ReCCur：面向开放与边缘场景下鲁棒视觉语言理解的递归式边界案例筛选框架

Wei Yihan[1], Shenghai Yuan[1], Tianchen Deng[2], Boyang Lou[3], Enwen Hu[3]南洋理工大学 [2]上海交通大学 [3]北京邮电大学{yihan005, shyuan}@e.ntu.edu.sg, tianchen.deng@sjtu.edu.cn, {woshiluobodan, owen.hu}@bupt.edu.cn

arXiv:2601.03011v1 [cs.CV] 6 Jan 2026

摘要

边缘案例——导致现实世界故障的罕见或极端场景——难以大规模整理：网络数据噪声较多，标签不可靠，且边缘部署无法进行大规模重新训练。我们提出***ReCCur***（递归式边缘案例整理），一种低计算资源、核心免训练框架，通过多智能体递归流水线将含噪网络图像转化为可审计的细粒度标签。首先，大规模数据采集与过滤 利用视觉语言模型（*VLM*）扩展领域词汇表、爬取网络数据，并通过少量人工抽查确保三模态（图像/描述/关键词）一致性，从而筛选出精炼候选项。接着，混合专家知识蒸馏 使用互补编码器（如 *CLIP/DINOv2/BEiT*）进行 k*NN* 投票，并结合双重置信度激活和不确定性采样，最终收敛至高精度集合。最后，区域证据视觉语言模型对抗式标注将提议器（多粒度区域与语义线索）与验证器（全局+局部链式一致性）配对，以生成可解释标签并形成闭环。在逼真的边缘案例场景（如泡水车检测）中，*ReCCur* 可在消费级*GPU*上运行，持续提升纯度和可分性，且只需极少人工监督——为资源受限条件下的下游训练与评估提供实用基础。我们的代码和边缘案例数据集将在论文录用后开源。

1. 引言

边缘案例——那些罕见或极端、会打破常见训练假设的情形——在开放世界动态和分布偏移下，正日益成为决定真实场景可靠性的关键因素。模型必须在缺少或完全没有领域内标签的情况下保持稳健，而边缘部署则受限于计算资源，无法频繁进行大规模重新训练。这引出了一个核心问题：如何以低成本、尽量减少人工投入，持续采集、清理并在语义层面丰富边缘案例数据？近期少样本和开放集研究呈现出三条发展方向：**(i)** 在冻结或轻度调优空间中稳定决策边界（间隔优化、局部增强）；**(ii)** 利用外部语义或检索进行增强（名称扩展、直推式推理、检索增强证据）；**(iii)** 进行轻量级跨领域自适应。这些方法在经过整理的数据环境中虽效果显著，但通常需要训练或微调，并假设数据/语义具有较高质量且可控。在噪声较多的网络数据采集中，主题漂移和标签噪声会削弱收益；大规模筛选和重新标注仍然消耗大量计算资源和人力，在边缘硬件上尤为如此。网络监督/自监督整理（置信度重加权、聚类去噪、在线/主动选择）虽能降低成本，但仍依赖常规的大批量训练动态和稳定先验——当边缘案例占主导且资源紧张时，这些条件便显得脆弱。

这些限制凸显了三大核心挑战：可获取性——扩展领域词汇，并以高召回率爬取开放网络，同时控制分布偏移和长尾噪声；可信度——通过多模态一致性和类内对齐筛选海量候选项，无需大规模重新训练，同时保留稀有正例并剔除虚假正例；可部署性——在计算资源有限、人力预算极少的情况下形成闭环，生成易于审计和复用、可解释且细粒度的标签。

我们提出 **ReCCur**(递归式边缘案例整理），这是一种免训练核心的多智能体框架（图1），通过运行递归处理流程——爬取 → 过滤 → 提炼 →基于证据验证的重新标注——将含噪网络图像转化为高置信度语义标签。视觉语言模型（VLM）扩展词汇表（支持多种语言）、驱动大规模爬取，并衡量图像／描述／关键词之间的三模态一致性；聚类结合小批量人工抽查对数据进行划分，并迭代改进提示词，从而得到精炼后的数据池。一个基于向量索引的混合专家（互补视觉编码器）执行 kNN 投票，并通过主题／标签置信度激活进行决策；不确定性采样器则筛选出对齐度低和边界案例，仅需少量人工审核，最终收敛至一个高精确率数据集。最后，区域证据阶段使用一个 *Pro-*

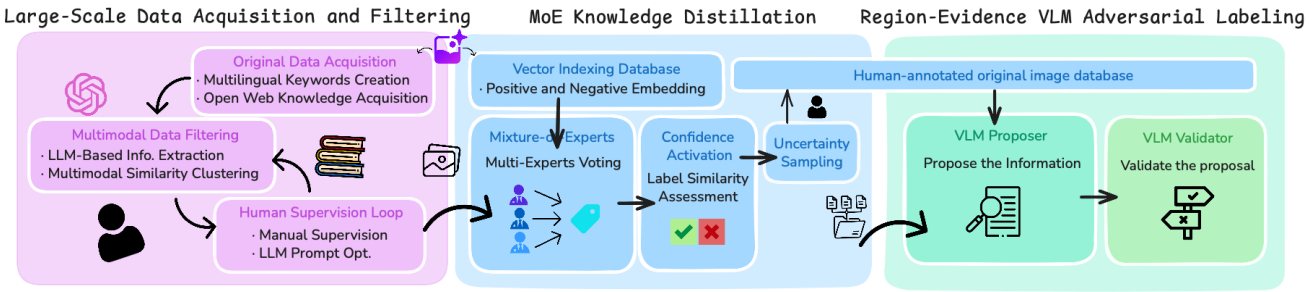


图 1. **ReCCur 概览**。通过多模态采集与筛选、混合专家标注（结合置信度 + 不确定性）以及区域证据视觉语言模型，逐步提纯数据并优化标签。

poser 以生成多粒度区域和语义线索，并由验证器执行全局+局部链式一致性检查，生成可审计的细粒度标签，并将其反馈到循环中。

贡献。

• 我们提出**ReCCur**，一种用于边缘案例数据整理的递归式核心免训练框架，能够在计算资源有限且仅需极少人工监督的条件下，实现高召回率网络采集和高精度筛选。

•我们结合三模态一致性、带双重置信度激活的混合专家向量投票，以及由不确定性驱动的人在回路机制，在严格控制标注成本的同时最大限度地提高数据纯度。

•我们提出一个区域证据对抗式重标注阶段，以生成可解释的细粒度语义，并完成数据整理闭环，以便下游复用。

2. 相关工作

2.1. 视觉领域中的边缘案例与小样本学习

视觉领域的小样本识别大体沿着三条路线发展。*(i)* 特征空间稳定化假设骨干网络固定或仅经过轻微调优，并通过重塑嵌入，使决策边界在标签稀缺时仍保持可靠——方法包括实例级最大间隔目标、直接在冻结空间中合成虚拟支持集的操作、强调局部判别性区域的基于注意力的正则化，以及扩展类内覆盖同时保持类间可分性的特征／语义通道增强 [1, 2, 13, 20, 30, 45]。*(ii)* 外部语义与检索通过利用大语言模型扩展类别名称、在测试批次上进行传导式优化，以及从大型检索语料库聚合邻域或进行分阶段检索增强微调来扩大覆盖范围；开放世界变体则通过双向图像↔文本对齐和提示学习来增强 CLIP，以用于检测和开放集识别，从而提升对未知类别的鲁棒性 [9, 15, 16,28, 29, 34, 38, 41]。*(iii)* 参数高效适配通过轻量级仿射／正交映射或可学习的注意力模块应对领域偏移，以最少的更新动态调整特征分布 [31, 32]。当语义具有可迁移性时，这些方向能显著提升小样本／开放集准确率；然而，它们通常需要（重新）训练和经整理的资源，并且在噪声较多的长尾网络采集和严格的边缘计算预算下，效果可能会减弱——而这些条件恰恰是边缘案例占主导的场景。

2.2. 数据爬取与自监督数据集构建

大规模网络／自监督数据整理旨在通过高效去噪实现高召回率。各类流程会评估图像–文本一致性，通过自一致性／置信度学习将可靠性内化，对噪声标签重新加权，并结合聚类与平衡采样，在无需大量人工标注的情况下提升语义纯度 [23, 37, 39]。为扩展覆盖范围和可迁移性，文本引导的增量扩展与外部记忆检索利用网络规模的嵌入，在存在大量噪声和标签漂移的情况下增强判别能力 [21, 27]。从评估和缩放角度来看，统一的候选池／协议将筛选／重采样效应系统化；数据集增长强调质量×多样性，并结合（重新）生成图注与多语言增强；缩放定律揭示了筛选强度、数据集大小和可用算力之间的耦合关系，表明有必要采用预算感知型数据整理策略 [2, 11, 14, 33]。可学习／在线／主动式流程进一步将选择转化为可训练筛选器，并利用不确定性、信息增益或相关性–特异性信号，在资源有限的情况下发现高价值子集 [25, 36]。尽管取得了进展，许多方法仍假设存在规律的大批量训练动态或稳定先验；在噪声严重且主题不断变化的边缘场景中，它们仍未能解决如何以免训练或最少监督的方式管理边缘案例数据这一问题。

2.3. 大语言模型智能体与多智能体投票

除简单的多数投票外，多智能体大语言模型框架还强调协作式审议和过程控制。

圆桌协议先进行独立推理，再通过相互质询达成共识 [5]。对于长上下文任务，智能体链通过明确的交接来分解问题，从而在较长的推理窗口中提升稳定性 [43]；反思和逐步审查则增加自检与交叉检查，以减少错误累积 [4]。另一条互补路线则将论证和裁决形式化。辩论智能体引出不同的假设，并对各项假设进行压力测试 [7, 26]；评判模型为论点打分，将分歧转化为监督信号 [10, 12, 19, 42, 44]。角色专门化系统（如 MetaGPT）引入软件工程抽象概念——角色、接口和制品——使协作过程可追踪、可审计 [18]。视觉以外的应用（金融决策；文化敏感型对齐）凸显了在异质先验下实现概念落地和价值平衡的重要性 [24, 40]。遵循这些原则，我们协调互补专家给出基于检索的判断，并采用双置信度激活和不确定性显化；同时，区域感知型视觉语言模型组件提供局部细粒度证据，用于集成前的一致性检查。这种协调能够调和嘈杂信号、节省人力，并在计算资源有限的情况下迭代提升边缘案例标签质量。

3. 方法

3.1. 大规模数据获取与筛选

整体结构。 图 2 中的整体处理流程包含三个部分：数据采集、数据筛选和人在回路中的监督。给定 $\mathcal{D}$，视觉语言模型将图像编码为特征描述 H_i。对于每个样本，三元组$\mathbf{u}_i = (X_i,\ H_i,\ Q_i)$ 可用于多模态相似度粗筛。聚类模块将增强后的嵌入 $\mathbf{z}_i$进行分组，随后利用一个小型标注集将这些簇划分为三个类别。根据这些标注，重新向视觉语言模型提供经过优化的指令，生成精炼数据集 Π。

原始数据采集。 给定一小组样本图像，每个类别 c由 $\mathcal{D}_c = \{X_{c,1}, X_{c,2}, \ldots, X_{c,N_c}\}$表示，其中 $X_{c,\ i}$表示类别 c 的第 i 张图像，N_c 表示样本数量。每个类别都由视觉提示和文本提示共同指定，后者编码了数据集上下文和任务目标。

使用视觉语言模型，将这些信息扩展为一组高度相关的爬虫关键词描述 $\mathcal{K}_{\mathrm{V}} = \mathcal{G}_{\mathrm{V}}(\mathcal{D}_c)$，适用于大规模网络采集。

为最大限度地进行数据采集，我们采用双重策略：除了根据视觉-语言联合提示生成爬虫关键词外，我们还单独使用文本提示来获取 $\mathcal{K}_{\mathrm{T}} = \mathcal{G}_{\mathrm{T}}(\mathcal{D}_c)$，从而增加采集数据的总量。为进一步提高爬虫的多样性，我们使用多语言词库 $\mathcal{L}(\cdot)$ 扩充关键词，

