计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(11): 2986−3001, 2026

粒度引导多模型协作的生成式图像分类方法

营 鹏[1,2] 李仲年[1,2] 袁宏广[1] 魏 孟[1] 许新征[1,2]

[1](中国矿业大学计算机科学与技术学院/人工智能学院 江苏徐州 221116)

[2](矿山数字化教育部工程研究中心(中国矿业大学) 江苏徐州 221116)

(pengying@cumt.edu.cn)

Generative Image Classification Method via Granularity-Guided Multi-Model Collaboration

Ying Peng[1,2], Li Zhongnian[1,2], Yuan Hongguang[1], Wei Meng[1], and Xu Xinzheng[1,2]

[1](*School of Computer Science and Technology / School of Artificial Intelligence*, *China University of Mining and Technology*, *Xuzhou*, *Jiangsu* 221116)

[2](*Mine Digitization Engineering Research Center of the Ministry of Education* (*China University of Mining and Technology*), *Xuzhou*, *Jiangsu* 221116)

Abstract In recent years, large vision-language models (VLM) have demonstrated remarkable zero-shot prediction capabilities in image classification tasks. Existing VLM-based classification approaches typically rely on a set of predefined category labels or external textual knowledge bases as semantic prior to align and match visual content with language. However, in real-world applications, such predefined labels often fail to cover the vast and dynamic concept space of the open world, thereby limiting the adaptability and generalization ability of these models. To address this challenge, a more open-ended task, generative image classification (GIC), is introduced, in which no prior semantic labels are provided, and the target categories reside in an open-ended, potentially unbounded semantic space, substantially increasing the classification difficulty. To tackle the complexity of GIC, a granularity-guided multi-model collaboration framework is proposed, designed to fully leverage the strengths of VLM in both image understanding and text generation. In this framework, prompt templates are carefully crafted to guide VLM in generating candidate labels at varying levels of semantic granularity. Subsequently, the candidate label space is expanded by a language model that associates semantically related categories based on the generated image descriptions at each granularity level. Finally, the most relevant label is identified through a semantic similarity retrieval process in the embedding space. The effectiveness and superiority of the proposed method for the GIC task are validated through extensive experiments on multiple fine-grained image classification benchmarks.

Key words generative image classification; vision-language models; multi-model collaboration; open vocabulary; multi-class

摘 要 近年来,大型视觉-语言模型(vision-language model, VLM)在图像分类任务中展现出卓越的零样

收稿日期: 2025-09-30; **修回日期**: 2026-09-07

基金项目: 国家自然科学基金项目(61976217, 62306320); 江苏省自然科学基金项目(BK20231063); 中国矿业大学研究生创新计划项目(2026WLKXJ116); 江苏省研究生科研实践创新计划项目(26CXJH4835)

This work was supported by the National Natural Science Foundation of China (61976217, 62306320), the Natural Science Foundation of Jiangsu Province (BK20231063), the Graduate Innovation Program of China University of Mining and Technology (2026WLKXJ116), and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (26CXJH4835).

通信作者: 许新征(xxzheng@cumt.edu.cn)

本预测能力，推动了图像识别技术的发展。现有 VLM 普遍依赖于一组预定义的类别标签或外部文本知识库作为语义先验，以实现图像与语义之间的对齐与匹配。然而，在实际应用中，预定义标签往往难以覆盖开放世界中的所有概念，导致模型的适应性和泛化能力受到限制。为应对这一挑战，提出了一种更具开放性的任务——生成式图像分类（generative image classification, GIC），该任务不提供任何先验的语义标签，而目标类别潜藏于一个开放的、可能无限扩展的语义概念空间中，极大地增加了分类难度。针对该任务的复杂性，提出了一种基于粒度引导的多模型协作框架，充分发挥 VLM 在图像理解与文本生成方面的优势。具体而言，该框架通过设计高效的提示模板，引导 VLM 在不同语义粒度下生成潜在标签；随后，语言模型基于图像描述自动关联同一粒度内的相关类别，扩展候选标签空间；最后，通过在嵌入空间中进行语义相似度检索，筛选出最相关的最终标签。在多个细粒度图像分类基准数据集上进行的实验验证了所提方法在 GIC 任务中的有效性与先进性。

