# Towards Sensor-Free Emission Monitoring for SMEs With Vision-Language Appliance Detection and Operational Context Inference

Muhammad Azeem[1] and Yining Hu[2]

[1] CyberGen
[2] University of Technology Sydney,
muhammad.azem03@gmail.com, Yining.Hu@uts.edu.au,

**Abstract.** Emission reporting has become increasingly compulsory for enterprises around the globe, including Small and Medium Enterprises (SMEs). Available tools–typically leveraging costly sensors, manual audits, and complex software–remain inaccessible to businesses with limited resources. This calls for lightweight, low-cost, and easily accessible technologies to help users monitor their day-to-day carbon footprints and identify inefficiencies.

In this preliminary study, we explore the feasibility of using Vision-Language Models (VLMs) for detecting appliances and inferencing their operational context in coffee shop environments using only images. We finetuned a YOLOv8s model using a synthesized dataset containing 169 images each capturing common coffee shop appliances such as refrigerators, coffee machines, etc, to perform appliance detection, achieving an overall Mean Average Precision at Intersection over Union (IoU)=0.50 (mAP50) of 0.865 and Mean Average Precision at IoU=0.95 (mAP50-95) of 0.623 on our test set with the strongest performance observed for coffee machines and toasters, reaching an mAP50 of 0.995 and 0.994 respectively. To infer an appliance' operational context, we used prompt-based semantic reasoning with a pre-trained transformer Contrastive Language Image Pre-training (CLIP) ViT-B/32, obtaining plausible similarity scores for coffee machines, dishwashers placed next to heat sources, and lighting on during daytime between 0.56 and 0.99. Overall, this vision-language framework demonstrates a feasible first step toward a low-cost, sensor-free, and easily-accessible solution that helps SMEs with emission reporting and identifying effective emission reduction measures.

**Keywords:** Vision-Language Models (VLMs), Appliance Detection, Zero-Shot Inference, SME Carbon Footprints, YOLOv8, Contrastive Language Image Pre-training (CLIP)

## 1 Introduction

Climate change is one of the most critical challenges of our time. Australia officially started enforcing mandatory emission reporting for large entities in 2025, requiring them to disclose Scope 1, 2, and 3 emissions annually [1]. Despite being

currently exempted from direct reporting obligations, SMEs may still face indirect reporting obligations to fullfill the Scope 3 reporting requirements of their larger enterprise partners [2]. In Australia, more than 2.5 million SMEs employ over 8 million people. The food service sector, comprised of coffee shops, restaurants, and small hospitality businesses, is energy-intensive, with dominance in Heating, Ventilation, and Air Conditioning (HVAC) systems (40–50%), refrigeration (20–30%), and cooking equipment (15–25%) [3].

Despite their aggregate impact, SMEs face significant barriers to monitoring and reporting their carbon footprints with 77% citing financial constraints, 63% citing knowledge gaps, and 59% reporting organizational inertia as barriers to energy management adoption [4, 5]. Half of SMEs have never been through energy audits, resulting in no baseline visibility into their consumption patterns [6]. Utility bills only contains total consumptions of resources, with no details on appliance-level usage, making it impossible to identify the precise sources in case of significant consumption. Traditional energy assessment and audits undertaken by external consultants are expensive and provide only a static snapshot [6]. Comprehensive energy monitoring requires data to be continuously collected and stored from multiple sources such as sensors, meters, manual records, and expertise to analyze consumption patterns [7].

SMEs need low-cost, sensor-free, easily-accessible means to understand and track their own carbon footprints at the appliance level. Deep Learning (DL)-based Computer Vision (CV) models, particularly You Only Look Once (YOLO), reached >90% accuracy in real-time object detection and enabled deployment on resource-constrained edge devices [8]. VLMs like CLIP, which are trained on 400 million image–caption pairs, achieve zero-shot classification by matching visual features to natural language descriptions [9]. These capabilities suggest the possibility to use images and videos captured by mobile phones or off-the-shelf cameras to automatically detect appliances, their operational context, and monitor their usage [10, 11].