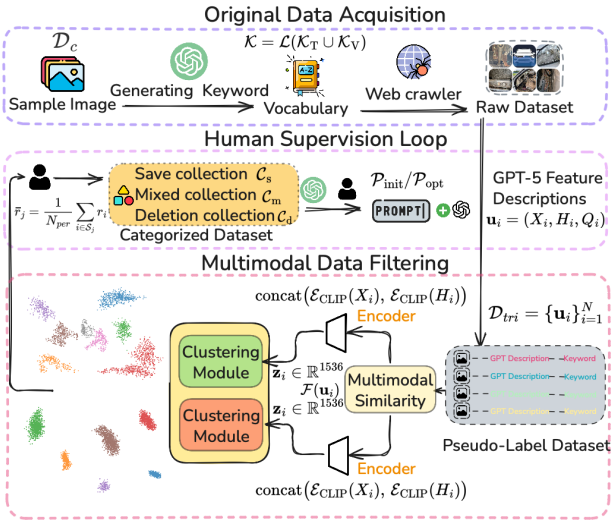


图2. **大规模数据获取与筛选的总体框架。**由大语言模型扩展的关键词指导爬取；人工监督循环优化提示词、添加描述词，并将样本分类为保留/混合/删除；视觉语言聚类剔除噪声，生成带伪标签的高质量数据集。

最终得到关键词语料库

$$\mathcal{K} = \mathcal{L}(\mathcal{K}_{\mathrm{T}} \cup \mathcal{K}_{\mathrm{V}}).$$

基于这一扩充语料库，我们执行关键词引导的网络爬取，以从开放网络来源构建一个预期的大规模数据集 $\mathcal{D}$。

多模态数据过滤。开放网络数据噪声较多：尽管使用高相关性关键词 Q_i 检索到 $\mathcal{D}$，许多图像仍然无关或相关性较弱，因此我们对其进行筛选。使用提示词 $\mathcal{P}_{\mathrm{init}}$ 或 $\mathcal{P}_{\mathrm{opt}}$ 与 ChatGPT-5 配合，为每个 $X_i \in \mathcal{D}$生成特征描述 H_i，形成图像–特征–关键词三元组

$$\mathbf{u}_i = (X_i, H_i, Q_i), \quad \mathcal{D}_{\mathrm{tri}} = \{\mathbf{u}_i\}_{i=1}^{N}.$$

借助这些元组，我们使用预训练 CLIP 对样本进行嵌入，并分别计算图像与描述、图像与关键词以及描述与关键词之间的余弦相似度，从而得到三个互补的指标。随后通过模态特定权重将它们融合为一个多模态相似度分数。

$$\mathrm{Sim}_{\mathrm{mm}}(\mathbf{u}_i) = \mathcal{F}(\mathbf{u}_i).$$

多模态相似度低于 τ_{mm} 的样本进入低相似度集合。高于 τ_{mm} 的样本则保留在高相似度集合中，用于聚类/筛选，从而去除大多数不相关或关联较弱的图像，并得到 Π。最后，将H_i 与预训练 CLIP 编码器得到的图像 X_i 嵌入拼接，形成一个向量来表示每个样本，从而得到语义增强嵌入

$$\mathbf{z}_i = \text{concat}\big(\mathcal{E}_{\text{CLIP}}(X_i),\, \mathcal{E}_{\text{CLIP}}(H_i)\big) \in \mathbb{R}^{1536}.$$

通过对 Π中所有样本的特征向量$\mathbf{z}_i$ 进行聚类，我们得到

$$\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_J\}.$$

人工监督循环。 聚类后，我们得到 $\mathcal{C}$，其中每个 $\mathcal{C}_j$ 都包含语义相似的图像。为减轻人工负担，每个簇中仅对 N_{per} 张图像进行人工标注。根据每个簇中 N_{per} 已标注图像组成的小批次对簇进行评分，得到

$$\bar{r}_j = \frac{1}{N_{per}} \sum_{i \in \mathcal{S}_j} r_i, \quad \mathcal{S}_j \subset \mathcal{C}_j,$$

其中， $r_i = 1$ 表示高相关性， $r_i = 0$ 表示低相关性，$\mathcal{S}_j \subset \mathcal{C}_j$ 是 N_{per}图像样本。具有 $\bar{r}_j \geq 0.8$ 的簇被分配至强集合 $\mathcal{C}_{\text{s}}$，具有$0.2 < \bar{r}_j < 0.8$ 的簇构成混合集合 $\mathcal{C}_{\text{m}}$，具有$\bar{r}_j \leq 0.2$ 的簇构成丢弃集合 $\mathcal{C}_{\text{d}}$。

通过这一分类过程，我们可以提炼出应予保留的显著特征，刻画混合或模糊案例，并可靠地识别出不符合标准、应予移除的实例。通过将人工监督与视觉语言模型相结合，原始提示 $\mathcal{P}_{\text{init}}$ 得以优化，从而获得改进后的提示$\mathcal{P}_{\text{opt}}$ 。

3.2. 混合专家知识蒸馏

整体结构。 混合专家知识蒸馏（图 3）对第一阶段的输出进行优化 Π ，以筛选出符合数据集要求的图像。该流程包含四个部分：(1) 构建向量索引数据库 $\mathcal{I}$；(2) 执行混合专家判断；(3) 应用置信度激活；(4) 通过选择性人工标注进行筛选。候选项通过 $\mathcal{I}$进行评估，由专家标注，经激活模块验证，并对少量路由子集进行人工标注，以形成供下一轮使用的高精度池。经过多轮迭代后，具有预测标签 $\hat{}y$ 的 $\mathcal{Y}$ 图像具有较高的准确率；汇总所有已处理数据即可得到数据集 π。

向量索引数据库。 向量索引数据库$\mathcal{I}$由一个小型人工标注参考数据集 $\mathcal{A} = \{(X_i,\ y_i)\}_{i=1}^{N}$构建而成，其中 X_i 表示一张带标签的图像， y_i 表示其对应的类别。为获取互补信息，采用了三种预训练图像嵌入模型：**CLIP**、**DINOv2**和**BEiT**。

具体而言，CLIP 为参考数据 $\mathcal{I}_{\text{CLIP}} = \{(y_i,\ \mathcal{E}_{\text{CLIP}}(X_i)) \mid X_i \in \mathcal{A}\}$ 提供语义表示，DINOv2 则编码空间视觉特征 $\mathcal{I}_{\text{DINOv2}} = \{(y_i,\ \mathcal{E}_{\text{DINOv2}}(X_i)) \mid X_i \in \mathcal{A}\}$，而 BEiT 捕获通用视觉表示 $\mathcal{I}_{\text{BEiT}} = \{(y_i,\ \mathcal{E}_{\text{BEiT}}(X_i)) \mid X_i \in \mathcal{A}\}$。每个模型都会对所有已标注图像进行嵌入，从而生成三个存储类别-向量对的子数据库。随后，将这些子数据库的并集定义为完整的向量索引数据库：

$$\mathcal{I} = \{\mathcal{I}_{\text{CLIP}}, \mathcal{I}_{\text{DINOv2}}, \mathcal{I}_{\text{BEiT}}\}.$$

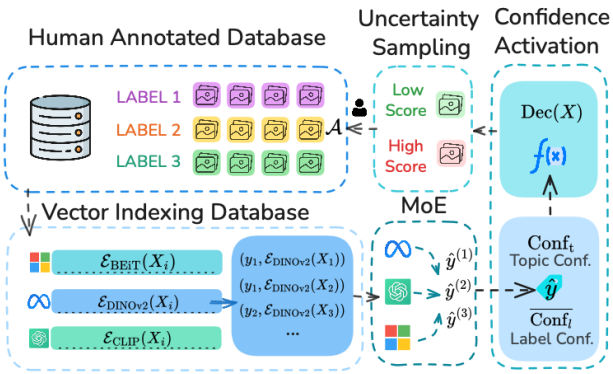


图 3. **混合专家知识蒸馏的整体结构**。 人工标注的图像被嵌入向量索引中。混合专家预测$\hat{y}$ 及主题/标签置信度；不确定性采样会将低置信度案例升级交由标注人员处理，而置信度激活则解码出最终决策。

混合专家。 三位专家 $\mathcal{M}$ ={CLIP、 DINOv2、BEiT} 分别查询向量索引中的各自子索引，以独立作出判断。对于每张测试图像，专家 m会对其进行嵌入，并检索排名前K 的邻居，从而得到

$$\mathcal{N}^{(m)} = \{(\xi_k^{(m)}, y_k^{(m)})\}_{k=1}^{K}, \quad m \in \mathcal{M},$$

其中， $K \in \mathbb{N}$是检索邻居的数量， $\xi_k^{(m)}$ 表示测试图像与专家 m子数据库中第 k个最近邻之间的相似度，$y_k^{(m)}$ 是该邻居的类别标签。

温度参数 T 将相似度$\{\xi_k^{(m)}\}_{k=1}^{K}$ 转换为归一化的 Softmax 权重。专家 m 的预测标签为

$$\hat{y}^{(m)} = \arg\max_{c \in \mathcal{Y}} \sum_{k=1}^{K} \frac{\exp(\xi_k^{(m)}/T)}{\sum_{j=1}^{K} \exp(\xi_j^{(m)}/T)} \mathbf{1}[y_k^{(m)} = c],$$

其中， $\mathbf{1}[\cdot]$ 是指示函数。

根据三位专家的预测 $\hat{}y^{(m)}$，通过多数投票得出最终标签 $\hat{}y$ 。若出现三方冲突，则优先采用 CLIP 的结果，同时将图像转交给**不确定性采样**模块，利用互补的语义、结构和通用视觉先验，提升标签分配的鲁棒性和可靠性。

置信度激活。对于每张测试图像，我们使用两项指标评估其已分配标签 $\hat{y}$ 的可信度：主题置信度 $\mathrm{Conf}_{\mathrm{t}}$ 和标签置信度 $\mathrm{Conf}_{\mathrm{l}}$。

主题置信度是来自一个专家的 Top-K 个向量的相似度，

$$\mathrm{Conf}_{\mathrm{t}}(X) = \frac{1}{K}\sum_{k=1}^{K}\xi_k,$$

用于捕捉 X 与其检索邻居之间的亲和性。

标签置信度是图像嵌入与类别均值嵌入之间的余弦相似度，

$$\mathrm{Conf}_{\mathrm{l}}^{(m)}(X) = \cos\left(\mathcal{E}^{(m)}(X), \boldsymbol{\mu}_{\hat{y}}^{(m)}\right),$$

用于量化预测标签 $\hat{y}$ 的可信度。

最终的标签置信度通过对相应的专家分数取平均得到：

$$\overline{\mathrm{Conf}_{\mathrm{l}}}(X) = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\mathrm{Conf}_{\mathrm{l}}^{(m)}(X).$$

最后，决策函数 $\mathrm{Dec}(X)$ 定义如下：

$$\mathrm{Dec}(X) = \begin{cases} \hat{y}, & \text{if } \mathrm{Conf}_{\mathrm{t}}(X) \geq \theta_{\mathrm{t}} \text{ and } \overline{\mathrm{Conf}_{\mathrm{l}}}(X) \geq \theta_{\mathrm{l}}, \\ \text{non-target}, & \text{otherwise.} \end{cases}$$

不确定性采样。为降低标注成本并最大化监督数据效用，我们采用不确定性采样模块，优先选择信息丰富的样本。对于每个预测类别，我们计算特征对齐分数（FAS）——图像嵌入与索引中类别均值嵌入 $\boldsymbol{\mu}_{\hat{y}}$ 之间的余弦相似度，并对各专家的分数取平均：

$$\overline{\mathrm{FAS}}(X,\hat{y}) = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\frac{\mathcal{E}^{(m)}(X)\cdot\boldsymbol{\mu}_{\hat{y}}^{(m)}}{\|\mathcal{E}^{(m)}(X)\|\,\|\boldsymbol{\mu}_{\hat{y}}^{(m)}\|}.$$

