Computer Science and Application 计算机科学与应用, 2026, 16(7), 125-139
Published Online July 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.167247

Hans汉斯

面向形态多变垃圾的轻量化视觉语言模型细粒度分类方法研究

孙紫阳，刘　柱，王向前，李　娜，熊旭辉*

湖北师范大学人工智能与计算机学院，湖北　黄石

收稿日期：2026年6月19日；录用日期：2026年7月17日；发布日期：2026年7月24日

摘　要

针对垃圾分类场景下垃圾形态多变、细粒度类别特征相近、训练样本稀缺以及边缘部署算力不足等难题，本文基于轻量化视觉语言模型SmolVLM-256，开展小样本细粒度垃圾分类的性能优化研究。结合真实场景中垃圾挤压、遮挡、复杂背景干扰等特点，本文构建了包含8类生活垃圾的细粒度数据集并搭建标准化实验评估框架。通过超参数调优、语义提示引导、数据增强与分类头结构改良等多维度优化实验，探究各策略对模型分类效果的影响。实验结果显示，优化后的SmolVLM-256分类准确率可达87.92% ± 1.56%，相较基线模型性能大幅提升。横向对比ViT、CLIP、BLIP、Qwen3-VL等模型可知，该轻量化模型在保障优良分类精度的同时，具备参数量小、推理成本低的优势，更适配边缘设备部署。研究表明，合理的训练与结构优化策略，能够有效提升轻量化视觉语言模型在复杂小样本垃圾场景的识别能力，可为边缘智能垃圾分类系统的研发与落地提供技术参考。

关键词

轻量化视觉语言模型，垃圾细粒度分类，形态多变垃圾，小样本学习，边缘部署

Research on Fine-Grained Classification Method of Lightweight Vision-Language Model for Morphologically Variable Waste

Ziyang Sun, Zhu Liu, Xiangqian Wang, Na Li, Xuhui Xiong*

School of Artificial Intelligence and Computer, Hubei Normal University, Huangshi Hubei

Received: June 19, 2026; accepted: July 17, 2026; published: July 24, 2026

Abstract

Aiming at the practical challenges of variable garbage morphology, similar fine-grained feature

*通讯作者。

文章引用: 孙紫阳, 刘柱, 王向前, 李娜, 熊旭辉. 面向形态多变垃圾的轻量化视觉语言模型细粒度分类方法研究[J]. 计算机科学与应用, 2026, 16(7): 125-139. DOI: 10.12677/csa.2026.167247

distribution, limited training samples, and insufficient computing resources for edge deployment in garbage classification scenarios, this paper conducts a performance optimization study on small-sample fine-grained garbage classification based on the lightweight vision-language model SmolVLM-256. Considering real-world interferences such as garbage extrusion, occlusion, and complex background environments, a fine-grained dataset containing eight types of domestic garbage is constructed, and a standardized experimental evaluation framework is established. Multiple optimization experiments including hyperparameter tuning, semantic prompt guidance, data augmentation, and classification head structure improvement are carried out to explore the influence of different strategies on model classification performance. The experimental results show that the optimized SmolVLM-256 achieves a classification accuracy of 87.92% ± 1.56%, which obtains a significant performance improvement compared with the baseline model. Comparative experiments with ViT, CLIP, BLIP, and Qwen3-VL demonstrate that the lightweight model maintains competitive classification accuracy with fewer parameters and lower inference cost, making it more suitable for edge deployment. The research verifies that reasonable training strategies and structural optimization can effectively enhance the recognition capability of lightweight vision-language models in complex small-sample garbage scenarios, providing technical references for the design and practical application of edge intelligent garbage classification systems.

Keywords

Lightweight Vision-Language Model, Fine-Grained Waste Classification, Deformable Waste, Few-Shot Learning, Edge Deployment

1. 引言

随着智慧环保与边缘人工智能技术发展，智能垃圾分类成为研究热点[1]。实际场景中垃圾存在折叠、挤压、遮挡等形态变化，且不同品类纹理、外观高度相似，属于典型细粒度、小样本识别任务[2] [3]。传统卷积网络、大型视觉 Transformer 虽具备一定识别能力，但大型视觉语言模型参数量大、推理成本高，难以部署在智能垃圾桶、巡检终端等边缘设备[4]。

轻量化视觉语言模型兼顾跨模态语义理解与低算力消耗，是解决上述问题的有效方向。目前 SmolVLM 这类轻量化多模态模型在垃圾细粒度分类任务中的超参数适配、提示词机制、数据增强策略等规律仍缺乏系统探究[5] [6]。

基于此，本文选取 SmolVLM-256 作为核心模型，自建真实场景垃圾数据集，从训练策略、网络结构、多模态融合等角度完成优化实验，同时与多款主流视觉/多模态模型开展横向对比。本文主要创新点：

1) 构建贴合真实应用的小样本细粒度垃圾数据集[7]；
2) 明确轻量化视觉语言模型在本任务下的最优超参数与优化组合[8]；
3) 验证双层分类头、加权语义融合、空间数据增强的协同增益[9] [10]；
4) 证明 SmolVLM-256 在精度与部署成本之间的平衡优势[4] [5]。

2. 相关技术与理论基础

细粒度图像分类要求模型捕捉类别间微小特征差异，而形态多变垃圾进一步提升了类内特征的学习难度[2]。小样本学习通过迁移学习、数据增强、正则化、提示学习等方法，缓解标注数据不足引发的过

拟合问题[3] [11]。

Vision Transformer (ViT)依靠自注意力机制建模图像全局特征，迁移能力优于传统卷积网络，但边缘部署成本较高[12]-[14]。视觉语言模型(VLM)通过图文预训练实现跨模态语义对齐，可借助文本语义辅助细粒度识别，典型代表包括 CLIP、BLIP、Qwen3-VL 等大型模型[15]-[18]。