This preliminary study investigates the feasibility of using lightweight to detect appliance and VLMs to infer their operational context in realistic, visually complex settings. We trained a YOLOv8s object detector on a custom dataset of 169 images each containing five common coffee shop appliances, i.e., customer-facing refrigerators, coffee machines, dishwashers, toasters, and ceiling-mounted lights. The model achieved an overall mAP50 of 0.865 and an mAP50-95 of 0.623 across all appliance categories. We also used a pre-trained CLIP ViT-B/32 VLM to infer operational context associated with potential inefficiencies, such as appliance placement, lighting conditions, and context. CLIP generated plausible similarity scores for these visual cues, which surface contextual factors informative for downstream energy diagnostics. More specifically, this paper makes the following major contributions:

1. Validates appliance detection accuracy using fine-tuned YOLOv8s in visually complex coffee shop environments.

2. Demonstrates CLIP-based semantic reasoning can identify visual factors associated with placement in direct sunlight, daytime lighting, and proximity to heat sources with similarity scores between 0.56 and 0.99.
3. Implements a low-cost, sensor-free prototype combining the fine-tuned YOLOv8s model for appliance detection and a CLIP VLM for context inference, which processes 640×640-pixel smartphone images in under two seconds, enabling daily monitoring without specialized equipment.

The rest of this paper is organized as follows. Section 2 presents the proposed Vision-Language appliance detection and operational context inference pipeline, and dataset construction and annotation. Section 3 presents results of appliance detection and operational context inference on synthesised images and qualitative analysis on real-world images. Section 4 reviews related work and justify the unique position of this study. Finally, Section 5 concludes the paper and indicates future directions.

## 2 Methodology

### 2.1 Vision-Language Framework

The proposed framework consists of four stages as is shown in Figure 1. 1) Detection: The fine-tuned YOLOv8s model processes the input image with a confidence threshold of 0.4, returning bounding boxes for detected appliances; 2) Region Extraction: For each detection, an extended region is extracted by enlarging the original bounding box by 1.5× to capture surrounding contextual information; 3) Context Inference: The cropped region is encoded using CLIP's image encoder, and cosine similarity is computed between the resulting visual embedding and the text embeddings of each context prompt [9]; 4) Context Assignment: The similarity scores are passed through softmax normalization to produce probability-like confidence measures ranging from 0 to 1. The operational context with the highest normalized score is assigned to the detected appliance. These confidence measures represent the model's relative certainty across candidate contexts—higher scores indicate stronger alignment between the visual content and the corresponding textual description. Given the absence of ground-truth operational context labels, we present a qualitative analysis to evaluate the plausibility and consistency of the model's semantic reasoning.

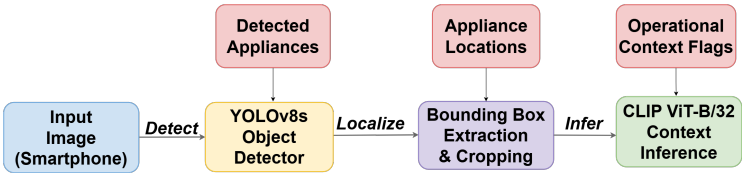


Fig. 1: Vision-Language Appliance Detection and Operational Context Inference.

### 2.2 Dataset Construction and Annotation

We curated a custom dataset of 169 images each containing five key coffee shop appliances including commercial espresso and brewing equipment or *coffee machines*, *customer-facing refrigerators* for displaying beverages and food items, commercial dishwashing units or *dishwashers*, countertop *toasters*, and ceiling-mounted *light fixtures*. Each image contains a varying subset of these classes, with the majority featuring all five categories to maximize annotation density. forming typical, complex coffee shop interiors with different layouts and configurations such as lighting conditions, camera perspectives (14mm ultra-wide to 85mm telephoto), interior design styles (minimalist Scandinavian, rustic wood-and-stone, industrial stainless-steel, modern clinical), and varied appliance spatial arrangements.