按此分数对各类别中的样本进行排序。根据排序结果，我们构建一个低分池（每个类别中分数最低的 $\alpha\%$ c_k），并从中随机抽取 K_L 个样本，以反映较高的不确定性。

对于每个非目标样本 X，我们计算其与所有类别的对齐分数：

$$\overline{\mathrm{FAS}}(X,c_k), \quad c_k \in \{c_1, c_2, \ldots, c_{|\mathcal{Y}|}\}.$$

其中，$|\mathcal{Y}|$ 表示类别总数。随后，我们将 X 的边界强度和边界归因类别定义为

$$B(X) = \max_{k=1,\ldots,|\mathcal{Y}|}\overline{\mathrm{FAS}}(X,c_k),$$

$$y^{\star}(X) = \arg\max_{k=1,\ldots,|\mathcal{Y}|}\overline{\mathrm{FAS}}(X,c_k).$$

$B(X)$衡量按类别划分的最强对齐程度，而$y^{\star}(X)$用于识别对齐类别；根据 $B(X)$对非目标样本进行排序，并将排名前 K_H 的样本作为边界候选样本。随后，对来自两个样本池的样本进行人工标注，并将所得标签整合到向量索引数据库中，以支持下一轮专家迭代。

3.3. 区域证据视觉语言模型对抗式标注

整体流程。在大规模采集/筛选和混合专家标注的基础上，该流程为数据集 π生成了质量较高但并不完美的标签。为进一步提升准确性和语义粒度，我们引入了两阶段区域证据视觉语言模型（RE-VLM）：视觉语言模型提议器（图 4）提取全局和多区域特征，视觉语言模型验证器（图5）则根据这些线索推断并验证类别和属性，从而生成针对特定图像的专有标签。

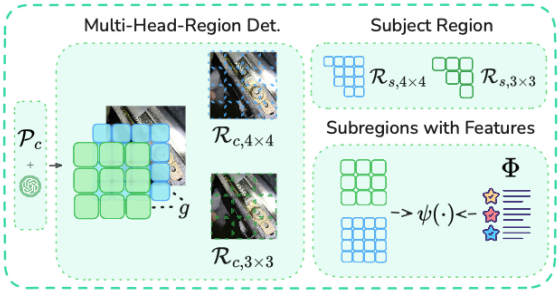


图4. **视觉语言模型提议器的架构**。提议器输出多头区域网格，分离主体和子区域，并为每个区域标注语义线索 $\boldsymbol{\Phi}$。

视觉语言模型提议器。作为信息提议器，提议器首先确定图像主体。在先验有限的边缘场景中，仅有类别标签是不够的。因此，我们根据标签编写通用且具有区分性的类别级特征描述，以构建视觉语言模型提示 $\mathcal{P}_c$。对于语义更丰富的标签，我们通过对主体进行多头区域检测来强化空间关注，为验证器提供更精细的证据。具体而言，我们使用两个头（3×3 和 4×4）。类别 c在粒度 g 下的子区域集合定义为

$$\mathcal{R}_{c,g} = \{\mathbf{r}_{c,g}^{(1)}, \mathbf{r}_{c,g}^{(2)}, \ldots, \mathbf{r}_{c,g}^{(N_g)}\}, \quad g \in \{3\times3, 4\times4\}$$

其中 $\mathcal{R}_{c,\ g}$ 表示在粒度 g下获得的子区域集合，N_g 是相应的子区域数量。

最后，将类别特征描述、输入图像 X、主体区域（下文记为 $\mathcal{R}_{s,g}$）以及所有特征属于特征集 $\boldsymbol{\Phi}$ 的子区域提供给模型：

$$\text{Prop}(X, \mathcal{P}_c) \rightarrow \left(\mathcal{R}_{s,g}, \{\psi(\mathbf{r}_{c,g}^{(j)}, \mathbf{\Phi})\}_{j=1}^{N_g}\right),\ \mathcal{R}_{s,g} \subset \mathcal{R}_{c,g},$$

其中，$\psi(\cdot, \mathbf{\Phi})$ 表示子区域 $\mathbf{r}_c^{(j)},_g$ 是否包含来自 $\mathbf{\Phi}$ 的特征。

视觉语言模型验证器。 验证器执行全图推理，以生成 $\hat{}$ $y_{\text{sem-1}}$ under $\mathcal{P}_c$，随后接收提议器提供的主区域和多粒度子区域，并将其与 $\mathcal{P}_c$ 对齐，以获得 $\hat{}y_{\text{sem-2}}$。为确保可靠性，我们采用链式一致性检查：在 $\mathbf{\Phi}$中检测特征并推断其是否存在；利用提议器的输出，根据 $\mathcal{R}_{c,\ g}$定位并评估子粒度区域；最后融合这两部分证据，得到 $\hat{}y_{\text{sem}}$。

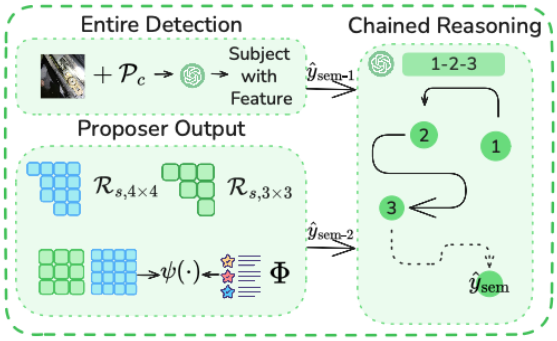


图5。**视觉语言模型验证器的架构**。 验证器执行全局推理，通过链式推理融合证据，并输出$\hat{y}_{\text{sem}}$。

4. 实验

本节通过详细实验验证整个架构的有效性。实验旨在通过构建一个包含积水车辆边缘案例、迷幻蘑菇边缘案例和墙体损坏边缘案例的新数据集，测试该架构在不同条件下的可行性。

4.1. 实现

设置 在实验中，所有与 VLM 相关的组件均使用 ChatGPT-5（公共 API；检查点：2025年-08-07）。我们采用了三种预训练视觉嵌入模型——CLIP ViT-L/14、DINOv2-L 和 BEiT-L/16——作为所有对比研究中的视觉骨干网络。针对每种骨干网络，我们分别构建了一个检索数据库，以生成专家标签。为确保在计算资源有限的情况下仍具可行性，端到端处理流程主要在一台笔记本电脑上运行，该电脑配备 NVIDIA GeForce RTX 3060（6 GB）和 Intel Core i7-11800H 中央处理器；在这台机器上难以实际执行的工作负载，则在服务器级 NVIDIA A10 GPU 上运行。

评估指标。 对于混合专家知识蒸馏，我们使用宏平均精确率、召回率和 F1 评估所有方法，这些指标根据真正例/假正例/假负例计算，除非另有说明，均在各类别上进行宏平均；此外，还采用两项专门针对筛选的指标：噪声去除率（NRR），即经人工验证的噪声样本中被剔除的比例；以及干净数据保留率（CDRR），即经人工验证的干净样本中被保留的比例。对于区域证据 VLM 对抗式标签标注，我们还使用精确率和召回率评估预测标签，以验证语义特征保真度。

基线选择。 我们将本方法与四类具有代表性的方法进行比较，涵盖度量学习、噪声剪除、开放词汇识别和语言引导推理。为确保公平性和可复现性，所有实现均取自官方代码库和公开发布版本。所有基线均在相同配置下运行，包括数据划分、输入分辨率、提示词/文本词汇表和计算预算，并遵循我们的评估协议与指标（精确率/召回率/F1、NRR、CDRR）。这一设计能够比较基于聚类的方法、开放词汇模型和视觉语言模型问答变体。

4.2. 案例研究：泡水车检测

背景。 极端天气导致车辆进水日益常见，在一些市场中，水灾损坏车辆经翻新后会再次出售。这些车辆会危害安全并损害消费者保护，而买家往往缺乏识别水损指标的专业知识。因此，我们采用 ReCCur 构建泡水车辆检测点数据集，其内容与认证机构认可的诊断部件相对应。ReCCur 从噪声网络数据中提炼高质量样本，并生成针对特定部件的特征感知语义标签，从而形成专用数据集，支持在边缘场景中从头进行验证。图6 概述了任务配置和实验目标。

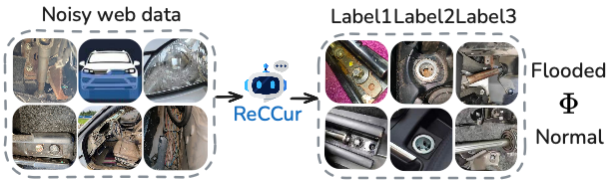


图6。**任务配置与目标**。ReCCur生成细粒度部件级语义标签，并支持比较不同方法在识别和过滤性能方面的表现。

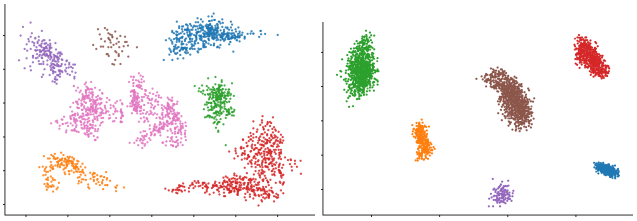
实现细节。 根据涉水车辆检查报告，我们选取常见痕迹点作为种子，构建 $\mathcal{K}_{\text{V}}$ 和 $\mathcal{K}_{\text{T}}$，形成用于原始数据采集的 $\mathcal{K}$；

表1. **Flooded-Car 研究的方法比较。** 比较各方法在两个阶段的表现，以评估处理流程的整体改进。“-”表示该阶段的指标不可用（不适用、执行失败或未报告）。检索增强型 GPT-5 使用与 ReCCur 相同的人类标注预算。

方法	图像识别与噪声过滤阶段					语义标注阶段			
	精确率	召回率	F1 分数	NRR	CDRR	精确率	召回率	F1 分数	完全匹配
KNN	0.788	0.898	0.832	0.785	0.982	—	—	—	—
K均值算法	0.811	0.683	0.691	0.497	1.000	—	—	—	—
Vo等人 [37]	0.886	0.576	0.685	0.777	0.815	—	—	—	—
YOLO-World v2[6]	0.253	0.129	0.094	0.968	0.032	—	—	—	—
SigLIP2[35]	0.748	0.474	0.393	0.412	1.000	—	—	—	—
CaSED[8]	0.774	0.583	0.481	1.000	0.173	—	—	—	—
Qwen-VL 2.5（指令问答）	0.846	0.799	0.783	0.784	0.831	0.520	0.830	0.506	0.182
GPT-5（指令问答）	0.892	0.863	0.846	0.839	0.886	0.718	0.884	0.776	0.649
检索增强型 GPT-5	0.902	0.885	0.873	0.852	0.879	0.725	0.870	0.801	0.661
ReCCur（我们的方法）	**0.954**	**0.972**	**0.963**	**0.956**	**0.990**	**0.883**	**0.851**	**0.867**	**0.801**

通过多语言、多渠道网络爬取得到 $\mathcal{D}$。在聚类前，应用多模态相似度阈值 $\tau_{\text{mm}} = 0.4$（例如，权重：0.5 图像–描述、0.3 描述–关键词、0.2 图像–关键词）筛选噪声样本。利用多模态数据过滤和人工监督循环，我们手动处理每个簇中的 N_{per}=5 张图像，在提示 $\mathcal{P}_{\text{opt}}$下获得 Π。标签空间 $\mathcal{Y}$ 包含七个进水点（A：安全带，B：座椅滑轨，C：线束，D：备胎槽，E：泡沫垫，F：12V 电源插座，G：转向柱）和一个噪声类别 H。在混合专家蒸馏中，我们设置 $K = 7$、 $\alpha = 20\%$、 $K_L = 3$和 $K_H = 3$，并使用阈值 θ_{t}=0.65 和 θ_{l}=0.45，以得出 π。对于视觉语言模型特征检测， $\mathbf{\Phi}$ 包含“锈蚀”“灰尘和沙土”“霉菌”“老化”和“无”。所有类别和痕迹均由人工标注一次，之后处理流程会生成语义标签。更详细的任务描述请参阅补充材料。