SmolVLM-256 为面向边缘场景的轻量化视觉语言模型，采用精简 Transformer 架构与参数共享机制，在保留图文融合能力的同时大幅缩减参数量[5] [19]。分类头负责将高维特征映射至类别空间，多层结构可增强非线性表达，提升相似类别区分能力；数据增强与 Dropout、权重衰减等正则化手段，是小样本任务中提升泛化能力、抑制过拟合的核心方法[11] [20]。

3. 实验方案与方法

3.1. 实验环境配置

实验基于 Ubuntu 22.04 系统、NVIDIA RTX 4060 显卡搭建，采用 PyTorch 2.4.1 深度学习框架，统一软硬件配置保证实验可复现，环境参数如表 1 所示。

Table 1. Experimental environment configuration
表 1. 实验环境配置

配置项	参数
操作系统	Ubuntu 22.04
CPU	Intel Core i7
GPU	NVIDIA RTX 4060
显存	8 GB
CUDA	12.1
Python	3.10
PyTorch	2.4.1
Transformers	4.51.3
Batch Size	8
优化器	AdamW

3.2. 数据集构建

本文自建细粒度垃圾数据集，选取 8 类外观易混淆的生活垃圾，模拟折叠、遮挡、复杂光照等真实形态变化。数据集总计 240 张图像，每类 30 张，按照 8:2 划分为训练集(192 张)与测试集(48 张)，严格保证类别分布均衡。数据集已开源，地址：https://github.com/1583118323/Waste_FineGrained_Dataset [7]。数据集类别如表 2 所示。

Table 2. Categories of datasets
表 2. 数据集类别

类别编号	类别名称
1	aluminum_can
2	newspaper

续表

3	packaging_bag
4	paper_box
5	plastic_bottle
6	plastic_box
7	surgical_mask
8	Toilet_paper

3.3. 数据预处理与增强策略

统一完成图像缩放、张量转换、像素归一化预处理。本文设置随机翻转、旋转、颜色抖动、随机裁剪四类数据增强策略，通过对照实验筛选最优方案[10] [11]。

3.4. 对比模型选择

选取 ViT、CLIP、BLIP、Qwen3-VL 作为对照模型，覆盖纯视觉、经典多模态、大型多模态三类主流架构[12] [15]-[17]。针对 SmolVLM-256，采用控制变量法完成超参数、Prompt 语义融合、分类头、正则化的递进式消融实验[6] [8]；文本与视觉特征采用加权融合策略，设置多组融合权重完成消融测试[9]。

3.5. 评价指标

采用准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1 分数作为评价指标[21]。SmolVLM-256 与 Qwen3-VL 采用五折交叉验证，结果以均值±标准差形式呈现，其余模型采用固定训练/测试集评估。

4. 实验结果与分析

本章采用横向模型对比 + 纵向模块消融的双层实验架构，基于自建小样本细粒度生活垃圾数据集开展测试，统一实验环境与参数，系统分析不同模型及优化策略的分类效果。

4.1. 实验总体设计

本次实验分为两大维度：横向选取 ViT、CLIP、BLIP、Qwen3-VL 四类主流模型完成基准测试，明确不同架构模型的性能差异；纵向以 SmolVLM-256 为核心，围绕超参数、Prompt 融合、数据增强、分类头及正则化开展递进式消融实验，量化各优化模块的贡献。通过多组对照实验，明确各类方法对小样本细粒度分类任务的影响规律，形成可复现的实验结论。

4.2. Vit

4.2.1. 模型概述

Vision Transformer (ViT)是将 Transformer 应用于图像分类的经典模型[12]，通过图像分块与自注意力机制建模全局视觉特征[13]。相较于 ResNet [22]、MobileNets [23]等传统卷积网络，ViT 全局特征建模能力更强。本文选用 ImageNet 预训练的 ViT-B/16 作为纯视觉基准模型，在自建数据集上完成迁移微调，为多模态模型提供对照依据，同时参考基于轻量化卷积网络的垃圾分类相关研究[24]开展横向对比。

4.2.2. 实验结果

1) 预训练模型零样本性能(未微调)

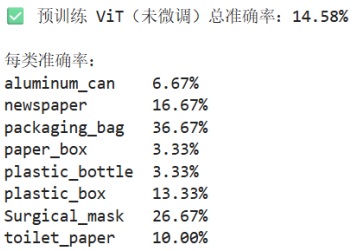

Figure 1. Zero-shot classification report of pre-trained ViT-B/16 model
图 1. 预训练 ViT-B/16 模型零样本分类报告

由图 1 可知，未经过领域微调的 ViT 模型测试准确率仅为 14.58%，仅略高于八分类随机猜测概率。这表明通用图像预训练特征无法适配形态多变的垃圾细粒度识别任务，领域微调十分必要。

2) 小样本微调性能

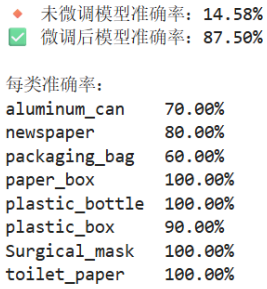

Figure 2. Few-shot fine-tuning classification report of ViT-B/16 model
图 2. ViT-B/16 模型小样本微调分类报告

如图 2 所示，使用少量样本完成微调后，ViT 准确率提升至 87.50%，性能涨幅达 72.92%。证明 ViT 具备优秀的跨域迁移能力，可通过少量标注数据学习垃圾材质、纹理等细粒度特征，也验证了该任务的可学习性。

4.2.3. 结果分析与小节结论

微调后的 ViT 模型分类准确率达 87.50%，充分体现了 Transformer 架构在细粒度视觉任务中的优势[25]。该模型迁移学习能力较强，但参数量偏大，边缘部署能力一般；同时因缺少文本语义辅助，对高度相似垃圾类别的区分效果有限。本实验为后续多模态模型研究提供了可靠基线，也说明轻量化 + 多模态结合是该场景下的重要研究方向。