All 169 images are synthetic, generated using Google Gemini. Prompts specified appliance types, materials, angles, and lighting, e.g., "An ultra-wide 14mm photorealistic industrial coffee shop scene at sunrise showing six fully separated appliances..." This approach addressed the difficulty of capturing all five appliance classes in real-world single frames. The dataset was partitioned into a training set of 118 images (70%), a validation set of 34 images (20%), and a test set of 17 images (10%). All images were manually annotated in YOLO format, where every bounding box is represented by normalized center coordinates of the box, width, and height. The total number of annotated instances in the curated set was 267 with light fixtures being the most frequent class (83 instances, 31.1%), followed by coffee machines (53 instances, 19.9%), customer-facing refrigerators (52 instances, 19.5%), toasters (41 instances, 15.4%), and dishwashers (38 instances, 14.2%). These annotations were distributed proportionally into the training, validation and testing sets.

### 2.3 Appliance Detection with YOLOv8s

*Model Architecture* We chose YOLOv8s as the object detection backbone as it offers a good trade-off between its detection accuracy and computational efficiency, and is potentially suitable for edge deployment scenarios [10]. The architecture design of YOLOv8s has a total of 11.1 million parameters, 129 layers, and 28.7 GFLOPs. Then, a Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) were used to fuse the multi-scale features [11]. Note that three scale predictions (P3, P4, P5) are used in the detection head to accommodate objects in different sizes. We initialized the model with pre-trained weights from the COCO dataset [12], which consists of 80 object categories in 330,000 images and adapted the final detection layer from 80 classes to 5 classes corresponding to the target appliance categories.

*Training Configuration* The model was trained using the Ultralytics YOLOv8s framework (version 8.3.228)[3] with the configuration in Table 1. Data augmen-

[3] https://github.com/ultralytics/ultralytics

tation was performed at training time to help the model generalize better with the configurations detailed in Table 1.

Table 1: YOLOv8 Training Configuration.

| **Parameters** | | **Data Augmentation** | |
|---|---|---|---|
| **Parameter** | **Value** | **Parameter** | **Configuration** |
| Input image size | 640×640-pixel | Mosaic augmentation | ON for the first 40 epochs, OFF for the last 10 epochs |
| Batch size | 8 | Random Horizontal Flip | Probability: 0.5 |
| Epochs | 50 | HSV color space augmentation | hue: ±0.015, saturation: ±0.7, value: ±0.4 |
| Optimizer | AdamW (auto-selected) | Scale augmentation | ±0.5 |
| Learning rate | 0.001111 (auto-determined) | Translation | ±0.1 |
| Momentum | 0.9 | Random erasing | probability of 0.4 |
| Weight decay | 0.0005 | | |
| Warmup epochs | 3 | | |
| IoU threshold | 0.7 | | |
| Confidence threshold | 0.25 (inference) | | |

*Evaluation Metrics* We report the standard object detection metrics as illustrated in Table 2 for the appliance detection performance [13]. The model that achieved the highest validation mAP50 was saved as the best checkpoint to proceed with further inference.

Table 2: Evaluation Metrics for Object Detection.

| **Metric** | **Description** |
|---|---|
| Precision (P) | Ratio of true positive detections to all positive predictions. |
| Recall (R) | Ratio of true positive detections to all ground truth objects. |
| mAP50 | Mean Average Precision at an IoU threshold of 0.50. |
| mAP50–95 | Mean Average Precision averaged across IoU thresholds from 0.50 to 0.95 in steps of 0.05, providing a more stringent evaluation of localization accuracy. |

### 2.4 Operational Context Inference with CLIP

*VLM Selection* For operational context inference, we selected CLIP ViT-B/32 [9], a VLM that supports zero-shot classification through vision-language alignment. CLIP was chosen for its ability to reason semantically without fine-tuning for each task, enabling flexible prompt-based queries about operational contexts that are difficult to capture with classical approaches [9], suitable for making inferences on various operational contexts without labeling training data in each scenario.

*Context-Aware Prompt Design* For each appliance class, we developed context-specific text prompts to capture following best practices for prompt engineering in vision–language models [14]: 1) Factors of environmental placement: whether appliances are installed in direct sunlight or shaded areas, affecting thermal load and, thereby, energy consumption; 2) Proximity to heat sources: detection of appliances placed near ovens or other heat-generating equipment, possibly increasing cooling requirements; 3) Lighting conditions: identifying artificial lighting operational during daylight hours, signifying possible energy waste. Given an individual detected appliance, CLIP calculates the similarity score between the cropped image region and a set of descriptive text prompts. Table 3 enumerates the complete set of context prompts for each appliance class.