第一阶段进行了十二轮迭代，第二阶段进行了五轮。验证标注由具备汽车专业知识的标注人员按照统一评估标准完成，以确保标签一致并提高标注人员间的一致性。



（a）高相似度（$\geq \tau_{\text{mm}}$）。 （b）低相似度（$< \tau_{\text{mm}}$）。

图7。 **用于数据筛选的聚类结果** 按多模态相似度阈值划分 τ_{mm} 聚类前（HDBSCAN）。**结果。** 在各轮中，第一阶段均实现了清晰的簇间分离和紧凑的簇内结构（图 7），表明该方法有效剪除了偏离主题且证据不足的样本。从定量结果来看，ReCCur 达到了较高的

在保持数据效用的同时实现了较高的识别准确率：NRR 和 CDRR 均超过 0.9（表1），两个阶段中仅有 4% 的样本被升级至人工审核。混淆矩阵呈现明显的对角线优势，非对角线元素总量始终较低（图8），这表明跨类别混淆较少，且在监督有限的情况下表现稳定。

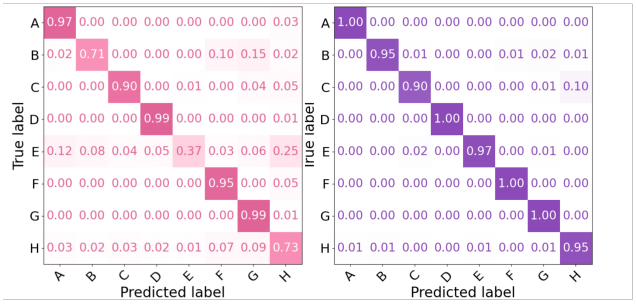


图8。 **Flooded-Car 研究的混淆矩阵** 第一轮（左）和最后一轮（右）。

4.3. 消融实验

我们在架构的每个阶段都进行了消融实验，以验证其有效性。

大规模数据获取与筛选。 在此阶段，我们对多模态相似度筛选进行了消融，移除了多模态相似度和增强的空间表示 $\mathbf{z}_i$，仅使用图像嵌入 $\mathcal{E}_{\text{CLIP}}$ 来验证性能。图 9 显示，消融后这三组数据波动很大，难以收敛。

混合专家知识蒸馏。 在此阶段，我们对以下配置进行消融实验：(i) 单专家（$|\mathcal{M}| = 1$）与多专家（$|\mathcal{M}| = 3$），以及 (ii) 置信度激活的开启与关闭。如表 2 所示，混合专家和置信度激活配置提升了检测性能，并增强了噪声去除能力以及干净数据的保留效果。

表2。 **R的消融研究　　eCCur组件**。 分别分析多专家、置信度激活和区域证据的影响。

模块				图像识别与噪声过滤阶段					语义标注阶段			
MoE	CA	VLM	RE-VLM	精确率	召回率	F1	NRR	CDRR	精确率	召回率	F1	完全匹配
✗	✗	✗	✗	0.807	0.864	0.832	0.904	0.906	—	—	—	—
✓	✗	✗	✗	0.880	0.872	0.876	0.912	0.925	—	—	—	—
✓	✓	✗	✗	0.954	0.972	0.963	0.956	0.990	—	—	—	—
✓	✓	✓	✗	0.954	0.972	0.963	0.956	0.990	0.731	0.812	0.769	0.659
✓	✓	✗	✓	0.954	0.972	0.963	0.956	0.990	0.894	0.859	0.856	0.802

表3. **M　　致幻蘑菇研究的方法比较**。基于GPT的方法与 ReCCur（精简版）　　.

方法	图像识别与噪声过滤			语义标注		
	精确率/召回率/F1值	NRR	CDRR	颜色精确率/召回率/F1值	形状精确率/召回率/F1值	完美
GPT-5	0.312/0.257/0.266	0.236	0.354	0.726/0.748/0.735	0.753/0.701/0.714	0.782
检索增强型 GPT-5	0.569/0.337/0.356	0.138	0.410	0.738/0.735/0.736	0.747/0.692/0.702	0.774
ReCCur（我们的方法）	**0.947/0.941/0.941**	**0.984**	**0.916**	**0.940/0.892/0.913**	**0.816/0.824/0.800**	**0.875**

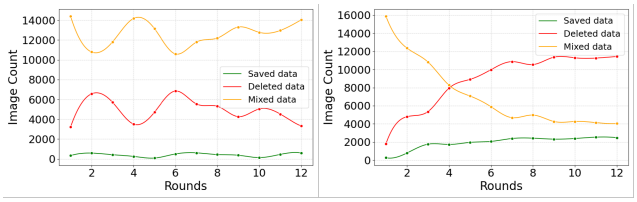


图9。 **消融后（左）与消融前（右）各集合的图像数量对比**

区域证据视觉语言模型对抗式标注。 我们通过比较独立视觉语言模型与加入区域证据的视觉语言模型，对此阶段进行了验证。结果表明，引入区域证据能够增强痕迹定位能力，减轻主体与背景之间的混淆，并生成更准确的语义标签。更详细的任务描述见补充材料。

4.4. 案例研究：致幻蘑菇

背景。 在一些地区，采集和食用野生蘑菇仍很常见。中毒事件经常伴随幻觉——通常被描述为看到“彩虹人影”——而此类事件的大量发生也引起了广泛的网络关注。这进一步凸显了可靠的毒蘑菇识别的必要性。

实现。 为了更好地测试含噪的网络场景，我们进一步构建了一个小型“迷幻蘑菇”子集：许多介绍蘑菇的网页会将真实的有毒物种与迷因风格或高度风格化的图像（鲜艳、彩虹色、“类似 LSD”）混在一起。从视觉上看，这些样本与真实蘑菇相去甚远，但仍会通过“有毒蘑菇”或“魔幻蘑菇”等文本查询检索到，因此是多模态整理中的典型边缘案例。在实验中，我们将蘑菇图像粗略分为四类：(i) 逼真的致命/有毒蘑菇，(ii) 普通/可食用蘑菇，(iii) 有条件可食的蘑菇，以及 (iv) 来自网络的迷幻或迷因风格蘑菇（噪声）图像。我们还使用 Qwen-VL2.5 作为 RE-VLM 后端，以验证第三阶段的通用性。

结果。 仅经过两轮、使用67个人工标签，ReCCur便在所有指标上超越了其他方法（表 3）。在RE-VLM阶段将VLM替换为Qwen-VL2.5后，语义召回率和完全匹配准确率均取得稳定提升，证实了其在低监督条件下与模型无关的通用性。ReCCur无需重新设计即可迁移到蘑菇领域，但其性能取决于VLM嵌入中的痕迹级线索。

4.5. 案例研究：墙体损伤检测

我们进一步在墙体损伤检测任务上测试了 ReCCur；其仍能保持准确性，区分墙体类型，并检测细微痕迹。详情及完整结果请参阅补充材料。

5. 结论

我们提出了**ReCCur**，这是一种模块化整理–标注框架，整合了多模态筛选、专家混合蒸馏和区域证据视觉语言模型验证。在车辆被淹和蘑菇研究中，ReCCur 以极少的人工工作量实现了较高的识别准确率、强大的噪声抑制能力和较高的干净数据保留率。标准化提示词/嵌入与置信度激活相结合，可产生稳定、可审计的决策，并高效地将不确定样本分流。由此得到的数据集呈现出清晰的结构，并支持更高的下游性能，展现了该框架的跨领域适用性和实际可部署性。

ReCCur：一种面向开放与边缘场景下鲁棒视觉语言理解的递归式边界案例筛选框架

补充材料

6. 符号列表

为便于理解后续推导，我们在表 15 中汇总了本节中使用的符号。

7. VLM基线配置

我们所使用的视觉语言模型的详细参数见表 4。

8. 硬件处理速度

表 5 展示了使用免训练 PyTorch 实现时，ReCCur 框架在各种硬件平台上的推理性能。在 Intel i7-11700 CPU 上，ReCCur 达到每秒 1.6 张图像的处理速度，作为基准。NVIDIA RTX 3060 将吞吐量提升至每秒 4.1 张图像，同时占用 5.6GB 内存，展现出消费级GPU上的良好性能。在数据中心级 NVIDIA A10 GPU 上，ReCCur 以 16 的批量大小处理每秒 8.2 张图像，并占用 5.9GB 内存。A10 上有限的吞吐量增益表明，ReCCur 并不严重受限于计算资源，并且 GPU 并行开销极小。

9. Flooded-Car 研究详情

9.1. 任务描述

洪水损坏车辆的金属部件和隐藏隔间会呈现细微的视觉痕迹。在我们的数据集中，每个区域都被标注为五种痕迹类型之一：“锈蚀”、“灰尘和沙土”、“霉菌”、“老化”和“无”。我们将“老化”定义为严重损坏，即可见痕迹覆盖主体结构的 50% 以上。ReCCur 以含噪网页图像作为输入，并通过图 10中的三个阶段，逐步移除离题和不可靠的样本，同时分配这些细粒度痕迹标签以及最终的正常/洪水状态。

9.2. 详细结果

如图 11和表 6所示，ReCCur 在各类别上均取得了良好结果：所有 F1 分数均为 ≥ 0.9。表现最佳的类别是 D（F1= 0.9936）、F（0.9890）和 A（0.9841）。四个类别达到了完美召回率（A/D/F/G = 1.0000）；其中，G 的精确率较低（0.9327），表明存在更多假阳性。表现最弱的是 C 类（F1 = 0.9057），原因是其召回率略低（0.8955）。总体而言，各类别的精确率都较高。

总体性能稳定（Micro F1 = 0.8669，Macro F1 = 0.8564），但类别不平衡现象明显。灰尘/沙子的精确率较高（0.9845），但召回率较低（0.5134），表明遗漏了许多正例。老化的召回率较高（0.9452），但精确率较低（0.7263），这表明模型可能存在过度预测，并与无或锈蚀类别混淆。无的召回率非常高（0.9720），但精确率中等（0.7871），这意味着模型过度使用“无痕迹”标签（表 7）。锈蚀和霉菌仍保持着均衡且良好的表现。这些结果表明，改进灰尘/沙子的精细纹理线索，并针对老化与无类别调整阈值或区域证据，可能有助于改善类别平衡。尽管总体结果令人满意，但分辨率降低可能会压缩特征表达能力，并对性能产生轻微影响；对亚像素痕迹的敏感性仍有提升空间。

如表 8所示，我们最初通过网络爬取得到 17,943 张图像，其中只有约 10% 是真正的目标样本，约 60% 与主题有一定关联，另有 30% 明显偏离主题。这种失衡的构成凸显了建立专门数据整理流程的必要性，以便在下游训练前抑制噪声，并筛选出干净、高精度的数据集。

9.3. 涉水车辆识别的下游评估

为进一步评估我们整理的数据是否确实有益于实际下游应用，我们开展了两项反映实际检查场景的涉水车辆识别任务：（i）涉水汽车部件的多类别分类；（ii）水损痕迹的二元检测。这两项任务均基于正文中介绍的部件与痕迹分类体系，所有结果均采用留出测试集上的 Top-1 准确率进行报告。

实验设置。 对于每项任务，我们采用两种训练条件训练两种常用的分类骨干网络——YOLOv8n-cls[22]和 ResNet-18[17]：使用整理后的干净 数据训练的模型，以及使用网络采集的原始脏数据训练的模型；后者包含质量参差不齐的数据和错误标签。我们的目标是量化数据整理对下游可靠性的影响，尤其是在安全相关场景中。为确保公平比较，同一模型系列中的所有超参数和训练方案均保持一致。

部件分类结果。 表9 显示，在干净数据集上训练时，两种架构都达到了近乎饱和的性能（YOLOv8n-cls：0.9729，ResNet-18：0.9884）。然而，当模型