4.3. BLIP

4.3.1. 模型概述

BLIP 是经典视觉语言预训练模型，依托双编码器实现图文联合表征与跨模态对齐[15]，在细粒度识别任务中优势明显。本文采用官方预训练 BLIP 模型，在自建数据集上完成监督微调。

4.3.2. 实验结果

模型训练前后的分类指标如表 3 所示。随机初始化状态下，模型准确率仅为 10.42%；经过微调后，总体准确率提升至 90.00%。其中医用口罩、塑料瓶等类别识别精度达 100%，仅包装袋因形态多样，F1 分数相对偏低。

Table 3. Comparison of model classification results before and after training
表 3. 训练前后模型分类结果对比

类别	指标	训练前	训练后
Surgical_mask	Precision/Recall/F1	0.00/0.00/0.00	1.00/1.00/1.00
aluminum_can	Precision/Recall/F1	0.00/0.00/0.00	1.00/0.83/0.91
newspaper	Precision/Recall/F1	0.75/0.50/0.60	1.00/0.83/0.91
packaging_bag	Precision/Recall/F1	0.00/0.00/0.00	0.60/1.00/0.75
paper_box	Precision/Recall/F1	0.00/0.00/0.00	1.00/0.67/0.80
plastic_bottle	Precision/Recall/F1	0.00/0.00/0.00	1.00/1.00/1.00
plastic_box	Precision/Recall/F1	0.00/0.00/0.00	1.00/0.83/0.91
toilet_paper	Precision/Recall/F1	0.25/0.33/0.29	0.86/1.00/0.92
-	总体准确率	10.42%	90.00%
-	宏平均	0.12/0.10/0.11	0.93/0.90/0.90

4.3.3. 结果分析与小节结论

实验表明，BLIP 预训练习得的跨模态知识可快速迁移至垃圾分类场景，图文融合特征有效缓解了类别混淆问题[15]。该模型分类性能优异，但参数规模大、推理成本高，难以部署在边缘设备[1]。

4.4. CLIP

4.4.1. 模型概述

CLIP 基于图文对比学习构建统一语义空间[16]，具备强大的零样本推理与跨域迁移能力。本文选用 ViT-B/32 作为其视觉编码器，开展零样本测试与线性探测微调实验。

4.4.2. 实验结果

为全面评估 CLIP 模型在小样本垃圾分类任务上的性能，本文分别进行了 Zero-Shot 零样本测试与小样本微调实验。

1) Zero-Shot 零样本分类性能

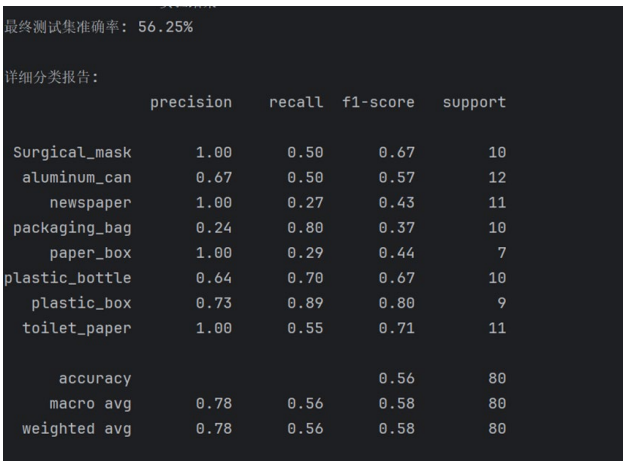

Figure 3. Zero-shot inference classification report of CLIP (ViT-B/32) model
图 3. CLIP (ViT-B/32)模型零样本推理分类报告

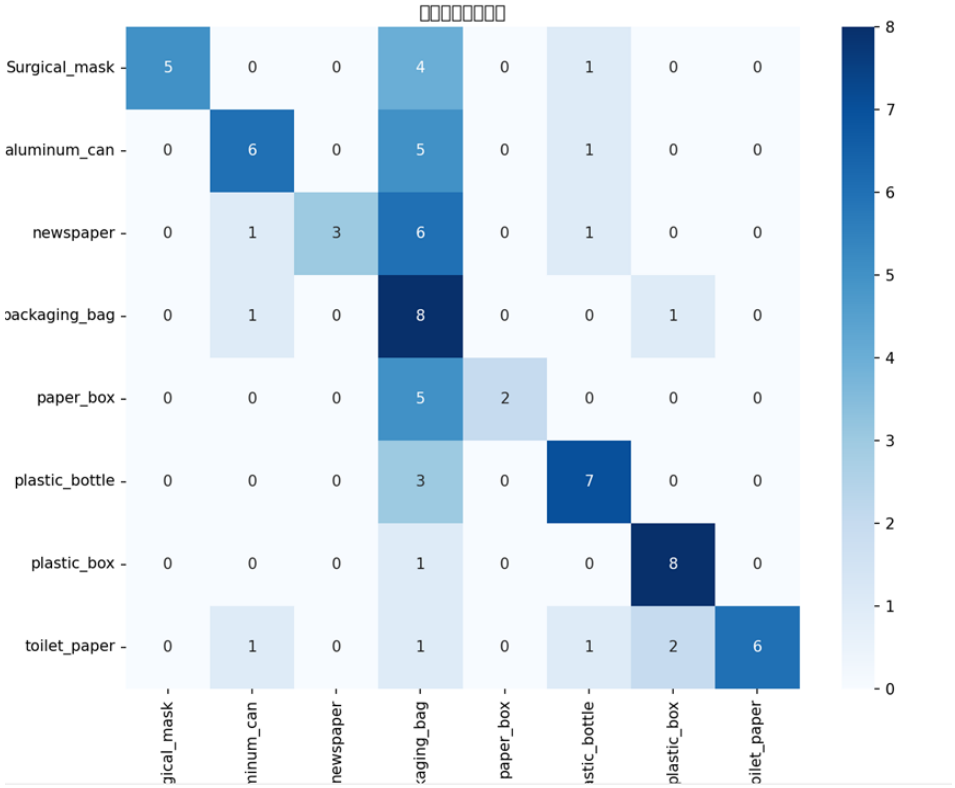

Figure 4. Confusion matrix for zero-shot inference classification of CLIP (ViT-B/32) model
图 4. CLIP (ViT-B/32)模型零样本推理分类混淆矩阵

