LLaVA-v1.5-7b, and others on the Smart Clean dataset. Each image is paired with a prompt (e.g., "Identify the suitable cleaning mode based on dirt and floor condition"). The model returns one of three classes: Vacuum_Cleaning, Mopping, or Vacuum_Cleaning_And_Mopping. These predictions are then passed to the RVC's control module, which configures mechanical subsystems accordingly.

To benchmark our system, we compare it against:

· MobileNetV2 [1], ResNet152V2 [2] and VGG19 [3]: CNN-based baselines commonly used in image classification tasks.

· MSER [4] based Image Processing: A classical approach that fails in cases like reflections or partial dirt detection.

E. Challenges Addressed

The proposed method is designed to handle real-world edge cases that commonly break traditional systems:

· **Reflections**: Accurate detection of dirt on glossy surfaces.

· **Dirt Size Variations**: Classification of small or partially visible dirt patches.

· **Complex Floor Patterns**: Distinguishing between dirt and intricate floor textures.

· **Blurred Images**: Robustness to motion blur caused by robot movement.

By addressing these challenges, our system ensures reliable cleaning mode classification across diverse environments, significantly improving operational efficiency and user satisfaction.

IV. Experimental Results

A. Fine-Tuning Setup

We fine-tuned a few Large Vision Models (LVMs) on the Smart Clean dataset. The models were trained for multi-class classification to predict one of the three cleaning modes: *Vacuum_Cleaning, Mopping, or Vacuum_Cleaning_And_Mopping.* All images were preprocessed using augmentation transformations, such as blur, zoom, noise addition, altering brightness, rotations and inversion, to simulate real-world robot scenarios. We employed cross-entropy loss as the training objective.

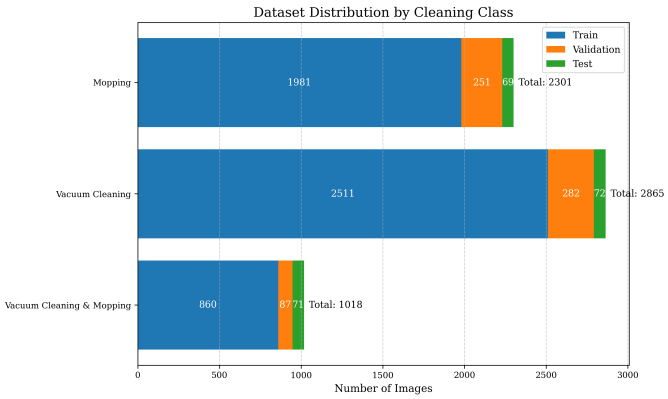

Fig. 3: Dataset Distribution by Cleaning Modes.

The dataset was split into training, validation, and test sets as detailed in Figure 3. The experimental setup was defined by key hyperparameters—in Table III including the number of epochs, learning rate, optimizer, and batch size—which were individually tuned for each model to maximize performance. To evaluate the models on the test set, we employed standard metrics, including **Accuracy, Precision, Recall, F1-Score, Confusion Matrices, and per-class classification reports.** These metrics were computed across all models to systematically gauge classification quality and generalization.

TABLE III: Fine-tuning Configuration

Parameter	Value
Learning Rate	1e-4
Batch Size	32
Epochs (CNN Baselines)	30 (Till Early Stopping)
Epochs (Transformer-based Vision Models)	2
Loss Function	Cross-Entropy loss
Optimizer	Adam
Initialization	Transfer Learning

B. Performance Comparison

Fine-tuning significantly improved vision-language models, with Phi-3 achieving 94% accuracy and LLaVA reaching 91%, outperforming traditional models like MobileNetV2 [1], ResNet152V2 [2] and VGG19 [3].We benchmarked the performance of our fine-tuned LVMs against MobileNetV2 [1], ResNet152V2 [2], VGG19 [3] (baseline CNN architecture).Both LVMs show a significant improvement over best-performing CNN-based deeplearning approach MobileNetV2 [1], indicating their superior capability in complex real-world environments. As we can observe from Table IV, Figure 4 and Figure 5, the fine-tuning consistently improved the performance of all tested models, as evidenced by their enhanced accuracy, precision, recall, F1 scores, confusion matrix, per-class classification charts.