表4. **视觉语言模型基线配置**。 在相同人工监督预算下，所有基于视觉语言模型的基线所采用的模型、监督方式、温度设置及检索用法汇总。

方法	模型	详情	训练	Temp	检索
基于GPT的问答	GPT-5	问答（单阶段）	No	默认值	No
检索增强型 GPT-5	GPT-5 + CLIP 索引	检索问答	No	默认验证集	Yes
ReCCur（我们的方法）	GPT-5/Qwen-VL 2.5 + CLIP/DINOv2/BEiT	混合专家 + 区域证据	No	默认验证集	Yes

表5. **中央处理器和图形处理器平台上的推理性能比较**。结果使用基于 PyTorch 的免训练模型获得。

设备	框架	批量大小	速度（图像/秒）	GPU内存（GB）	备注
Intel i7-11700 CPU	PyTorch	1	1.6	–	CPU基准性能
Intel i7-11700 CPU+NVIDIA RTX 3060 GPU	PyTorch	16	4.1	5.6	消费级GPU
英特尔至强(Ice Lake) Platinum 8369B+NVIDIA A10 GPU	PyTorch	16	8.2	5.9	数据中心卡

表6. **Flooded-Car 研究的各类别精确率/召回率/F1值**。

类别	精确率	召回率	F1值
A	0.9686	1.0000	0.9841
B	0.9070	0.9512	0.9286
C	0.9160	0.8955	0.9057
D	0.9873	1.0000	0.9936
E	0.9550	0.9745	0.9646
F	0.9781	1.0000	0.9890
G	0.9327	1.0000	0.9652
H	0.9891	0.9569	0.9727

表7。**Flooded-Car 研究的痕迹类型结果**。

痕迹类型	精确率	召回率	F1值
rust	0.9892	0.9387	0.9633
灰尘/沙子	0.9845	0.5134	0.6749
aged	0.7263	0.9452	0.8214
mold	0.9836	0.9231	0.9524
none	0.7871	0.9720	0.8698
宏平均	0.8942	0.8585	0.8564

表8. **洪水损害网页图像的爬取统计**。 该数据集在长尾网页场景中噪声极大。

类别	比例
目标图像	10%
主题相关	60%
无关噪声	30%

在原始噪声数据集上训练时（YOLOv8n-cls：0.8992，ResNet-18：0.8798），准确率明显下降。8–11 个百分点的降幅表明，即使是中等程度的粗粒度部件-

表9。**下游洪水淹没车辆部件分类（Top-1 准确率）**。我们比较使用经过筛选的（“干净”）网络数据与原始含噪（“脏”）网络数据训练的模型。

模型	训练集	Top-1 准确率
YOLOv8n-cls	干净	0.9729
YOLOv8n-cls	脏	0.8992
ResNet-18	干净	0.9884
ResNet-18	脏	0.8798

表10。**下游洪水淹没车辆轨迹二分类（Top-1准确率）**。在整理后的（“干净”）数据与原始噪声（“脏”）数据上训练的任务。

模型	训练集	Top-1 准确率
YOLOv8n-cls	干净	0.9519
YOLOv8n-cls	脏	0.5667
ResNet-18	干净	0.9444
ResNet-18	脏	0.5148

等级识别对标签噪声很敏感，凸显了数据精炼的必要性。

痕迹二分类结果。对于细粒度痕迹检测，噪声的影响显著加剧（表10）。使用干净数据训练时，YOLOv8n-cls 和 ResNet-18 均表现出色（分别为 0.9519 和 0.9444）。相比之下，使用脏数据训练会导致性能降至接近随机水平（0.5667 和 0.5148），这表明噪声监督无法为细微的痕迹线索提供有意义的决策边界。这种显著差距表明，下游安全关键任务需要高质量的标注；仅凭噪声网络数据并不足够。

摘要。 在两种架构和两项任务中，经整理的干净数据都能持续带来显著且可靠的提升。

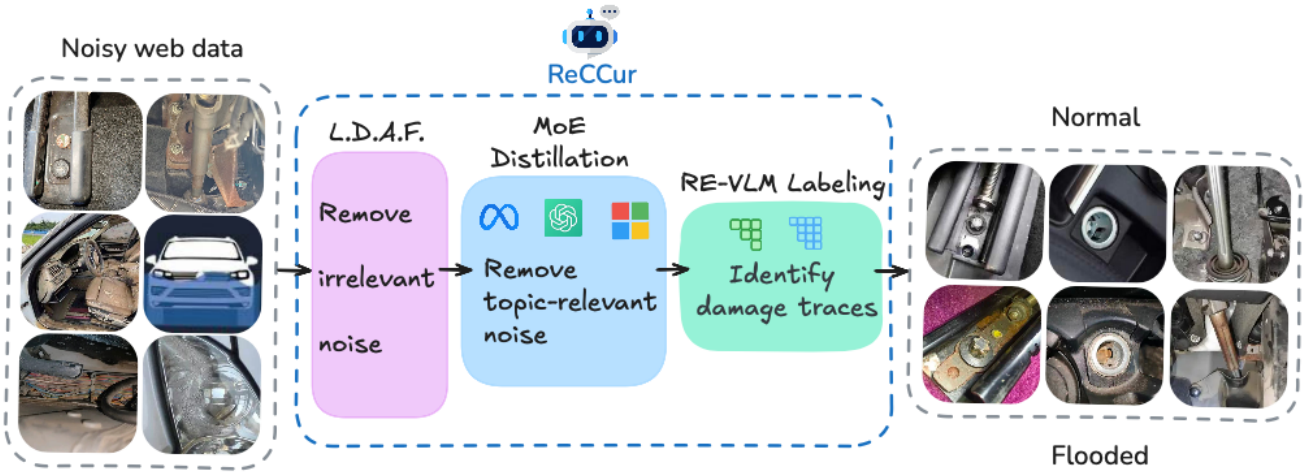


图10。**任务与处理流程详情**。从含噪声的网络爬取图像开始，大规模数据获取与筛选（L.D.A.F.）阶段会剔除视觉上无关和离题的样本。混合专家（MoE）蒸馏阶段会进一步去除与主题相关的噪声，并分配可靠的粗粒度标签。最后，区域证据视觉语言模型（RE-VLM）模块执行细粒度推理，以识别腐蚀、泥浆残留物、生锈螺栓和线束氧化等损伤痕迹，从而使系统能够区分正常车辆与遭受洪水损坏的车辆。

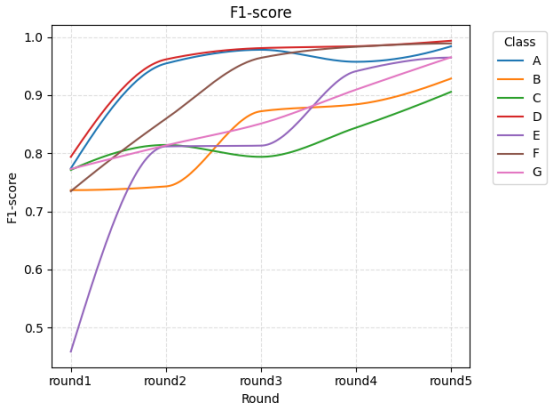


图11。**迭代轮次中的F1分数变化**。各类别（A–G）的F1 分数在五轮迭代中稳步上升。

尤其是在痕迹分类任务中，数据整理将一个几乎无法使用的模型转变为高度可靠的模型。这些结果验证了我们的数据清洗流水线不仅提升了数据集质量指标，还为下游的洪水淹没车辆检查模型带来了实质且显著的收益。

10. 迷幻蘑菇研究详情

10.1. 任务描述

在本实验中，颜色和形状属性均在封闭标签集内定义，以确保一致性和可比性。颜色集合包括白色、红色、黄色，棕色和黑色，代表了蘑菇物种中观察到的主要视觉色调。形状集合涵盖条带状、伞状、球形、不规则和层片状，分别对应蘑菇菌盖和菌柄最常见的形态模式。所有预测都限定在这些类别中，同义词或变体术语（例如，发白的 →白色，菌盖 → 伞状）会在评估过程中自动归一化，以保持标签的一致性。借助 ReCCur，我们进一步将原始网络爬取数据划分为这些语义类别，并赋予明确的毒性和属性标签。如图12所示，蘑菇数据集在所有四个类别中都包含真实样本和风格化样本，因此这项任务在视觉上模棱两可，容易受到噪声引发的混淆。

我们进行了两轮大规模数据筛选。由于蘑菇图像的语义差异相对细微，且网络上有大量有效样本，因此只需少量筛选轮次，即可获得噪声极少的数据集。

10.2. 详细结果

经过两阶段筛选后，我们对精炼数据集进行了聚类，聚类图像的可视化结果如图13。

如图 14所示，少量可食用样本被误分类为噪声，导致数据保留率略有下降。表 11 展示了各组的详细信息。



图12. **蘑菇类别示例**。四个类别的代表性样本：有条件可食、致命、可食用以及噪声/风格化。真实图像与风格化图像之间的视觉相似性，体现了该任务的挑战。

10.3. 致幻蘑菇分类的下游评估

迷幻蘑菇领域对于数据整理和下游识别而言，是一个特别具有挑战性的场景。与被洪水淹没的汽车数据集或墙体损坏数据集不同，从网络收集的蘑菇图像在风格上表现出很大的差异：真实照片、野外指南扫描图、高饱和度插图、人工智能生成的艺术作品，以及类似“迷幻”风格的表情包图像经常共存。

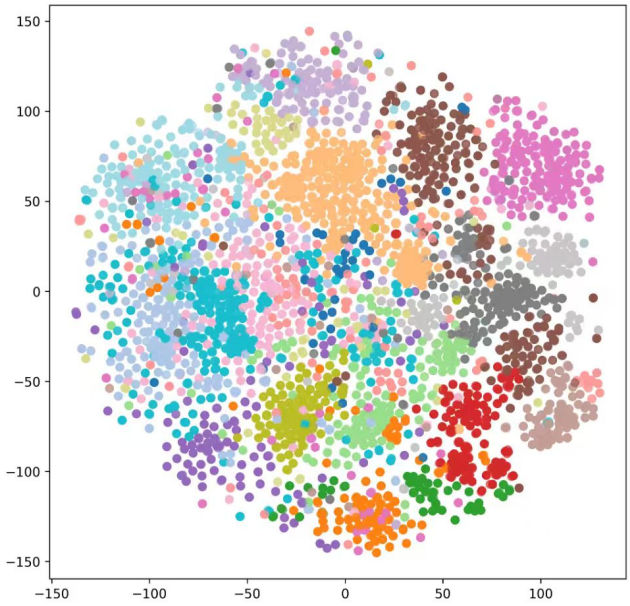


图13。**经过两轮筛选后，蘑菇数据集中的聚类样本**。

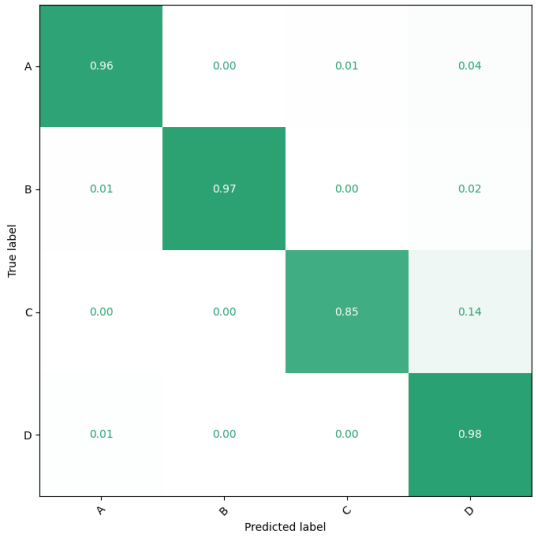


图14. **蘑菇分类的归一化混淆矩阵**。大多数样本位于对角线上，类别 C 和 D 之间存在少量混淆。（A：有条件可食，B：有毒，C：可食用，D：噪声）

尽管许多图像属于本质上不同的语义类别（真实与伪造），但它们具有相似的整体色彩和纹理模式，因此该领域是理想的测试平台，可用于评估 ReCCur 能否缓解噪声引起的歧义，并支持可靠的下游分类。