由图 3、图 4 可知，CLIP 零样本推理准确率为 56.25%，虽优于随机猜测，但报纸与纸盒、塑料瓶与塑料盒等相似类别混淆严重，反映通用语义与垃圾分类领域存在明显偏差。

2) 小样本微调(Linear Probe)性能

详细分类报告:

	precision	recall	f1-score	support
Surgical_mask	0.91	1.00	0.95	10
aluminum_can	1.00	0.90	0.95	10
newspaper	0.67	0.60	0.63	10
packaging_bag	0.80	0.80	0.80	10
paper_box	0.73	0.80	0.76	10
plastic_bottle	0.89	0.80	0.84	10
plastic_box	0.91	1.00	0.95	10
toilet_paper	1.00	1.00	1.00	10
accuracy			0.86	80
macro avg	0.86	0.86	0.86	80
weighted avg	0.86	0.86	0.86	80

Figure 5. Linear probe fine-tuning classification report of CLIP (ViT-B/32) model
图 5. CLIP (ViT-B/32)模型线性探测微调分类报告

完成线性探测微调后，模型准确率提升至 86.00% (见图 5)，各类别识别效果显著改善，证明 CLIP 提取的特征可有效适配本任务。

4.4.3. 结果分析与小节结论

CLIP 在零样本与微调场景下均表现出良好的迁移能力[26]，文本语义可辅助提升细粒度识别效果。但该模型仍属于大型多模态架构，算力需求高，边缘落地难度大，也进一步印证了轻量化改造的必要性。

4.5. Qwen3-VL

4.5.1. 模型概述

Qwen3-VL 是新一代大型视觉语言模型，擅长融合图文特征完成细粒度识别。本文采用 Qwen3-VL-2B 模型，结合 LoRA 高效微调方式开展实验。

4.5.2. 实验结果

1) 五折交叉验证结果

在自建垃圾分类数据集上，Qwen3-VL-2B 经过 LoRA 微调后取得了较优的分类性能。五折交叉验证结果如表 4 所示。

Table 4. 5-Fold cross-validation results of Qwen3-VL
表 4. Qwen3-VL 五折交叉验证结果

Fold	Accuracy (%)
Fold1	95.83
Fold2	100.00
Fold3	95.83
Fold4	93.75
Fold5	95.83
Mean ± Std	96.25 ± 2.04