Table 3: Context Prompts for Operational Inference.

| **Appliance Class** | **Context Prompts** |
|---|---|
| **Coffee Machine** | 1. coffee machine in use; 2. coffee machine in idle; 3. coffee machine in sunlight; 4. coffee machine near heat source |
| **Customer Fridge** | 1. fridge in sunlight; 2. fridge in shade; 3. fridge near heat source; 4. fridge door open |
| **Dishwasher** | 1. dishwasher running; 2. dishwasher closed; 3. dishwasher near heat source; 4. dishwasher door open |
| **Light Fixture** | 1. lights on daytime; 2. lights on dark room; 3. lights off; 4. lights near window |
| **Toaster** | 1. toaster in use; 2. toaster idle; 3. toaster in sunlight; 4. toaster near heat source |

## 3 Results and Discussion

*Model Training* Figure 2 demonstrates steady convergence of all loss components, with bounding box loss decreasing from 2.0 to approximately 0.4 and classification loss declining from 2.5 to 0.5 by epoch 50.

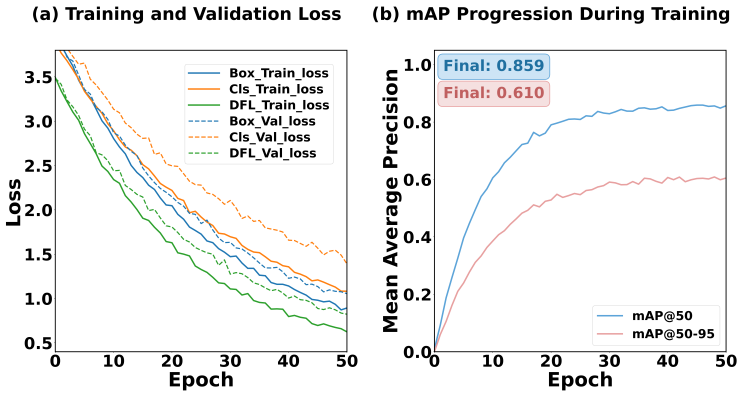


Fig. 2: YOLOv8s Training Progress Over 50 Epochs.

*Appliance Detection:* The fine-tuned YOLOv8s model demonstrated excellent detection performance for all five appliance categories. Figure 3 presents the per-class detection performance of the fine-tuned YOLOv8s model. Our multi-class classifier achieved high true positive rates across appliance categories: coffee machines (0.98), toasters (0.97), dishwashers (0.89), and customer fridges (0.85). Dishwashers were sometimes confused with customer fridges (4% misclassification), likely due to similar stainless-steel surfaces and rectangular geometries. Light fixtures showed the highest misclassification rate (0.60) correct and 0.30 misclassified as background, likely due to the inherent difficulties of detecting ceiling-mounted fixtures against variable architectural backgrounds.

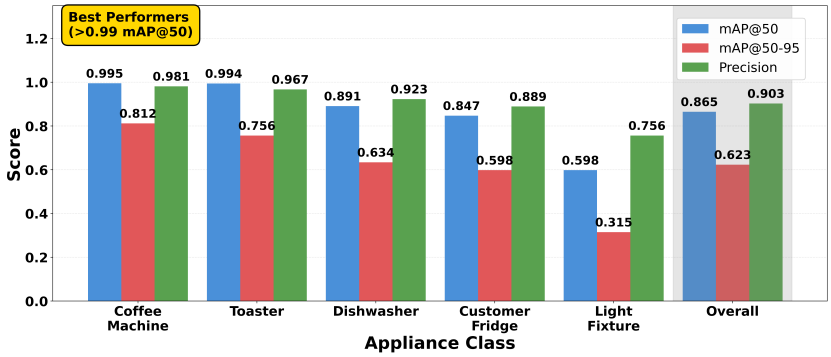


Fig. 3: YOLOv8s Detection Performance by Appliance Class.

*Operational Context Inference:* To assess the zero-shot semantic reasoning of CLIP, the similarity scores were evaluated across four types of operational context: daytime lighting, sunlight exposure, heat proximity, and active usage state. Figure 4 illustrates the distribution of CLIP similarity scores and their relevance across different appliance-context combinations.