关键词 生成式图像分类；视觉-语言模型；多模型协作；开放词汇；多类别

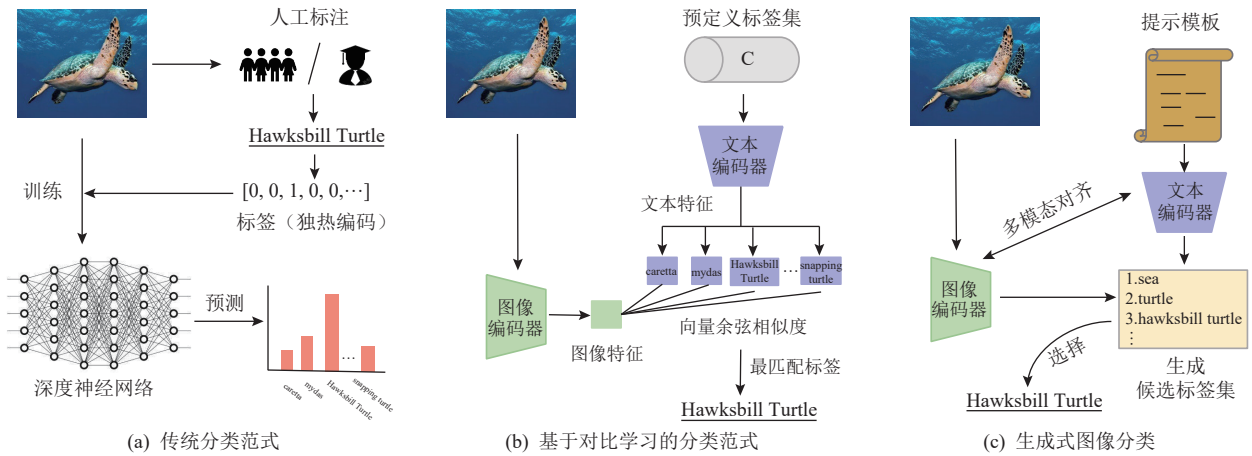
中图法分类号 TP18

DOI: 10.7544/issn1000-1239.202550731 **CSTR:** 32373.14.issn1000-1239.202550731

近年来，图像分类技术在计算机视觉领域得到了广泛关注，尤其是在开放世界图像识别、弱监督学习和多模态融合等方向取得了显著进展。其中，多类别图像分类（multi-class image classification, MIC）任务旨在为输入图像分配一个唯一且语义准确的类别标签，是众多视觉任务的基础环节之一[1-3]。如图 1（a）所示，传统 MIC 方法依赖于大量人工标注的训练样本，通过训练判别式分类模型学习图像特征与预定义类别标签之间的映射关系[4]。这一范式在封闭标签空间及类别数量有限的任务中表现稳定，但其高度依赖人工标注，导致构建训练数据集成本昂贵，且难以扩展至开放标签空间下的大规模图像识别任务[5]。

为了解决上述问题，研究者开始关注如何借助预训练模型进行零样本（zero-shot）或少样本（few-shot）的图像分类，从而减少对高质量人工标签的依赖。受大语言模型（large language model, LLM）成功应用的启发，近年来兴起了一类融合图像与文本模态的大型视觉-语言模型（vision-language model, VLM），代表模型包括 CLIP[6] 和 RAM[7] 等。这类模型通常基于图文对比学习机制，在海量图像-文本对上进行联合预训练，从而在视觉与语言模态之间学习统一的嵌入空间，并具备良好的迁移能力与通用性[8-9]。如图 1（b）所示，VLM 通常接受 2 种模态的输入：图像模态与文本模态，其中文本模态以“提示词（prompt）”的形式提供语义先验。具体而言，用户需为候选类别构造一组文本标签作为输入提示，模型通过计算图像与这些提示语之间的语义相似度，选出匹配度最高的类别标签作为最终预测结果。由于 VLM 在训练阶段已见过大量自然图像与语言描述，其在零样本分类任务中表现出强大的泛化能力，显著减轻了传统方法对人工标注的依赖。

然而，构建这样一组预定义的类别标签集合往



(a) 传统分类范式 (b) 基于对比学习的分类范式 (c) 生成式图像分类

Fig. 1 Illustration of generative image classification

图 1 生成式图像分类示意图

往需要大量的时间、人力与领域知识投入,在实际应用中面临显著的成本与可扩展性挑战。尤其是在开放世界图像识别场景中,类别集合可能呈现动态变化、分布稀疏甚至语义未知的特征,使得固定标签设计难以满足多样化任务需求。此外,尽管已有部分研究工作尝试在更为开放的语义设定下进行图像分类探索[10-11],但这些方法仍然依赖于外部数据库或知识库,用以限定候选标签的语义范围。这种依赖使得模型性能在很大程度上受限于辅助数据库的完备性与语义质量,难以实现真正意义上的开放标签识别。

为进一步突破当前方法在标签可拓展性与语义开放性方面的限制,本文提出一种更具挑战性且更贴近开放世界需求的图像识别新任务:生成式图像分类(generative image classification, GIC)。与以往方法不同,GIC 任务设定下不提供任何预定义标签集合或外部知识库,模型需完全依靠 VLM 自身对图像的理解能力与语义生成能力,自动生成候选标签并完成分类预测,如图 1(c)所示。值得注意的是,当前主流的 VLM 普遍在包含数百万图文对的大规模语料上进行预训练,已内嵌丰富的通用知识表示。因此,在 GIC 任务中,VLM 不仅作为图像特征提取器,更作为一个语义知识库,通过生成式机制为图像自动赋予开放语义标签,提升在未知类别下的识别能力与适应性。