2) 各类别分类性能

图 6 展示了 Qwen3-VL 模型在各类别上的 Precision、Recall 和 F1-score 评估结果。

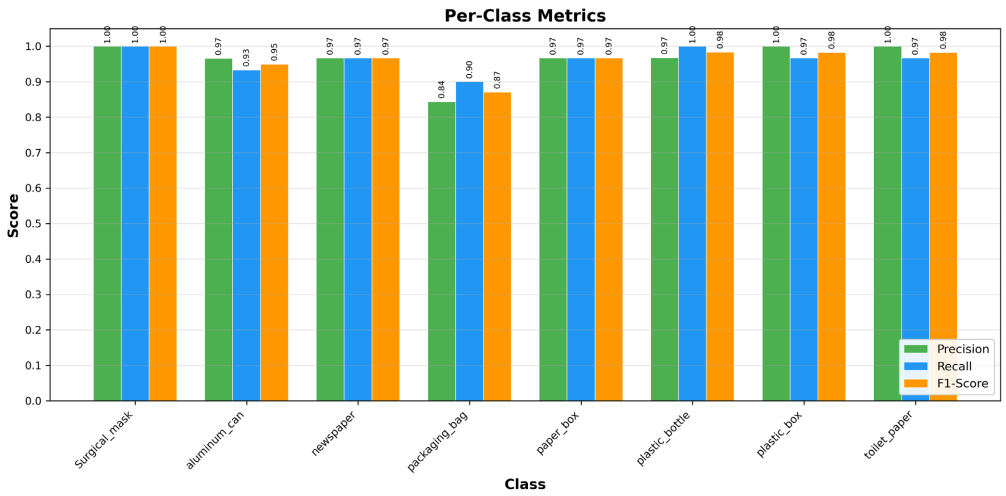

Figure 6. Evaluation of each category
图 6. 各类别评估

图 7 展示了模型在全部验证样本上的混淆矩阵。

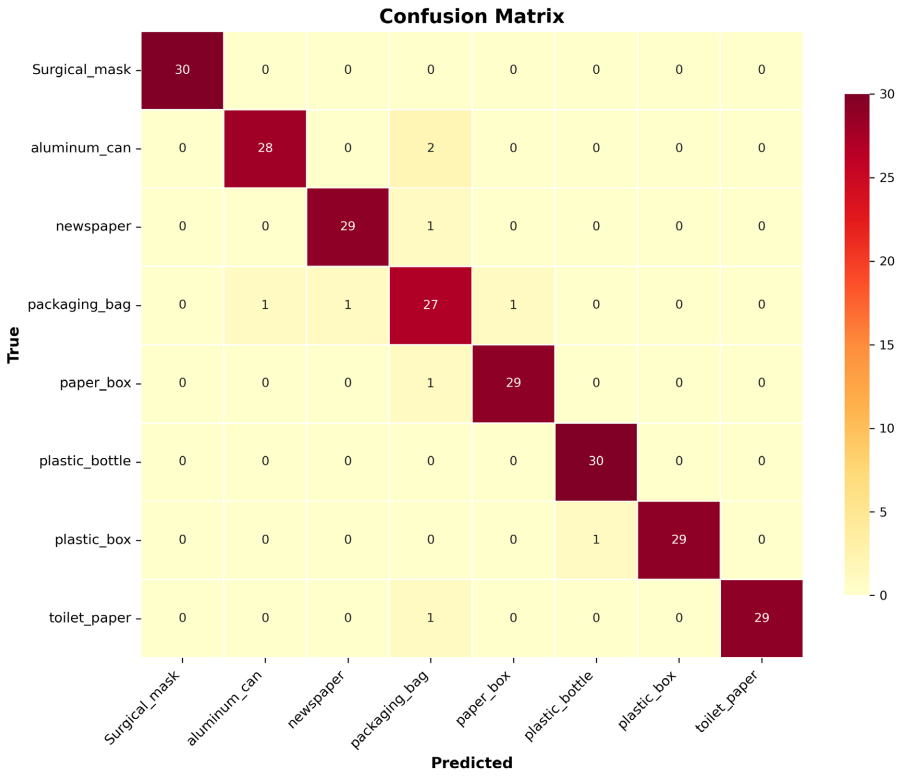

Figure 7. Confusion matrix of each category
图 7. 各类别混淆矩阵

模型五折交叉验证结果如表 6 所示，平均准确率达 96.25%±2.04%，整体稳定性优异。结合图 6、图 7 可知，除包装袋识别效果一般外，其余类别精度均保持较高水平。

4.5.3. 结果分析与小节结论

Qwen3-VL 凭借海量预训练知识，取得所有模型中最优分类精度，细粒度判别能力突出。模型五折实验标准差仅 2.04%，泛化性能良好。但该模型参数量极大，对硬件要求严苛，无法应用于低算力边缘终端[3]，这也是本文研究轻量化模型的核心原因。

4.6. SmolVLM_256

4.6.1. 模型概述

SmolVLM-256 是面向边缘场景的轻量化视觉语言模型[5]，采用精简 Transformer 结构与参数压缩方案[19]，兼顾图文融合能力与低推理开销，是本文的核心研究对象。

4.6.2. 实验结果

1) 超参数优化实验

为探究超参数对 SmolVLM-256 垃圾分类性能的影响，本文采用控制变量法，在固定数据集、模型结构、优化器的前提下，依次分析基础配置、学习率、训练轮数、Dropout 正则化对模型准确率的作用规律，筛选适配垃圾分类小样本任务的最优超参数组合。

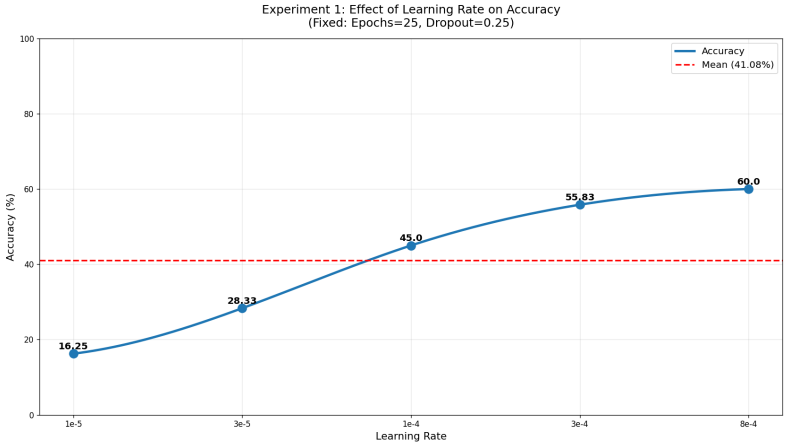

Figure 8. Curve of the influence of Learning Rate (LR) on classification accuracy of SmolVLM-256
图 8. 学习率(LR)对 SmolVLM-256 分类准确率的影响曲线

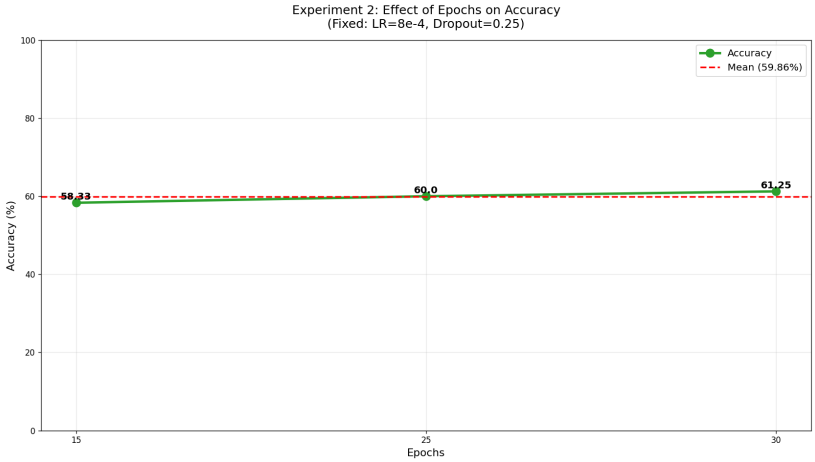

Figure 9. Curve of the influence of training Epochs on classification accuracy of SmolVLM-256
图 9. 训练轮数(Epochs)对 SmolVLM-256 分类准确率的影响曲线