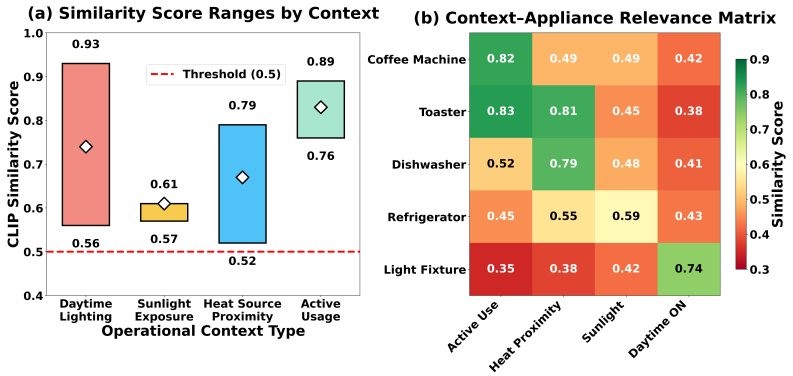


Fig. 4: CLIP Similarity Scores for Operational Context Inference.

The relatively high similarity scores for daytime lighting conditions (median: 0.75) and active usage inference (median: 0.83) suggest that this model effectively captures visual cues indicating energy-relevant conditions. The lower performance on sunlight exposure (median: 0.59) may reflect challenges in distinguishing direct sunlight from other bright indoor lighting conditions, especially when sunlight is diffused through windows or reflected off surfaces.

*VLM Accuracy Analysis:* Table 4 summarizes CLIP similarity scores across appliance-context combinations evaluated on real-world images. Confidence levels are categorized as: high ($\geq$0.70), medium (0.60–0.69), and low ($<$0.60). Coffee machines achieved the highest scores (0.81–0.99) due to distinctive visual cues during active brewing. Dishwashers, toasters, and refrigerators demonstrated reliable medium-to-high confidence (0.58–0.74) for operational state and heat proximity detection. Light fixtures achieved scores up to 0.70 for lighting condition inference, successfully identifying energy-relevant contexts such as artificial lighting during daytime. Overall, CLIP demonstrates sufficient accuracy across all appliance categories to flag actionable energy inefficiencies, validating its suitability for sensor-free operational context inference in SME environments.

Table 4: CLIP ViT-B/32 Similarity Scores by Appliance and Context.

| **Appliance** | **Context** | **Range** | **Confidence** |
|---|---|---|---|
| Coffee Machine | In use / Idle | 0.81–0.99 | High |
| Toaster | In use / Idle | 0.58–0.73 | Medium–High |
| Dishwasher | Running / Heat proximity | 0.63–0.73 | Medium–High |
| Customer Fridge | Heat source proximity | 0.64–0.74 | Medium–High |
| Light Fixture | Daytime / Dark room | 0.62–0.70 | Medium–High |

*Qualitative Analysis Across Diverse Environments* Despite training on synthetic data only, the model generalizes well to real-world images. To validate deployment feasibility, we implemented the proposed framework as a web application that processes smartphone imagery and outputs appliance detections with inferred operational contexts and energy consumption estimates; the source code is publicly available.[4] Figure 5 presents the framework applied to two real-world coffee shop images not seen during training, where green boxes indicate high-confidence contexts ($\geq$0.70); red boxes indicate medium-confidence contexts (0.60–0.69). In Figure 5a, the model detects all five appliance classes with high confidence: coffee machine (0.84), customer fridge (0.74), toaster (0.71), dishwasher (0.80), and light fixtures (0.71–0.73). CLIP identifies operational contexts consistent with potential energy-waste conditions, such as lights being on during daytime and appliances in active use. Figure 5b shows performance in a rustic coffee shop environment characterized by complex lighting and architectural elements. Notably, the model identifies a customer fridge positioned near a heat source (conf: 0.76), demonstrating the framework's capacity to flag placement issues relevant to energy efficiency. These results indicate that despite training exclusively on synthetic data, the model generalizes effectively

[4] https://github.com/Muhammadazeem-eng/Vision_Appliance_Detection_APP

to real-world environments with diverse lighting conditions, visual clutter, and appliance configurations.