任务动机。在实际应用中，与蘑菇相关的图像

表11. **蘑菇数据集分类结果**

类别	精确率	召回率	F1值
A	0.9786	0.9567	0.9675
B	0.9961	0.9716	0.9837
C	0.9926	0.8522	0.9170
D	0.8209	0.9842	0.8951
宏平均	0.9470	0.9412	0.9409

集合常受到合成或风格化内容的污染，包括卡通图像、奇幻插画以及经过大量编辑、无法反映真实形态的图像。无论是科学分析还是数据集构建，区分真实蘑菇照片与非摄影或虚假图像都很重要。同时，我们也关注：当训练数据中存在此类虚假内容时，经过清理的数据是否有助于模型保留不同蘑菇类别之间的细粒度差异。因此，我们评估两类下游任务，以衡量 ReCCur 精选数据的实际效果。

实验设置。 我们考虑两项互补的下游任务：

• **多类别识别：** 对有条件可食、致命、可食用和噪声/风格化样本进行四类别分类。

•**真假二分类识别：** 对真实蘑菇照片与虚假/非照片类图像（包括卡通图像、人工智能生成图像以及高度风格化或模因类内容）进行二分类。这反映了区分真实生物图像与合成或插画类内容的能力。

对于这两项任务，我们使用两种有代表性的分类骨干网络——YOLOv8n-cls[22]和 ResNet-18[17]，在两种数据设置下进行训练：(i) 我们通过 ReCCur 筛选整理的干净数据集，以及 (ii) 原始脏网络爬取数据集，其中包含错误标注样本，且风格化或虚假图像占比较大。为确保公平比较，同一架构下的超参数、数据增强和训练计划均保持一致。所有指标均以留出的干净测试集划分上的 Top-1准确率报告。

多类别识别结果。表12 报告了四分类设置下的 Top-1 准确率。在干净数据上训练能显著提升模型的可靠性：YOLOv8n-cls 的准确率从 0.852（脏数据）提高到 0.948（干净数据），ResNet-18 则从 0.875 提高到 0.964。这些提升表明，原始网络数据仍会引入不可忽视的类别间混淆，尤其是在可食用、条件可食用和迷幻风格图像之间；这些图像的颜色和形状模式与真实物种相似。通过移除含糊、无关主题和风格化程度过高的样本，ReCCur 使两种架构都能形成更清晰的决策边界，并减少

表12. **下游蘑菇多类别分类（Top-1准确率）**。

模型	训练集	Top-1 准确率
YOLOv8n-cls	干净	0.948
YOLOv8n-cls	脏	0.852
ResNet-18	干净	0.964
ResNet-18	脏	0.875

表13. **下游蘑菇真伪分类（Top-1准确率）**。

模型	训练集	Top-1 准确率
YOLOv8n-cls	干净	0.972
YOLOv8n-cls	脏	0.914
ResNet-18	干净	0.981
ResNet-18	脏	0.923

风格偏差导致的系统性错误标注。

真实与虚假二分类结果。噪声会明显影响真实与虚假分类任务；在该任务中，模型必须区分逼真的生物图像与卡通图像、人工智能生成艺术以及经过大量编辑的“迷幻风格”内容。如表13所示，在干净数据上训练可获得非常出色的性能，YOLOv8n-cls 和 ResNet-18 的 Top-1 准确率分别为 0.972 和 0.981。在脏数据集上训练时，这两种架构的准确率仍然很高（分别为 0.914 和 0.923），但与使用干净数据训练的模型相比仍低约 6 个百分点。这一持续存在的性能下降表明，训练集中的混杂、未标注虚假内容会扰动真实图像与非真实图像之间的决策边界，导致风格化蘑菇与真实标本之间出现更多混淆。ReCCur 的精选数据集通过在训练前筛除非摄影图像或语义不一致样本，缓解了这一问题。

总结。 在多类别和真假分类任务中，精选数据都能为 YOLOv8n-cls 和 ResNet-18 带来稳定且显著的提升。在真假分类设置下，ReCCur 还能让原本表现强劲的模型进一步接近性能饱和；与使用脏数据训练相比，错误率降低了一半以上。这些发现进一步表明，在风格噪声严重且包含合成内容的领域，结构化多模态筛选至关重要，而精选数据带来的益处也会直接传递到下游的蘑菇相关识别任务中。

11. 案例研究：墙体损伤检测

背景。 墙体老化是场景理解和缺陷检测中一种罕见且代表性不足的边缘案例，其表现形式多样（污渍、泛碱、剥落、微裂纹）。此类现象样本稀缺且标签噪声较大，会导致分布偏移——在完好表面上训练的模型会过拟合纹理/颜色先验——因此，对老化制品的鲁棒性仍缺乏充分研究，却又不可或缺。

实现细节。 我们基于 S2DS 墙体数据集 [3] 并应用 ReCCur 扩充数据，生成四种额外的墙体材料领域：砖、花岗岩、金属和木材，同时标注噪声样本。经过三轮识别与筛选后，我们获得最终数据集。所有设置均与洪水受损车辆实验中的设置保持一致。

结果。 在这一场景中，ReCCur 在数据分类和语义标注方面均表现出色，始终优于基于 GPT 的基线方法。增加细化轮次可进一步提高准确率，证实了 ReCCur 的有效性。归一化混淆矩阵显示，数值主要集中在对角线上，只有少量泄漏至非对角线，表明各类别区分良好。跨方法比较进一步表明，ReCCur 在精确率/召回率/F1 方面优于基线方法（表14）。结果显示，所有方法在检测大规模痕迹方面的表现都优于识别细微或小规模痕迹。得益于两阶段推理设计，与其他方法相比，ReCCur 的整体表现略胜一筹。

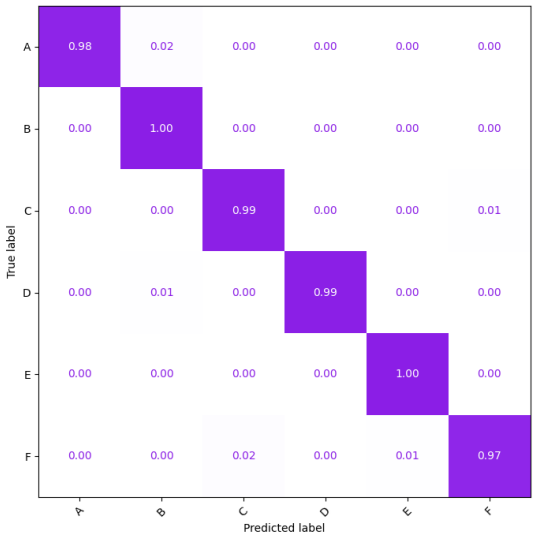


图 15. **墙体分类的归一化混淆矩阵。** A：砖，B：花岗岩，C：金属，D：混凝土，E：噪声，F：木材。

12. 局限性与未来工作

尽管 ReCCur 表明，免训练核心的多阶段数据整理流水线能够以适度的计算资源显著提升噪声网络数据的纯度和语义丰富度，但仍存在若干局限。

领域与模态范围。 我们的实证验证聚焦于三个案例研究（洪水受损车辆检查、迷幻蘑菇和墙体损伤检测），且均处于图像领域。尽管这些场景已呈现显著的长尾噪声和语义歧义，但尚未涵盖视频、3D 或时间维度的边缘案例，也未涉及高度交互的具身场景。将 ReCCur 扩展至多帧证据、深度/3D 表征和序列现象（例如渐进式损伤、时变场景）是下一步的重要工作。

对上游模型的依赖。 该处理流程依赖预训练视觉语言模型和视觉编码器（例如 ChatGPT-5、Qwen-VL 2.5、CLIP、DINOv2、BEiT），用于关键词扩展、多模态相似度评分和语义推理。因此，其失效模式会继承这些模型的偏见和盲点，并且在预训练语料库中代表性不足的领域，性能可能会下降。未来的工作包括使用更强和更弱的骨干网络对 ReCCur 进行压力测试、引入轻量级领域自适应，以及显式建模上游模型带来的认知不确定性。

残余噪声与覆盖缺口。 尽管 ReCCur 在我们的基准测试中实现了较高的精确率和较强的干净数据保留率，但它无法保证标签完全不含噪声，也无法穷尽覆盖所有罕见现象。具体而言，轨迹级模式（例如细微的灰尘/沙子、精细纹理伪影）仍然难以处理，而当前的置信度激活和不确定性路由阈值是通过启发式方法设定的。一个有前景的方向是将 ReCCur 与形式化校准目标及筛选质量的概率保证相结合，并根据下游性能进行自适应阈值设定。

人工监督与评估设计。 人机协同组件经过有意简化，但仍需要具备领域知识的标注人员（例如汽车领域专家），而这类人员可能并非普遍可得。此外，尽管我们确实提供了下游评估来验证精选数据的实际价值，但分析范围仍然有限。我们尚未将 ReCCur 精选的数据与更广泛的其他过滤基线或大型工业系统进行系统比较，而且下游任务仅涵盖了可能的现实世界应用中的一部分。未来工作包括整合...

表14. **墙体损伤研究的方法比较**。基于GPT的方法与 ReCCur 的比较。

方法	图像识别与噪声过滤阶段					语义标注阶段			
	精确率	召回率	F1 分数	NRR	CDRR	精确率	召回率	F1 分数	完全匹配
GPT-5	0.889	0.865	0.883	0.824	0.901	0.814	0.869	0.822	0.772
检索增强型 GPT-5	0.921	0.903	0.896	0.897	0.924	0.817	0.861	0.828	0.786
ReCCur（我们的方法）	0.968	0.965	0.937	0.965	0.984	0.862	0.887	0.863	0.854

将 ReCCur 集成到端到端训练处理流程中，衡量其对安全关键任务的下游影响，并探索让非专家能够提供可靠反馈的接口。

系统性与分布性偏差。由于我们的网络爬取数据反映了现有的在线内容，因此可能继承地域、社会经济或审美偏差（例如，某些汽车车型、建筑材料或蘑菇物种被过度代表）。ReCCur 目前能够改善标签质量，但尚未明确纠正这类不平衡。一个重要的扩展方向是，将兼顾公平性与多样性的目标整合到采集（关键词扩展、爬取）和筛选（采样、聚类选择）过程中，并为用户提供用于审计和重新平衡所整理数据集的工具。

13. 伦理与社会影响

ReCCur 被设计为一个面向边缘案例的通用型、低计算资源数据整理框架，旨在提升安全相关应用中的鲁棒性和透明度。我们的案例研究聚焦于具有明确消费者或公共安全动机的领域：涉水车辆检查、毒蘑菇识别和墙体损伤检测。在这些场景中，更高质量的边缘案例数据集能够通过支持更优的检测模型和更可靠的诊断工具，帮助降低安全风险（例如涉水车辆被转售、意外中毒或结构性失效）。

与此同时，系统地挖掘、筛选并对罕见视觉模式进行语义丰富化的能力也带来了双重用途风险。原则上，类似的处理流程可能被重新用于整理敏感或涉及隐私的数据，或构建有助于广泛监控或定向用户画像的数据集。我们的设计选择通过以下方式在一定程度上缓解了这些风险：(i) 将实验限制在明确的任务聚焦领域；(ii) 强调透明、基于证据的标签标注，而非隐蔽的用户画像；以及 (iii) 主要利用人工监督来验证技术标签（例如部件、痕迹、材料类型），而不是验证个人属性。

我们实验中使用的所有数据均来自公开可访问的网络内容或现有数据集，且我们避免爬取明显属于私人或受访问控制的来源。我们不会尝试推断个人的人口统计或敏感属性，也不会发布任何身份标注。今后若公开发布精选数据集，将排除主要描绘可识别个人的图像；对于偶然包含的个人身份信息（如面部或车牌），我们将尽最大努力进行模糊处理或遮挡。我们明确反对将 ReCCur 或其筛选数据集用于违反隐私、监控或数据保护相关地方法规的应用。

最后，应将我们的工作视为数据整理方面的早期研究，而非可直接部署的现成系统。任何现实场景部署——尤其是在保险、消费者保护、基础设施检查或医疗保健等受监管且高风险的领域——都必须经过单独的风险评估、领域特定验证以及法律与伦理审查。从业者仍有责任确保其对 ReCCur 及相关数据集的使用符合适用法律、机构审查标准和社会规范。