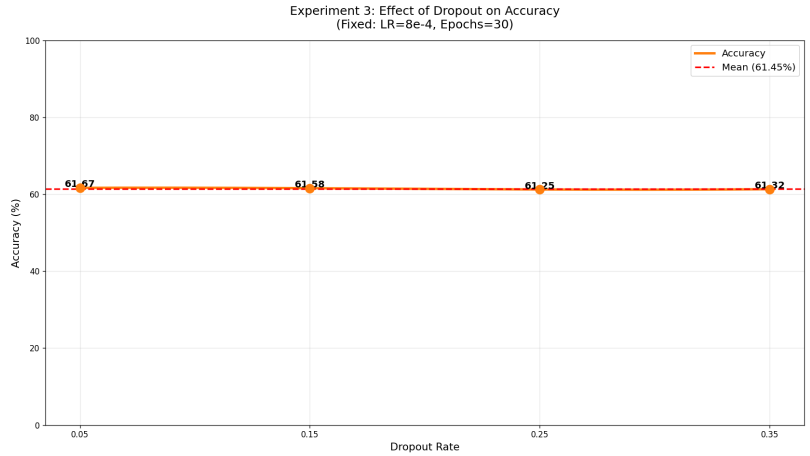

Figure 10. Curve of the influence of dropout ratio on classification accuracy of SmolVLM-256
图 10. Dropout 比例对 SmolVLM-256 分类准确率的影响曲线

模型默认配置下准确率仅为 61.25%，存在明显欠拟合问题。本文采用控制变量法完成参数遍历，最终确定最优组合：学习率 8e−4、训练轮数 30、Dropout 系数 0.05 [8]。

学习率：由图 8 可见，准确率随学习率上升稳步增长，8e−4 为最优取值，可兼顾收敛速度与训练稳定性。

训练轮数：如图 9 所示，训练轮数提升至 30 轮时，模型特征学习最为充分。

Dropout：结合图 10 结果，0.05 的低强度正则化可在抑制过拟合的同时保留特征表达能力。

2) Prompt 引导实验

Table 5. Prompt experimental results
表 5. Prompt 实验结果

Prompt 类型	Fusion 方式	Accuracy
No Prompt	Visual Only	71.67% ± 3.39%
Simple Prompt	Direct Concat	69.17% ± 2.43%
Simple Prompt	Weighted Fusion	71.34% ± 1.67%
Task Prompt	Weighted Fusion	71.02% ± 1.72%
Fine-grained Prompt	Weighted Fusion	70.83% ± 1.86%
Expert Prompt	Weighted Fusion	69.58% ± 2.83%

本文参照层级提示词设计思路[27]，搭建多类型 Prompt 开展对照实验，结果如表 5 所示。直接拼接文本特征会引入语义噪声，降低精度；加权融合可规避该问题。同时各类人工设计的静态 Prompt 均未带来明显性能提升，说明全局共享文本特征难以提供有效判别信息[6]。

3) 融合权重消融实验

为进一步研究文本语义在多模态融合中的影响，本文对加权语义融合(Weighted Semantic Fusion)中的融合权重参数 α 进行消融实验。实验通过设置不同的 α 参数，分析文本语义强度对模型分类性能的影响。

融合特征表达形式如下：

$$f = v + \alpha \boldsymbol{W} t \tag{1}$$

其中：v 表示视觉特征，t 表示 Prompt 文本特征，$\boldsymbol{W}$ 表示线性映射矩阵，α 表示文本语义融合权重。

Table 6. Fusion weight ablation experimental results
表 6. 融合权重消融实验结果

α	Accuracy
0.0	71.67%
0.1	71.34%
0.2	70.82%
0.5	70.12%
1.0	68.86%

文本语义融合权重 α 的测试结果如表 6 所示。随着 α 增大，模型准确率持续下降，证明本任务中视觉特征为分类核心，文本语义占比过高会造成干扰[9]，最优取值为 $\alpha = 0$。

4) 数据增强实验

在完成提示词语义引导消融实验的基础上，本文进一步探究不同图像增强策略对垃圾分类模型性能与泛化能力的影响，分别设置无增强、水平翻转、随机旋转、颜色抖动、随机裁剪 5 组消融实验组，实验结果如表 9 所示。

Table 7. Data augmentation experimental results
表 7. 数据增强实验结果

Augmentation	Accuracy
None	71.67% ± 3.39%
Horizontal Flip	77.50% ± 2.04%
Rotation	80.00% ± 3.39%
Color Jitter	77.50% ± 3.58%
Random Crop	85.83% ± 2.43%

各类数据增强方案的对比结果如表 7 所示。所有增强方式均可提升模型泛化能力，其中随机裁剪效果最优，能够有效模拟垃圾形变、位置变化等真实场景[10]；颜色抖动带来的增益相对有限。

5) 消融实验

为验证各优化模块对垃圾分类模型性能的独立贡献与协同增益，本文依次对学习率优化、训练轮次、Dropout、图像预处理、语义提示策略、数据增强、分类头结构、权重衰减等模块开展递进式消融实验，结果如表 8 所示。

Table 8. Ablation experimental results
表 8. 消融实验结果

实验编号	LR 优化	Epoch 优化	Dropout 优化	固定尺寸预处理	Prompt		Flip 增强	双层分类头	Weight-Decay	Accuracy
					Direct Concatenation Prompt	Weighted Semantic Fusion				
Exp-1	×	×	×	×	×	×	×	×	×	43.12% ± 4.15%
Exp-2	√	×	×	×	×	×	×	×	×	60.00% ± 7.38%
Exp-3	√	√	×	×	×	×	×	×	×	61.25% ± 4.86%
Exp-4	√	√	√	×	×	×	×	×	×	61.67% ± 4.49%
Exp-5	√	√	√	√	×	×	×	×	×	71.67% ± 3.39%
Exp-6	√	√	√	√	√	×	×	×	×	69.17% ± 2.43%
Exp-7	√	√	√	√	×	√	×	×	×	71.67% ± 1.67%
Exp-8	√	√	√	√	×	×	√	×	×	85.83% ± 2.43%
Exp-9	√	√	√	√	×	×	√	√	×	87.50% ± 1.32%
Exp-10	√	√	√	√	×	×	√	√	√	87.92% ± 1.56%