TABLE IV: Performance comparison among models

Model	Accuracy	Precision	Recall	F1 Score
ResNet152V2 (fine-tuned)	58%	60%	58%	55%
VGG19 (fine-tuned)	40%	26%	40%	28%
MobileNet V2 (fine-tuned)	67%	71%	67%	65%
llava-v1.5-7b(vanilla/base)	56%	70%	55%	53%
llava-v1.5-7b (fine-tuned)	91%	91%	91%	91%
Phi-3.5-vision-instruct (vanilla/base)	63%	69%	63%	61%
Phi-3.5-vision-instruct (fine-tuned)	94%	94%	94%	94%
Idefics2-8b (vanilla/base)	47.5%	45.68%	78.72%	57.81%
Idefics2-8b (fine-tuned)	86.89%	86.75%	86.79%	86.73%
Moondream2 (vanilla/base)	53.1%	51.42%	51.6%	51.53%
Moondream2 (fine-tuned)	81.60%	81.68%	81.60%	81.63%
Llama-3.2-11B-Vision (vanilla/base)	37.6%	47.7%	55.31%	51.23%
Llama-3.2-11B-Vision (fine-tuned)	77%	77%	77%	77%
Pixtral-12B (vanilla/base)	24.5%	54.41%	13.12%	21.14%
Pixtral-12B (fine-tuned)	69%	70.6%	69.8%	70.1%

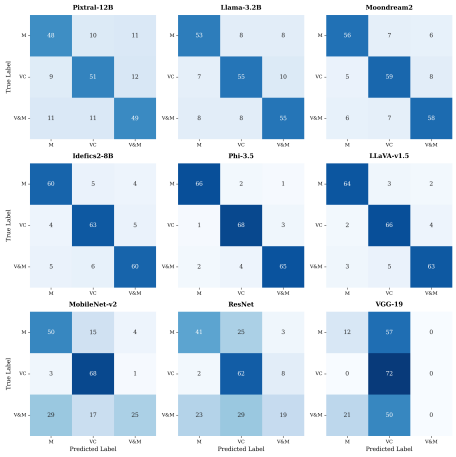

Fig. 4: Confusion Matrices of Different Trained Models

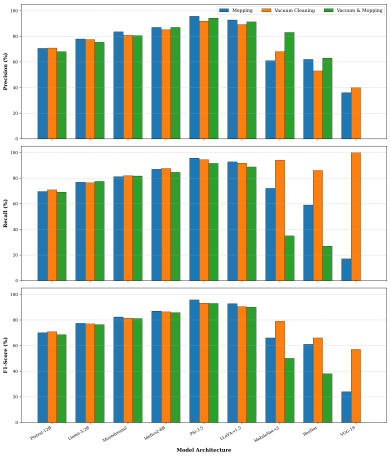

Fig. 5: Per class classification of Different Trained Models

C. Robustness in Challenging Scenarios

To test generalization, we evaluated models on specific dedicated test sets containing challenging cases such as Reflections on shiny floors, Varying dirt size, Blur due to robot movement, and Complex floor textures. Both Phi-3 [5] and LLaVA [11] performed exceptionally well, with Phi-3 [5] demonstrating high accuracy in detecting small dirt particles and in blurred images. LLaVA [11] also showed strong performance with a recall rate of 91.6% in complex patterned floors. Some of the sample images from challenging scenarios are provided in Figure 6.

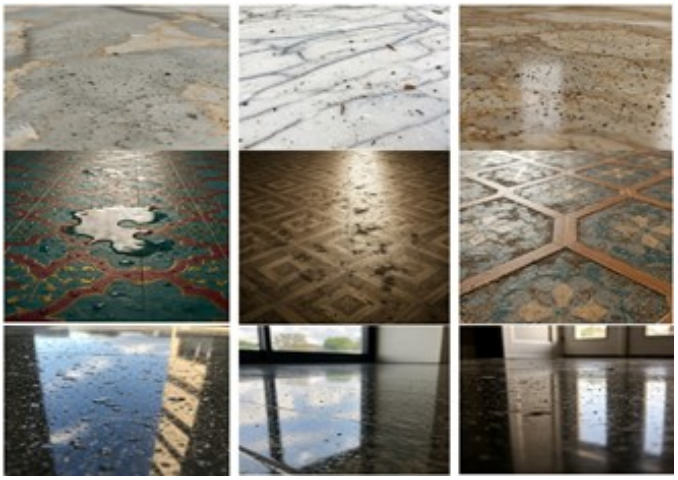

Fig. 6: Some of the Challenging Images

D. Comparison with Classical MSER

To assess the effectiveness of our proposed method in contrast to traditional image processing techniques, we conducted an in-depth evaluation against the MSER [4], which has been employed in prior dirt detection systems for robotic cleaners. While MSER [4] can identify high-contrast regions and segment dirt in controlled conditions, it often fails under real-world challenges such as very small dirt sizes which is very common in smart home environments.

To better understand this limitation, we performed a **targeted evaluation:**

· First, we extracted the dirt regions detected by MSER [4].

· We then created new modified images by masking or removing these detected regions, thus preserving only the undetected dirt regions missed by MSER [4].

These residual images were then passed to our fine-tuned Phi-3.5-vision-instruct model to evaluate whether it could still predict the correct cleaning mode using incomplete visual information. Despite having access to only partial visual cues, the model correctly identified the appropriate cleaning mode in most cases, demonstrating strong contextual understanding and generalization capabilities. This experiment highlights the model's ability to infer scene semantics even when significant dirt regions remain undetected by traditional image processing techniques. Figure 7 presents representative examples of these modified images, while Table V summarizes the corresponding performance results.