(a) Scene 1 (b) Scene 2

Fig. 5: Performance on Real-World Coffee Shop Images.

*Comparative Evaluation with State-of-the-Art:* Table 5 benchmarks our YOLOv8s model against prominent architectures used in energy and appliance monitoring. Our method achieves an $mAP_{50}$ of 86.5%, surpassing the 80.8% equipment usage accuracy reported by Tien et al. [15] using Faster R-CNN. While Faster R-CNN is robust, its two-stage pipeline introduces significant latency, making it impractical for real-time SME applications compared to the single-stage YOLO framework. Furthermore, compared to earlier YOLO generations (e.g., v5/v7) which rely on anchor boxes, our model leverages the anchor-free head in YOLOv8s, enhancing detection for smaller objects like toasters and light fixtures, as highlighted in recent YOLO reviews [16]. YOLOv8s provides an optimal trade-off for detecting objects at multiple scales within the same scene—ranging from large refrigerators to small toasters and light fixtures—which is critical for real-time monitoring in resource-constrained SME environments.

Table 5: Comparison with State-of-the-Art Models in Appliance Detection.

| **Model** | **Performance (%)** | **Architecture** | **Ref.** |
|---|---|---|---|
| Faster R-CNN | 80.8% | Two-stage | [15] |
| SSD | ∼78.0% | Single-stage | [17] |
| YOLOv5/v7 | ∼85.0% | Single-stage | [16] |
| **Proposed YOLOv8s** | **86.5%** | **Single-stage** | **Ours** |

YOLOv8s thus achieves the best balance when it comes to detecting objects within a single scene that range from large refrigerators to small toasters or light fixtures. This makes it a preferable option, compared to substantial computational latency with newer variants, for resource-constrained SME environments.

## 4 Related Work

*Carbon Footprints Monitoring for SMEs* Non-Intrusive Load Monitoring (NILM) has emerged as a prominent approach for appliance-level energy disaggregation. Liu et al. [7] combined multi-head self-attention with Gated Recurrent Units (GRU) for minute-level corporate carbon estimation, though requiring high-frequency electrical data unavailable to typical SMEs. Ayub and El-Alfy [18] developed Vision Transformer (ViT)-based appliance identification achieving F1 scores exceeding 97%, but relying on smart meter data. Zhao et al. [19] introduced a ViT with detection head for unknown appliances, while Tang et al. [20] proposed semi-supervised learning with pseudo-labels to reduce labeled data dependence. Athanasiadis et al. [21] developed lightweight Convolutional Neural Network (CNN) classifiers ( 54K parameters) for edge NILM. These approaches require smart meter infrastructure, presenting barriers for financially constrained SMEs [3, 4]. Building Energy Management Systems (BEMS) offer alternatives. Zhang et al. [22] proposed Large Language Model (LLM) Agent-based BEMS achieving 29.4% improvement in user comprehension via Reflection-Chain-of-Thought (RCoT) prompting, but assumes existing sensors. Papaioannou et al. [23] presented EnergiQ, an LLM-driven platform with Internet of Things (IoT) sensing achieving 94% accuracy and 91% expert agreement, yet requiring smart plug deployment. Computer vision has shown promise for energy management. Tien et al. [15] achieved 97.32% occupancy and 80.80% equipment detection accuracy using Faster Region-based Convolutional Neural Network (Faster R-CNN), demonstrating 65.75% equipment heat gain reductions; however, their work targets controlled office settings.

*Appliance Detection and Operational Context Inference* YOLO variants have achieved prominence for real-time detection. Kim et al. [8] demonstrated over 90% accuracy with edge deployment via concurrent multi-frame processing. Zheng et al. [24] proposed Ghost-Enhanced Bi-directional YOLO (GEB-YOLO) integrating GhostConv to reduce computation. Al Rajab et al. [25] developed EnergySense combining YOLO with augmented reality for residential optimization. Cao et al. [26] adapted YOLOv8 for photovoltaic defect detection, and Hussain and Khanam [16] reviewed YOLOv1-v10 improvements. YOLOv8s provides suitable balance between performance and speed for consumer-grade hardware. VLMs are powerful tools for visual understanding. CLIP, trained on 400 million image-caption pairs [9], enables zero-shot classification and is effective for semantic reasoning. Kalivarathan et al. [14] proposed Intelligence of Things (INOT) using Owl-ViT 2 for spatial context-aware device control. Despite these advances, VLM application to operational context inference for emissions estimation remains unexplored. Current works target anomaly detection [27], device control or general understanding [14]—none infer energy-relevant parameters like placement affecting thermal efficiency. Wen et al. [28] demonstrated visual processing for industrial assistants in manufacturing. Smart home literature [29, 30] shows vision-based recognition feasibility without emissions-relevant inference.