多模块递进消融结果如表 8 所示。学习率优化、图像预处理、随机裁剪带来的性能提升最为显著；

双层分类头可增强非线性表达，权重衰减进一步抑制过拟合[12]。全部模块叠加后，模型最终准确率达到87.92% ± 1.56%。

4.6.3. 结果分析与小节结论

优化后的轻量化模型 SmolVLM-256 识别效果可逼近大尺寸多模态模型。文本语义能辅助视觉区分相似垃圾类别，随机裁剪等几何增强手段提升模型对目标位置、拍摄角度变化的适配性。五折实验标准差仅 1.56%，模型在不同数据划分下性能波动小。该模型精度略低于 Qwen3-VL，但参数量与计算开销更低，更适配边缘设备落地。

本文针对该模型完成超参数、Prompt、特征融合、数据增强、分类头结构多组对比实验。实验证实学习率、迭代次数、Dropout [8]、文本提示加权融合方案[9]、空间几何图像增强[10]均能正向优化模型表现。调优完成后 SmolVLM-256 准确率为 87.92% ± 1.56% [4]，兼顾轻量化与识别精度。综上，轻量化视觉语言模型适用于细粒度垃圾分类任务，可为边缘智能垃圾分类设备提供技术方案。

4.7. 模型复杂度分析

本文结合模型架构、参数量与实验结果，对比各模型的识别效果与落地适配性，相关数据见表 9。

Table 9. Comparison of classification performance and deployment characteristics of different models
表 9. 不同模型分类性能与部署特性对比

Model	Accuracy	参数规模	边缘部署能力
ViT	87.50%	较大	一般
CLIP	86.00%	大	较低
BLIP	90.00%	大	较低
Qwen3-VL	96.25%	极大	较差
SmolVLM-256	87.92%	小	较强

准确率上，Qwen3-VL 表现最优，达到 96.25%；BLIP 次之，准确率为 90.00%；SmolVLM-256 准确率 87.92%，整体精度处于较高水平。部署层面，Qwen3-VL、BLIP、CLIP 参数量庞大，对算力和显存要求高，难以应用于边缘设备。SmolVLM-256 采用轻量化设计，在保证识别精度的同时资源消耗更低，更适配资源受限的边缘垃圾分类场景。

综合来看，大型多模态模型分类效果更佳，而 SmolVLM-256 实现了性能与部署成本的均衡，在实际落地场景中优势显著。

5. 结论与应用价值

本章总结全文研究成果，梳理核心结论，分析现有研究不足，并结合行业发展趋势提出后续研究方向。

5.1. 研究结论

1) SmolVLM-256 可有效适配小样本、形态多变的细粒度垃圾分类场景。经过全维度优化，模型准确率由 61.25%提升至 87.92% ± 1.56%，性能提升超 26 个百分点，轻量化模型具备充足的优化空间。

2) 训练策略与数据增强对模型影响显著，其中学习率调优、随机裁剪数据增强带来的性能增益最大[8] [10]，几何类增强可有效提升模型对垃圾形态变化的鲁棒性。

3) 本文采用的全局静态 Prompt 引导方式效果有限，文本语义无法有效辅助类别判别，模型主要依靠视觉特征完成分类，轻量化模型的多模态语义利用方式仍有待优化[6]。

4) 双层分类头结合权重衰减正则化，能够增强模型非线性表达能力、抑制小样本过拟合[20]，提升相似垃圾类别的区分效果。

5) 横向对比实验表明，SmolVLM-256 分类性能接近主流大型视觉语言模型，同时具备参数量小、算力开销低的特点[4] [5]，在边缘设备部署场景中实用性更强。

综上，本文形成了一套完整的轻量化模型优化方案。实验证明，借助针对性优化手段，轻量化视觉语言模型能够达到接近大型模型的分类水平，可为边缘端智能垃圾分类设备的研发提供参考。

5.2. 研究展望

受数据集规模、实验条件与研究周期限制，本研究仍存在拓展空间，结合多模态 AI 与智慧环保发展趋势，后续可从六个方向开展研究：

1) 扩充数据集品类与场景，新增厨余垃圾、有害垃圾及遮挡、极端光照、混合垃圾等复杂样本，构建通用型垃圾分类基准数据集[28]。

2) 探究 LoRA、Adapter、IA3 等参数高效微调技术，进一步降低模型训练与部署成本[29]。

3) 研发动态自适应 Prompt 机制，突破静态提示词的局限，提升文本语义的利用效率[30]。

4) 结合量化、剪枝、知识蒸馏等技术对模型二次压缩，实现嵌入式设备毫秒级实时推理[31] [32]。

5) 构建多模态持续学习框架，引入增量学习能力，使模型可快速识别新增垃圾品类[33]。

6) 将模型部署至树莓派、智能垃圾桶、巡检机器人等实体设备，开展实景测试，推动技术产业化落地[1]。

综上，轻量化视觉语言模型在智能垃圾分类领域具备广阔应用前景，伴随边缘人工智能与多模态技术的持续发展，轻量化、低成本的智能分类方案将为城市生态治理与智慧环保建设提供有力支撑。

基金项目

本课题受到“湖北师范大学 2025 年国家级大学生创新创业训练计划项目”(项目编号：D202505311134555171)的资助。

参考文献

[1] 肖克江, 陈亮, 高阔, 杨文齐, 庞世燕. 基于多模态数据融合的边缘设备轻量级垃圾分类方法和系统研究[J]. 工程科学学报, 2025, 47(9): 1905-1916.