TABLE V: Model performance on undetected cases in Image Processing MSER [4] approach

Model	Accuracy	Precision	Recall	F1 Score
Phi-3.5-Vision	66%	66%	100%	79.5%

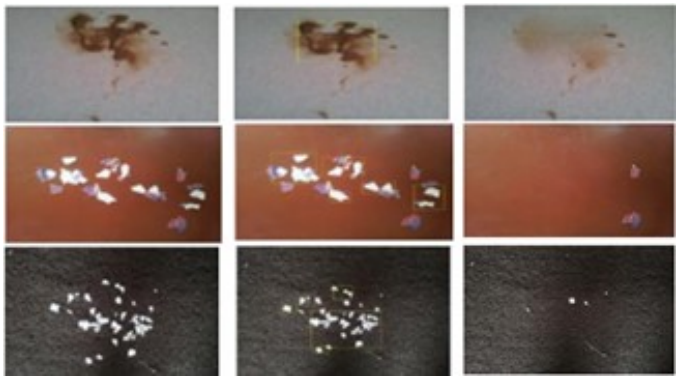

Fig. 7: Sample Images of Various Dirt from Home Environment (a. Original, b. MSER [4] Detected, c. Modified Images).

V. Results

We report the performance of our models evaluated on the Smart Clean dataset using standard classification metrics: Accuracy, Precision, Recall, and F1 score. The fine-tuned Large Vision Models (LVMs), e.g., Phi-3.5-vision-instruct and LLaVA-v1.5-7b significantly outperform both the classical image processing baseline (MSER [4]) and the lightweight convolutional model (MobileNetV2 [1], ResNet152V2 [2] and VGG19 [3]). Phi-3.5 demonstrates the best overall performance, showing substantial improvements of nearly 40% over the best performing CNN based deep learning model MobileNetV2 [1] in accuracy. The performance gap widens when comparing to MSER [4], a traditional image processing method, especially in failure-prone scenarios. Phi-3.5 achieves an F1 score of $\tilde{7}9.5\%$ on cases where MSER [4] failed to provide usable predictions, demonstrating LVMs' superior ability to generalize across complex floor-waste patterns. Overall, these results validate the effectiveness of our Smart Clean dataset and the proposed approach for real-world robotic cleaning scenarios, with LVMs showing strong potential for deployment in intelligent robotic vacuum systems.

VI. Conclusion

This paper presents Smart Clean [8][9][10], a novel image dataset designed to address the limitations of static, rule-based cleaning mode selection in robotic vacuum cleaners. Captured from real-world home environments using an on-board camera, the dataset includes over 6000 annotated images across diverse floor types and dirt categories, including challenging conditions such as reflections, motion blur, and complex floor textures. It serves as the foundation for evaluating vision-based cleaning mode classification systems. We fine-tuned a few large vision-language models on the Smart Clean dataset and benchmarked them against MobileNetV2 [1], ResNet152V2 [2] and VGG19 [3] and classical image processing (MSER [4]). The results show that fine-tuned LVMs deliver significant performance improvements over traditional models, especially Phi-3.5 [5] and LLaVA [11]. These improvements extend to edge-case scenarios, where models demonstrated high robustness across reflection handling, dirt size variation, and motion blur. On the curated test dataset, Phi-3.5 achieved 94% accuracy, while LLaVA achieved 91%. These findings highlight the potential of large vision models in enabling autonomous, context-aware cleaning systems that adapt in real-time to floor conditions. Our work demonstrates a viable path toward closing the perception-action gap in robotic vacuum cleaners using deep multimodal architectures.

References

[1] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," 2019.

[2] Keras, "Resnet152v2," 2016.

[3] ——, "Vgg19," 2014.

[4] T. Milinda and A. Madhusanka, "Multi-functional floor cleaning robot for domestic environment," 09 2018.

[5] Microsoft, "Phi-3.5-vision-instruct model card," 2024.

[6] Z. Yan, S. Schreiberhuber, G. Halmetschlager, T. Duckett, M. Vincze, and N. Bellotto, "Robot perception of static and dynamic objects with an autonomous floor scrubber," 2020.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[8] S. Gharatkar, "Smart clean dataset 1," 2025. [Online]. Available: https://huggingface.co/datasets/Shanky123/SmartClean1

[9] ——, "Smart clean dataset 2," 2025. [Online]. Available: https://huggingface.co/datasets/Shanky123/SmartClean2

[10] S. Khan, "Smart clean dataset 3," 2026. [Online]. Available: https://huggingface.co/datasets/Suhelkhan/SmartClean3

[11] H. L. et al., "Llava: Large language and vision assistant," 2024.