Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

Life Sciences und
Facility Management

Don’t waste Compute
Efficient AI Models for Waste Management

Masterthesis by Helene Benkert • Applied Computational Life Sciences • Spring Semester 2025
Supervisor: Dr. Martin Schüle

Introduction

The correct separation of household waste is a cornerstone of efficient recycling and sustainable resource management. However, for consumers, this task often presents a significant challenge due to complex, regional regulations.

Objective of this Thesis: To develop and evaluate modern AI models for creating a user-friendly, automated sorting assistant that provides precise, location-specific instructions.

Research Question

How do different AI paradigms (generative, discriminative, and detection models) compare in their effectiveness at classifying household waste and providing accurate recycling guidance?

Methodology

1. **Dataset:** Creation of a custom dataset with >5,800 images, reflecting the recycling rules of Lower Austria. 14 Material classes and 11 Disposal classes.
2. **Models:** Training and evaluation of four architectures:
   - CuMo: Generative Multimodal Model
   - ViLT: Discriminative Multimodal Model
   - ViT: Unimodal Image Classifier
   - YOLO11m: Object Detector
3. **Training:** Comparison of specialized (material-only/disposal-only) vs. combined training strategies.

Results

All models achieved excellent performance with F1-scores and mAP values ≥ 0.90.

- **CuMo (Generative):** Functionally superior. Generates complete, human-readable instructions.
- **ViT/ViLT (Discriminative):** Very high classification precision but limited to a predefined set of labels.
- **YOLOv11m (Detection):** Excellent at localizing objects with clear shapes (e.g., batteries), but weaker on amorphous materials (e.g., organic waste).

Experiment: AI vs. Human

A comparative study with 10 test items revealed:

- **Finetuned CuMo**: 10/10 correct answers
- **Humans (Local Residents):** 7/10 correct answers
- **ChatGPT-4:** 7/10 correct answers
- **Gemini 2.5 Pro:** 3/10 correct answers

Key Insight: Humans and general-purpose AIs fail on ambiguous cases (e.g., broken glass, aluminum foil). A specialized, fine-tuned model significantly surpasses average human knowledge and general AI performance.

Difficult Classes & Model Strengths

The Battery: A Key Example

- **ViT/ViLT (Classifiers):** Struggled significantly (low recall). Theory: Their global view of the image "overlooked" the small but critical features of the battery.
- **YOLO (Detector):** Perfect score. Its architecture is optimized for finding local features and ignoring the background.
- **CuMo (Generative):** Also a perfect score. Theory: Despite a ViT-based encoder, this points to the power of its CLIP pre-training and the downstream language model.

Composite Materials (e.g., Tetra Pak)

Often confused by ViT/ViLT with their constituent materials (cardboard, aluminum, plastic), highlighting the limits of purely visual texture recognition.

Conclusion

The targeted fine-tuning of a generative multimodal model (CuMo) is the most effective strategy for creating a reliable, practical expert system. It bridges the gap from simple recognition to situational problem-solving.

Proof of Concept

A functional prototype was implemented as a web app on Hugging Face Spaces. Users can upload an image and receive a direct, clear sorting instruction. This demonstrates the technical feasibility of the application.

Shown below are some example pictures, taken from the Dataset:

Composite Carton


White Glass & Metal


Rigid Plastic Container


Residual Waste


Plastic & Aluminium

Key References:
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv. https://arxiv.org/abs/2010.11929
Kim, W., Son, B., & Kim, I. (2021). *ViLT: Vision-and-language transformer without convolution or region supervision.* arXiv. https://arxiv.org/abs/2102.03334
Li, J., Wang, X., Zhu, S., Kuo, C.-W., Xu, L., Chen, F., Jain, J., Shi, H., & Wen, L. (2024). *CuMo: Scaling multimodal LLM with Co-Upcycled Mixture-of-Experts.* arXiv. https://arxiv.org/abs/2405.05949
Ultralytics YOLOv11 (Version 11.0.0) [Software]. Ultralytics. https://github.com/ultralytics/ultralytics