14. 许可与法律合规

论文被接收后，我们将在明确的许可和合规保障措施下，公开发布 ReCCur 代码和精选数据集，供研究和商业用途。完整的训练、评估和推理代码将以宽松的开源许可（例如 Apache 2.0）托管在 GitHub 上，允许用户将 ReCCur 用于学术项目和商业产品，并对其进行修改和集成。

ReCCur 筛选数据集将依据专门的“ReCCur 数据许可协议”发布，允许非商业和商业用途、修改及再分发，但用户必须：(i) 适当注明我们的工作；(ii) 将任何衍生数据集按照相同的数据许可条款再分发；以及 (iii) 遵守其所在司法管辖区内所有适用的版权、隐私和数据保护法。该许可将明确禁止将数据用于生物识别、大规模监控，或任何其他可能威胁个人隐私或公民自由的用途。

我们的数据整理流程基于第三方模型和数据集（例如 CLIP、DINOv2、BEiT、视觉语言模型 API 以及现有的缺陷数据集），并可能包含网络爬取图像。对于每项上游资源，我们将记录其来源、

版本和原始许可证，并在“许可证与归属”部分予以说明。仅当第三方内容的许可证允许时，我们才会重新分发该内容；否则，我们将提供下载链接和预处理脚本，以便用户自行复现我们的数据集。对于网络爬取内容，我们将精选数据视为衍生研究成果，并尽可能存储来源元数据，同时提供明确的下架渠道：经核实的原始内容所有者请求，将导致相应样本在后续版本中被移除。

ReCCur 或其衍生版本在现实世界决策处理流程中的任何部署，都必须遵守相关法律框架（例如数据保护、消费者保护、产品责任和行业专项法规）。因此，我们发布时将包含免责声明，声明作者和所属机构不对代码或数据集的滥用或非法应用承担任何责任，下游用户应对确保其所在司法管辖区内的法规合规承担全部责任。

参考文献

[1] Zhaochong An、Guolei Sun、Yun Liu、Runjia Li、Min Wu、Ming-Ming Cheng、Ender Konukoglu 和 Serge Belongie。多模态有助于少样本 3D 点云语义分割。载于 第 *13* 届国际学习表征会议论文集（*ICLR*），第 50163–50184 页，新加坡，2025年。国际学习表征会议。2[2] Andreas B¨ar、Neil Houlsby、Mostafa Dehghani 和 Manoj Kumar。面向少样本图像分类的冻结特征增强。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 16046–16057 页，2024年。2[3] Christian Benz 和 Volker Rodehorst。利用分层多尺度注意力进行基于图像的结构缺陷检测。载于 *DAGM*德国模式识别会议（*GCPR*），第 337–353 页。施普林格，2022年。6[4] Xiaohe Bo、Zeyu Zhang、Quanyu Dai、Xueyang Feng、Lei Wang、Rui Li、Xu Chen 和 Ji-Rong Wen。基于大语言模型的反思式多智能体协作。神经信息处理系统进展，37：138595–138631，2024年。3[5] Justin Chih-Yao Chen、Swarnadeep Saha 和 Mohit Bansal。Reconcile：圆桌会议通过多样化大语言模型之间的共识提升推理能力。*arXiv* 预印本 *arXiv:2309.13007*，2023年。3[6] Tianheng Cheng、Lin Song、Yixiao Ge、Wenyu Liu、Xing-gang Wang 和 Ying Shan。YOLO-World：实时开放词汇目标检测。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2024年。7[7] Hyeong Kyu Choi、Xiaojin Zhu 和 Yixuan Li。辩论还是投票：在多智能体大语言模型中，哪种方式能产生更好的决策？载于 神经信息处理系统进展，2025年。NeurIPS 2025，重点报告。3

[8] Alessandro Conti、Enrico Fini、Massimiliano Mancini、Paolo Rota、Yiming Wang 和 Elisa Ricci。无词汇图像分类。载于《神经信息处理系统进展》（NeurIPS），2023年。7[9] Arnav M Das、Gantavya Bhatt、Lilly Kumari、Sahil Verma 和 Jeff Bilmes。Cobra：用于少样本适应的组合检索增强。在计算机视觉与模式识别会议论文集上，第20534–20546页，2025年。2[10] Andrew Estornell 和 Yang Liu。多大语言模型辩论：框架、原则与干预措施。载于《神经信息处理系统进展》，第28938–28964页，2024年。3[11] Talfan Evans、Nikhil Parthasarathy、Hamza Merzi´c 和 Olivier J. H´enaff。通过联合选择样例进行数据整理，进一步加速多模态学习。载于《神经信息处理系统进展》，2024年。NeurIPS 2024，数据集与基准赛道（重点报告）。2[12] Zhaopeng Feng、Jiayuan Su、Jiamei Zheng、Jiahan Ren、Yan Zhang、Jian Wu、Hongwei Wang 和 Zuozhu Liu。M-MAD：用于高级机器翻译评估的多维多智能体辩论。载于第63届计算语言学协会年会论文集（第1卷：长篇论文），第7084–7107页，维也纳，奥地利，2025年。计算语言学协会。3[13] Minghao Fu 和 Ke Zhu。面向实用小样本识别的基于实例的最大间隔。载于IEEE/CVF计算机视觉与模式识别会议论文集，第28674–28683页，2024年。2[14] Samir Yitzhak Gadre、Gabriel Ilharco、Alex Fang、Jonathan Hayase、Georgios Smyrnis、Thao Nguyen、Ryan Marten、Mitchell Wortsman、Dhruba Ghosh、Jieyu Zhang 等。Datacomp：寻找下一代多模态数据集。《神经信息处理系统进展》，36：27092–27112，2023年。2[15] Zhaopeng Gu、Bingke Zhu、Guibo Zhu、Yingying Chen、Ming Tang 和 Jinqiao Wang。UniVAD：用于少样本视觉异常检测的免训练统一模型。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第15194–15203页。IEEE，2025年。2[16] Guangxing Han 和 Ser-Nam Lim。基于基础模型的少样本目标检测。载于IEEE/CVF计算机视觉与模式识别会议论文集，第28608–28618页，2024年。2[17] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于IEEE计算机视觉与模式识别会议（CVPR）论文集，第770–778页，2016年。1, 5[18] Sirui Hong、Mingchen Zhuge、Jonathan Chen、Xiawu Zheng、Yuheng Cheng、Jinlin Wang、Ceyao Zhang、Zili Wang、Steven Ka Shing Yau、Zijuan Lin、Liyang Zhou、Chenyu Ran、Lingfeng Xiao、Chenglin Wu 和 J¨urgen Schmidhu-ber。MetaGPT：面向多智能体协作框架的元编程。载于第十二届国际学习表征会议（ICLR 2024），2024年。3[19] Tianyu Hu、Zhen Tan、Song Wang、Huaizhi Qu 和 Tian-long Chen。面向大语言模型评审的多智能体辩论，具有自适应稳定性检测。

表15。**符号列表**

符号	类型/空间	含义
数据集/集合		
$\mathcal{D}$	$\mathcal{P}(\mathbb{R}^{H\times W\times 3})$	原始大规模网络爬取数据集
$\mathcal{D}_c$	$\mathcal{P}(\mathbb{R}^{H\times W\times 3})$	c类别的数据集
$\mathcal{D}_{\text{tri}}$	$\mathcal{P}\big(\mathbb{R}^{H\times W\times 3}\times\Sigma^*\times\Sigma^*\big)$	三模态样本三元组集合 $\{\mathbf{u}_i\}_{i=1}^{N}$
Π	$\mathcal{P}(\mathbb{R}^{H\times W\times 3})$	多模态筛选后的精炼数据集
$\mathcal{C}=\{\mathcal{C}_1,\dots,\mathcal{C}_j\}$	$\mathcal{P}\big(\mathcal{P}(\mathbb{R}^{1536})\big)$	通过对嵌入进行分组得到的聚类
$\mathcal{A}=\{(X_i,y_i)\}$	$\mathcal{P}\big(\mathbb{R}^{H\times W\times 3}\times\mathcal{Y}\big)$	人工标注参考数据集
$\mathcal{I}$	$(\mathcal{P}(\mathcal{Y}\times\mathbb{R}^{d_{768}}))\times(\mathcal{P}(\mathcal{Y}\times\mathbb{R}^{d_{1024}}))\times(\mathcal{P}(\mathcal{Y}\times\mathbb{R}^{d_{1024}}))$	由三个专家子数据库构成的向量索引数据库
π	$\mathcal{P}(\mathbb{R}^{H\times W\times 3})$	迭代后的最终整合数据集
$\mathcal{I}_{\text{CLIP}}$	$\mathcal{P}\big(\mathcal{Y}\times\mathbb{R}^{d_{768}}\big)$	CLIP 子数据库（标签-向量对）
$\mathcal{I}_{\text{DINOv2}}$	$\mathcal{P}\big(\mathcal{Y}\times\mathbb{R}^{d_{1024}}\big)$	DINOv2 子数据库（标签-向量对）
$\mathcal{I}_{\text{BEiT}}$	$\mathcal{P}\big(\mathcal{Y}\times\mathbb{R}^{d_{1024}}\big)$	BEiT 子数据库（标签-向量对）
样本 / 变量		
X_i	$\mathbb{R}^{H\times W\times 3}$	第 i 张图像张量
H_i	Σ^*	X_i 的 VLM 生成的文本描述
Q_i	Σ^*	搜索查询关键词
$\mathbf{u}_i=(X_i,H_i,Q_i)$	$\mathbb{R}^{H\times W\times 3}\times\Sigma^*\times\Sigma^*$	三模态样本三元组
$\mathbf{z}_i$	$\mathbb{R}^{1536}$	语义增强嵌入（CLIP-img ‖ CLIP-txt）
r_i	$\{0,1\}$	二元相关性标签（1=高， 0=低）
$\bar{r}_j$	$[0,1]$	平均簇相关性得分
y_i	$\mathcal{Y}$	真实类别标签
$\hat{y},\ \hat{y}^{(m)}$	$\mathcal{Y}$	预测标签（集成 / 专家 m）
$\hat{y}_{\text{sem-1}},\hat{y}_{\text{sem-2}},\hat{y}_{\text{sem}}$	$\mathcal{Y}\times\mathcal{P}(\mathbf{\Phi})$	推理阶段的语义标签
$\mathcal{Y}$	-	Dec(X) 使用的扩展标签空间
提示词 / 关键词		
$\mathcal{P}_{\text{init}},\ \mathcal{P}_{\text{opt}},\ \mathcal{P}_c$	Σ^*	初始 / 优化后 / 类别提示词
$\mathcal{K}_{\text{V}},\ \mathcal{K}_{\text{T}},\ \mathcal{K}$	Σ^*	视觉语言模型关键词 / 纯文本关键词 / 最终语料库
$\mathcal{G}_{\text{V}}(\cdot)$	$\mathcal{P}(\mathbb{R}^{H\times W\times 3})\times\Sigma^*\to\Sigma^*$	基于视觉语言模型的关键词生成
$\mathcal{G}_{\text{T}}(\cdot)$	$\Sigma^*\to\Sigma^*$	纯文本关键词生成
$\mathcal{L}(\cdot)$	$\Sigma^*\to\Sigma^*$	多语言词典扩展
模型 / 函数		
$\mathcal{E}_{\text{CLIP}},\ \mathcal{E}_{\text{DINOv2}},\ \mathcal{E}_{\text{BEiT}}$	$\mathbb{R}^{H\times W\times 3}\to\mathbb{R}^{d_m};\ \Sigma^*\to\mathbb{R}^{d_m}$	专家级图像/文本编码器
$\mathcal{F}(\cdot)$	$\mathbb{R}^3\to\mathbb{R}$	多模态相似度融合算子
相似度$_{\text{mm}}(\mathbf{u}_i)$	$\mathbb{R}$	多模态相似度分数
提议器$(X,\ \mathcal{P}_c)$	$\mathbb{R}^{H\times W\times 3}\times\Sigma^*\to(\mathbb{Z}^4)^{N_g}\times\{0,1\}^{N_g}$	视觉语言模型提议器（区域提取和特征标记）
$\psi(\mathbf{r}_c^{(j)},\ g,\ \mathbf{\Phi})$	$\{0,1\}$	区域属性存在指示器
Dec(X)	$\mathbb{R}^{H\times W\times 3}\to(\mathcal{Y}\cup\{\text{非目标}\})$	基于置信度阈值的最终决策规则
专家 / 嵌入 / 相似度		
$\mathcal{M}$	{CLIP,DINOv2, BEiT}	专家集成
$\mathcal{N}^{(m)}=\big((\xi_k^{(m)},\ y_k^{(m)})\big)_{k=1}^{K}$	$\big([-1,1]\times\mathcal{Y}\big)^K$	Top-K 邻居相似度–标签序列（按 ξ排序）
ξ_k	$[-1,1]$	余弦相似度值
$\overline{\text{FAS}}(X,\ c_k),\ \overline{\text{FAS}}(X,\ \hat{y})$	$(\mathbb{R}^{H\times W\times 3}\times\mathcal{Y})\to[-1,1]$	特征对齐分数（按类别 / 平均）
$\mu_c^{(m)}$	$\mathbb{R}^{d_m}$	专家 m下类别 c 的平均嵌入
置信度 / 阈值 / 分数		
置信$^{(m)}$ 度(X)	$[-1,1]$	专家 m的标签置信度
置信度$_{\text{t}}(X),\ \overline{\text{置信度}}_{\text{l}}(X)$	$[-1,1]$	主题置信度和标签置信度
$\theta_{\text{t}},\ \theta_{\text{l}}$	$[-1,1]$	置信度阈值
τ_{mm}	$\mathbb{R}$	多模态相似度筛选的阈值
$B(X)$	$[-1,1]$	边界强度（最大对齐度）
$y^*(X)$	$\mathcal{Y}$	边界归因类别
区域 / 特征		
$\mathcal{R}_{c,g}=\{\mathbf{r}_c^{(1)},g,\ \dots\dots,\ \mathbf{r}^{(N_g)_{c,g}}\}$	$(\mathbb{Z}^4)^{N_g}$	粒度为 g 的子区域边界框（每个以像素为单位表示为 (x,y,w,h)；通常为 $N_g=g^2$）
$\mathbf{r}_{c,g}^{(j)}$	$\mathbb{Z}^4$	第j个子区域框（$x,\ y,\ w,\ h$），单位为像素
N_g	$\mathbb{N}$	粒度为 g 时的子区域数量（针对 $g\times g$ 网格， $N_g=g^2$）
$\mathbf{\Phi}$	–	视觉属性集（例如：锈蚀、霉菌、泥浆）
$\mathcal{R}_{s,g}=\{\text{递归}_s^{(1)},g,\ \dots\dots,\ \text{递归}^{(N_g)_{s,g}}\}$	$(\mathbb{Z}^4)^{\ N_g}$	粒度为 g 时的主体子区域集（$\mathcal{R}_{s,\ g}\subset\mathcal{R}_{c,\ g}$）
$\mathbf{r}_{s,g}^{(j)}$	$\mathbb{Z}^4$	第 j 个主体子区域框（$x,\ y,\ w,\ h$）
标量 / 参数		
$K,\ K_L,\ K_H$	$\mathbb{N}$	邻居数 / 候选池大小
α	$(0,1]$	低对齐池的采样比例
$N,\ N_c$	$\mathbb{N}$	数据集大小 / 每类样本数
N_{per}	$\mathbb{N}$	每个簇中手动标注的图像数量
t	$\mathbb{N}$	推导中使用的类别数量（例如，边界排序）
T	$\mathbb{R}_{>0}$	Softmax 温度参数