A significant gap remains as no existing solution combines visual-language reasoning with appliance detection for emissions estimation in SMEs' day-to-day operations. Our approach differs by: 1) targeting SME environments; 2) combining detection with semantic context inference for emissions; 3) requiring only smartphone imagery; and 4) leveraging pre-trained models (YOLOv8s, CLIP ViT-B/32).

## 5 Concluding Remarks

This paper has presented a vision-language framework for appliance detection and operational context inference in SME settings, using a coffee shop as an example. We demonstrated that a fine-tuned YOLOv8s model achieves robust detection performance for five classes of appliances, resulting in an $mAP_{50}$ of 0.865 and $mAP_{50–95}$ of 0.623, with particularly high accuracy for detecting coffee machines ($mAP_{50} = 0.995$) and toasters ($mAP_{50} = 0.994$). We also established that CLIP ViT-B/32 can perform plausible inferences of operational context through prompt-based semantic reasoning, yielding similarity scores from 0.56 to 0.99 for appliances located close to heat sources, refrigerators exposed to direct sunlight, and artificial lighting during daytime. The full pipeline operates using accessible smartphone imagery without specialized sensors, sub-metering infrastructure, or manual audits, addressing commonly reported SME barriers to energy-management adoption. These findings demonstrate the technical feasibility of a low-cost, sensor-free, and accessible solution for achieving appliance-level usage monitoring in visually complex environments.

Several directions for future research remain for this proof-of-concept to be deployed to support SMEs in understanding and reporting their own carbon footprints. Firstly, integrating appliance-level energy consumption data with visual outputs would enable quantitative carbon-footprint estimation rather than purely qualitative inference. Secondly, empirical validation of the energy impact associated with visual context cues identified by CLIP (such as sunlight exposure and proximity to heat sources) would require longitudinal studies correlating imagery with measured consumption. Thirdly, extending the framework to additional SME domains beyond coffee shops—such as retail stores, restaurants, and small offices—and incorporating video-based temporal analysis would improve generalizability and real-world applicability.

## References

1. AccountantsDaily. Indirect climate reporting requirements to affect small businesses: Asic, 2025. [Accessed: Mar. 2025].
2. Australian Securities and Investments Commission. Regulatory guide 280: Sustainability reporting, March 2025.
3. J. Á. Jaramillo, J. W. Z. Sossa, and G. L. O. Mendoza. Barriers to sustainability for small and medium enterprises in the framework of sustainable development—literature review. *Business Strategy and the Environment*, 28(4):512–524, 2019.