[2] 魏秀参, 杨晓龙, 李泽超, 等. 细粒度图像分析研究进展与展望[J]. 计算机学报, 2023, 46(4): 861-891.

[3] 王晋东, 陈益强. 小样本学习研究综述[J]. 计算机学报, 2021, 44(11): 2441-2470.

[4] 张正, 王超, 李凯, 等. 深度学习模型轻量化技术综述[J]. 软件学报, 2023, 34(5): 2319-2348.

[5] Marafioti, A., Zohar, O., Farré, M., Noyan, M., Bakouch, E., Cuenca, P., *et al.* (2025) SmolVLM: Rede-Fining Small and Efficient Multimodal Models. https://arxiv.org/abs/2504.05299

[6] 刘知远, 孙茂松. 提示学习研究进展与展望[J]. 中国科学: 信息科学, 2023, 53(2): 247-273.

[7] 孙紫阳, 刘柱, 王向前, 李娜. 面向形态多变垃圾的细粒度生活垃圾数据集[EB/OL]. https://github.com/1583118323/Waste_FineGrained_Dataset, 2025-12-20.

[8] Bergstra, J. and Bengio, Y. (2012) Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, **13**, 281-305.

[9] Cao, Y., Liu, Y.Z., Wang, W.H., *et al.* (2024) MMFuser: Multimodal Multi-Layer Feature Fuser for Fine-Grained Vision-Language Understanding. *IEEE Transactions on Image Processing*, **33**, 4721-4734.

[10] Chen, Q.P., Jiao, L., Wang, F.M., Du, J.M., Liu, H.Y., Wang, X. and Wang, R.J. (2024) Integrating Foreground-Background

Feature Distillation and Contrastive Feature Learning for Ultra-Fine-Grained Visual Classification. *Pattern Recognition*, **150**, 110339. https://doi.org/10.1016/j.patcog.2024.110339

[11] Shorten, C. and Khoshgoftaar, T.M. (2019) A Survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, **6**, Article No. 60. https://doi.org/10.1186/s40537-019-0197-0

[12] 李玉洁, 马子航, 王艺甫, 等. 视觉 Transformer (ViT)发展综述[J]. 计算机科学, 2025, 52(1): 194-209.

[13] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[14] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al*. (2021) An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*, 3-7 May 2021, 1-16.

[15] Li, J., Li, D., Xiong, C. and Hoi, S. (2022) BLIP: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation. *Proceedings of the* 39*th International Conference on Machine Learning*, Baltimore, 17-23 July 2022, 12888-12900.

[16] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., *et al*. (2021) Learning Transferable Visual Models from Natural Language Supervision. *Proceedings of the* 38*th International Conference on Machine Learning*, 18-24 July 2021, 8748-8763.

[17] Bai, J., Bai, Y., Wang, Z., *et al*. (2023) Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. https://arxiv.org/abs/2308.12966

[18] 刘群, 李鹏, 王士进, 等. 视觉语言预训练模型研究综述[J]. 计算机学报, 2024, 47(2): 341-372.

[19] 李卫. 视觉 Transformer 模型设计及其轻量化研究[D]: [硕士学位论文]. 成都: 电子科技大学, 2023.

[20] Srivastava, N., Hinton, G., Krizhevsky, A., *et al*. (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, **15**, 1929-1958.

[21] Wu, W., Zhang, Y., Li, X., *et al*. (2023) Applications of Convolutional Neural Networks for Intelligent Waste Identification and Recycling: A Review. *Journal of Cleaner Production*, **397**, 136542. https://doi.org/10.1016/j.resconrec.2022.106813

[22] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[23] Howard, A., Zhu, M., Chen, B., *et al*. (2017) MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. https://arxiv.org/abs/1704.04861

[24] Chen, G., Li, Y. and Wang, H. (2025) RepVGG-MEM: A Lightweight Model for Garbage Classification Achieving a Balance between Accuracy and Speed. *IEEE Access*, **13**, 28764-28775. https://doi.org/10.1109/ACCESS.2025.3544631

[25] Wang, L., Zhang, Q. and Liu, H. (2023) Integrating Human Vision Perception in Vision Transformers for Classifying Waste Items. *Applied Sciences*, **13**, 13215.

[26] Chen, X., Li, Y., Zhang, H., *et al*. (2024) Zero-Shot Garbage Classification Using CLIP with Prompt Engineering. 2024 *IEEE International Conference on Multimedia and Expo*, Niagara Falls, 15-19 July 2024, 1-6.

[27] Zhang, C., Liu, Y. and Wang, H. (2025) Hierarchical Prompt Design for Fine-Grained Image Classification. 2025 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Seattle, 11-15 June 2025, 123-130.

[28] 黄凯奇, 陈晓棠, 康运锋, 等. 基于深度学习的垃圾分类方法研究进展[J]. 自动化学报, 2022, 48(3): 647-666.

[29] Hu, E.J., Shen, Y., Wallis, P., *et al*. (2022) LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations*, 25-29 April 2022, 1-17.

[30] Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E. and Singh, S. (2020) AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. *Proceedings of the* 2020 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), 16-20 November 2020, 4222-4235. https://doi.org/10.18653/v1/2020.emnlp-main.346

[31] Frantar, E., Ashkboos, S., Hoefler, T., *et al*. (2023) GPTQ: Accurate Post-Training Quantization for Generative Pre-Trained Transformers. https://arxiv.org/abs/2210.17323

[32] Hinton, G., Vinyals, O. and Dean, J. (2015) Distilling the Knowledge in a Neural Network. https://arxiv.org/abs/1503.02531

[33] Wang, Z., Li, Y. and Zhang, H. (2025) Continual Learning for Vision-Language Models: A Survey. https://arxiv.org/abs/2501.07654