应稳定性检测。载于神经信息处理系统进展，2025年。NeurIPS 2025，海报。3[20] Xilang Huang 和 Seon Han Choi。面向小样本开放集识别的特征语义增强网络。*Pattern Recognition*，156：110781，2024年。2[21] Ahmet Iscen、Alireza Fathi 和 Cordelia Schmid。通过从网络规模图文数据中检索来改进图像识别。载于*IEEE/CVF*计算机视觉与模式识别会议，第19295 19304页，2023年2–。

[22] Glenn Jocher、Ayush Chaurasia 和 Jing Qiu。Ultralyt- ics yolov8。https://github.com/ultralytics/ultralytics，2023年。访问日期：2025-11-18。1，5[23] Wooyoung Kang、Jonghwan Mun、Sungjun Lee 和 Byungseok Roh。基于网络爬取的图文数据进行噪声感知学习，用于图像描述生成。载于*IEEE/CVF* 国际计算机视觉会议论文集，第2942–2952页，2023年。2[24] Dayeon Ki、Rachel Rudinger、Tianyi Zhou 和 Marine Carpuat。多个大语言模型智能体通过辩论实现公平的文化

对齐。载于《计算语言学协会第63届年会论文集》（第1卷：长篇论文），第24841–24877页，奥地利维也纳，2025年。计算语言学协会。3[25] Chris Dongjoo Kim、Jihwan Moon、Sangwoo Moon、Heese- ung Yun、Sihaeng Lee、Aniruddha Kembhavi、Soonyoung Lee、Gunhee Kim、Sangho Lee 和 Christopher Clark。Re-spec：面向视频-文本数据流学习的、基于相关性与特异性的在线筛选。载于计算机视觉与模式识别会议论文集，第29040–29049页，2025年。2[26] Tian Liang、Zhiwei He、Wenxiang Jiao、Xing Wang、Yan Wang、Rui Wang、Yujiu Yang、Shuming Shi 和 Zhaopeng Tu。通过多智能体辩论鼓励大语言模型进行发散性思维。载于2024年自然语言处理经验方法会议论文集，第17889–17904页。计算语言学协会，2024年。3[27] Chang Liu、Giulia Rizzoli、Pietro Zanuttigh、Fu Li 和 Yi Niu。从网络学习：语言驱动用于语义分割的弱监督增量学习。载于欧洲计算机视觉会议论文集，第352–369页。施普林格，2024年。2[28] Tian Liu、Huixin Zhang、Shubham Parashar 和 Shu Kong。通过分阶段检索增强微调实现少样本识别。载于计算机视觉与模式识别会议论文集，第15086–15097页，2025年。2[29] S´egol`ene Martin、Yunshi Huang、Fereshteh Shakeri、Jean- Christophe Pesquet 和 Ismail Ben Ayed。传导式零样本和少样本CLIP。载于IEEE/CVF计算机视觉与模式识别会议论文集，第28816–28826页，2024年。2[30] Muhammad Anwar Ma’sum、Mahardhika Pratama、Lin Liu、Habibullah Habibullah 和 Ryszard Kowalczyk。联邦小样本类增量学习。载于第13届国际学习表征会议（ICLR）论文集，第101524–101553页，新加坡，2025年。国际学习表征会议。2[31] Naeem Paeedeh、Mahardhika Pratama、Muhammad Anwar Ma’sum、Wolfgang Mayer、Zehong Cao 和 Ryszard Kowl- czyk。通过自适应Transformer网络实现跨域小样本学习。《知识型系统》，288卷，111458，2024年。2：。[32] Rashindrie Perera 和 Saman Halgamuge。用于跨域小样本学习的判别式样本引导型参数高效特征空间自适应。载于IEEE/CVF计算机视觉与模式识别会议论文集，第23794–23804页，2024年。2[33] Ziheng Qin、Zhaopan Xu、Yukun Zhou、Zangwei Zheng、Zebang Cheng、Hao Tang、Lei Shang、Baigui Sun、Xiaojiang Peng、Radu Timofte等。数据集增长。载于欧洲计算机视觉会议论文集，第58–75页。施普林格，2024年。2[34] Shuai Shao、Yu Bai、Yan Wang、Baodi Liu 和 Yicong Zhou。DEIL：面向开放世界少样本学习的直接与逆向CLIP。载于IEEE/CVF计算机视觉与模式识别会议论文集，第28505 28514页，2024年。2 – 。

[35] Michael Tschannen、Alexey Gritsenko、Xiao Wang、Muham- mad Ferjad Naeem、Ibrahim Alabdulmohsin、Nikhil Parthasarathy、Talfan Evans、Lucas Beyer、Ye Xia、Basil Mustafa、Olivier H´enaff、Jeremiah Har msen、Andreas Steiner 和 Xiaohua Zhai。《Siglip 2：具备更强语义理解能力、定位能力和密集特征的多语言视觉语言编码器》。arXiv 预印本 arXiv:2502.14786，2025年。7[36] Vishaal Udandarao、Nikhil Parthasarathy、Muhammad Fer- jad Naeem、Talfan Evans、Samuel Albanie、Federico Tombari、Yongqin Xian、Alessio Tonioni 和 Olivier J H´enaff。《主动数据整理可有效提炼大规模多模态模型》。载于《计算机视觉与模式识别会议论文集》，第14422–14437页，2025年。2 – .[37]Huy V Vo、Vasil Khalidov、Timoth´ee Darcet、Th´eo Moutakanni、Nikita Smetanin、Marc Szafraniec、Hugo Tou- vron、Camille Couprie、Maxime Oquab、Armand Joulin 等。《自监督学习的自动数据整理：一种基于聚类的方法》。arXiv 预印本 arXiv:2405.15613，2024年。2，7[38] Zhaowei Wu、Binyi Su、Qichuan Geng、Hua Zhang 和 Zhong Zhou。《通过提示学习和鲁棒决策边界提升小样本开放集目标检测》。载于《第三十四届国际人工智能联合会议论文集》，IJCAI-25，第6606–6614页。国际人工智能联合会议组织，2025年。主会场。2[39] Jingkang Yang、Litong Feng、Weirong Chen、Xiaopeng Yan、Huabin Zheng、Ping Luo 和 Wayne Zhang。《利用自包含置信度进行网络监督图像分类》。载于《欧洲计算机视觉会议论文集》，第779–795页。施普林格，2020年。2[40] Yangyang Yu、Zhiyuan Yao、Haohang Li、Zhiyang Deng、Yupeng Cao、Zhi Chen、Jordan W. Suchow、Rong Liu、Zhenyu Cui、Zhaozhuo Xu、Denghui Zhang、Koduvayur Subbalak- shmi、Guojun Xiong、Yueru He、Jimin Huang、Dong Li 和 Qianqian Xie。《Fincon：一种通过概念性语言强化提升金融决策能力的合成大语言模型多智能体系统》。载于《神经信息处理系统进展》（NeurIPS 2024），2024年。3[41] Hai Zhang、Junzhe Xu、Shanlin Jiang 和 Zhenan He。《简单的语义辅助小样本学习》。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第28588–28597页，2024年。2[42] Shaokun Zhang、Ming Yin、Jieyu Zhang、Jiale Liu、Zhiguang Han、Jingyang Zhang、Beibin Li、Chi Wang、Huazheng Wang、Yiran Chen 和 Qingyun Wu。《哪个智能体导致了任务失败？何时导致？关于大语言模型多智能体系统的自动化失败归因》。载于《第42届国际机器学习会议论文集》，第76583–76599页。PMLR，2025年。3[43] Yusen Zhang、Ruoxi Sun、Yanfei Chen、Tomas Pfister、Rui Zhang 和 Sercan Arik。《智能体链：大语言模型协作处理长上下文任务》。载于《神经信息处理系统进展》，37：132208–132237，2024年。3

[44] Yilun Zhou、Austin Xu、PeiFeng Wang、Caiming Xiong 和 Shafiq Joty。《评估作为评价者的裁判：作为测试时扩展评价者的 LLM-as-judges 基准测试 jetts bench- mark》。载于《第*42*届国际机器学习会议（*ICML*）论文集》，2025年。海报。3[45] Xingyu Zhu、Shuo Wang、Jinda Lu、Yanbin Hao、Haifeng Liu 和 Xiangnan He。《通过注意力特征正则化提升小样本学习》。载于《*AAAI*人工智能会议论文集》，第7793–7801页，2024年。2