4. F. P. Privat and D. C. Guerrieri. Energy efficiency in small and medium-sized enterprises: a literature review approach. *Revista de Gestão – RGSA*, 18(11):e09687, 2024.
5. R. Agrawal, L. De Tommasi, P. Lyons, S. Zanoni, G. K. Papagiannis, C. Karakosta, et al. Challenges and opportunities for improving energy efficiency in smes: learnings from seven european projects. *Energy Efficiency*, 16, 2023.
6. Joanna Southernwood, Grigoris K. Papagiannis, Erudino Llano Güemes, and Luisa Sileni. Energy efficiency solutions for small and medium-sized enterprises. *Proceedings*, 65(1), 2020.
7. G. Liu, J. Liu, J. Zhao, J. Qiu, Y. Mao, Z. Wu, and F. Wen. Real-time corporate carbon footprint estimation methodology based on appliance identification. *IEEE Transactions on Industrial Informatics*, 19(2):1401–1412, 2023.
8. S. Kim, C. Kim, and J. Kim. Improving performance of real-time object detection in edge device through concurrent multi-frame processing. *IEEE Access*, 13:1522–1533, 2025.
9. 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. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 8748–8763. PMLR, 2021.
10. G. Jocher, A. Chaurasia, and J. Qiu. Ultralytics yolov8. version 8.0.0, 2023. [Online; accessed 2025].
11. S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8759–8768, 2018.
12. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, Zurich, Switzerland, 2014.
13. J. Zhang, J. Huang, S. Jin, and S. Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, 2024.
14. Sukanth Kalivarathan, Muhmmad Abrar Raja Mohamed, Aswathy Ravikumar, and S Harini. Intelligence of things: A spatial context-aware control system for smart devices, 2025.
15. P. W. Tien, S. Wei, and J. Calautit. A computer vision-based occupancy and equipment usage detection approach for reducing building energy demand. *Energies*, 14(1):156, 2021.
16. M. Hussain and R. Khanam. In-depth review of yolov1 to yolov10 variants for enhanced photovoltaic defect detection. *Solar*, 4(3):351–386, 2024.
17. Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, volume 9905, pages 21–37. Springer, 2016.
18. M. Ayub and E.-S. M. El-Alfy. Household appliance identification using vision transformers and multimodal data fusion. *IEEE Transactions on Consumer Electronics*, 71(2):2774–2792, 2025.
19. Q. Zhao, W. Liu, K. Li, Y. Wei, and Y. Han. Unknown appliances detection for non-intrusive load monitoring based on vision transformer with an additional detection head. *Heliyon*, 10(9):e30666, 2024.
20. T. Tang, K. Li, C. Su, and Z. Liu. Semi-supervised learning with flexible threshold for non-intrusive load monitoring. *Heliyon*, 10(14):e34457, 2024.

21. C. Athanasiadis, D. Doukas, T. Papadopoulos, and A. Chrysopoulos. A scalable real-time non-intrusive load monitoring system for the estimation of household appliance power consumption. *Energies*, 14(3):767, 2021.
22. S. Zhang, H. Zhao, J. Wei, Y. Xie, D. Xiang, X. Cai, Y. Cheng, and J. Zhao. Building energy management systems with llm agent enhanced natural language policy explanation. In *Proceedings of the 2025 5th Power System and Green Energy Conference (PSGEC)*, pages 409–414, 2025.
23. C. Papaioannou, I. Tzitzios, A. Papaioannou, A. Dimara, C.-N. Anagnostopoulos, and S. Krinidis. Energiq: A prescriptive large language model-driven intelligent platform for interpreting appliance energy consumption patterns. *Sensors*, 25(16):4911, 2025.
24. J. Zheng, H. Liu, Q. He, et al. Geb-yolo: A novel algorithm for enhanced and efficient detection of foreign objects in power transmission lines. *Scientific Reports*, 14(1):15769, 2024.
25. M. Al Rajab and S. Loucif. Sustainable energysense: a predictive machine learning framework for optimizing residential electricity consumption. *Discover Sustainability*, 5:74, 2024.
26. Y. Cao, D. Pang, Q. Zhao, Y. Yan, Y. Jiang, C. Tian, F. Wang, and J. Li. Improved yolov8-gd deep learning model for defect detection in electroluminescence images of solar photovoltaic modules. *Engineering Applications of Artificial Intelligence*, 131:107866, 2024.
27. Yuxuan Cai, Xinwei He, Dingkang Liang, Ao Tong, and Xiang Bai. Anomaly detection by adapting a pre-trained vision language model, 2024.
28. D. Wen, J. Zheng, R. Liu, Y. Xu, K. Peng, and R. Stiefelhagen. Snap, segment, deploy: A visual data and detection pipeline for wearable industrial assistants, 2025.
29. Climate Change Authority. 2023 review of the carbon credits (carbon farming initiative) act 2011, December 2023.
30. A. Macintosh, D. Butler, P. Larraondo, M. C. Evans, D. Ansell, M. Waschka, R. Fensham, D. Eldridge, D. Lindenmayer, P. Gibbons, and P. Summerfield. Australian human-induced native forest regeneration carbon offset projects have limited impact on changes in woody vegetation cover and carbon removals. *Communications Earth & Environment*, 5, 2024.