GIC 任务面临的最大挑战在于:目标类别嵌套于一个开放且无限扩展的语义概念空间中,缺乏可供检索的先验标签集合。在传统图像分类任务中,语义先验(如类别标签集合或外部知识库)通常为模型提供明确的对齐方向和约束边界,从而有效缩小检索空间。然而,在 GIC 的任务设定下,语义先验的缺失导致模型在生成候选标签时缺乏引导信息,难以精准定位目标语义,极大地增加了分类的不确定性和任务复杂度。在实验中观察到,直接向 VLM 输入简单提示词(例如:图中的物体是什么?)时,模型往往更倾向于生成语义粒度较粗的标签。虽然这种泛化能力在某些情境下具有一定的容错性,但在实际应用中可能引发严重后果。例如,给定一张图像,其真实类别为“玳瑁海龟(hawksbill turtle)”——一种濒危物种,而 VLM 可能仅返回“海龟(sea turtle)”这一粗粒度类别,忽略了其物种特异性,从而在生态保护等高敏感任务中引发误导甚至决策风险。这一现象表明,如何有效引入语义粒度控制机制,引导 VLM 生成更精准、细粒度的标签,是提升 GIC 性能的关键所在。

近期在多智能体系统的研究成果表明,在复杂任务中引入多个 LLM 之间的协作机制,能够显著提升整体系统的智能表现与任务完成效率[12-13]。受到该思想的启发,本文提出一种面向 GIC 任务的粒度引导多模型协作框架(granularity-guided multi-model collaboration framework, GMCF),该方法无需训练,能够在不依赖外部标签或知识库的前提下完成开放语义空间中的图像分类任务。具体而言,GMCF 首先基于语义粒度建模思想,将潜在标签划分为 3 个粒度层级,分别为:上位层级、粗粒度层级与细粒度层级。随后,引入 VLM(如 LLaVA[14])作为粒度生成器,根据输入图像分别生成 3 个层级下的候选标签。同时,VLM 还负责生成一段详细的图像描述文本,为后续标签扩展提供语义引导信息。接着,采用一个对比学习模型作为标签过滤器,通过计算图像嵌入与文本嵌入之间的语义相似度,初步筛选出最具代表性的粒度标签。为了进一步丰富候选标签空间,GMCF 还包含一个 LLM(如 LLaMA[15])作为标签扩展器,该模块基于前述图像描述与初始粒度标签,生成若干具有语义相关性但不完全重复的新标签,从而构建一个扩展后的标签池。最后,标签筛选器再次基于图文嵌入相似度,从扩展标签池中选择与图像语义最为匹配的标签作为最终分类结果。

综上所述,本文的主要工作和贡献包括:

1)提出并形式化定义了 GIC 任务,该任务不依赖预定义的类别标签或外部知识库,候选标签完全由视觉-语言模型生成,更贴近开放世界场景下的实际应用需求。

2)提出了一种基于语义粒度引导的多模型协作框架,在不进行额外训练的前提下,结合视觉-语言模型、对比学习模型和语言模型之间的功能协作,用以提升 GIC 任务中的标签生成质量与分类性能。

3)在 5 个细粒度图像分类基准数据集上对所提方法进行了评估,实验结果表明,GMCF 在多个数据集上取得了稳定且具有竞争力的性能,验证了其在无标签设定下的有效性。

1 相关工作

1.1 视觉语言模型

大型视觉-语言模型旨在建模图像与文本之间复杂的跨模态语义对齐关系,是多模态人工智能领域的核心研究方向之一。近年来,随着大规模图文数据的获取与多模态模型架构的发展,VLM 在多个典

型任务中取得了突破性成果，包括图像描述[16]、视觉问答、多模态检索等，其泛化能力与零样本学习能力受到广泛关注。其中，CLIP 作为开创性工作，采用跨模态对比学习框架，在包含 4 亿对图文配对数据的大规模数据集上进行联合预训练，实现了图像与文本在统一嵌入空间中的语义对齐，为下游任务提供了无需微调的零样本分类能力。InstructBLIP[17] 则在 BLIP 模型的基础上，引入指令调优机制，通过人类指令驱动模型生成过程，增强了对复杂自然语言指令的理解与执行能力，从而显著提升了模型在开放域问答、图像理解与人机交互等多任务场景下的性能。LLaVA 通过将视觉编码器与 LLM 进行桥接，并引入投影层将视觉特征映射至语言空间，使得模型能够以语言交互的方式理解视觉内容，具备较强的图文融合与多轮对话能力。此外，Florence-VL[18] 提出多层次多任务视觉建模方法，强化模型对空间结构与语义实体的感知能力；Qwen2-VL[19] 系列在多语言和多图像对话方面展现出卓越性能，具备跨文化与多模态交互能力；mPLUG-Owl2[20] 引入模态自适应机制以提升视觉和语言之间的协同建模能力；Alpha-CLIP[21] 则通过引入 Alpha 通道实现显著区域关注，从而提升模型在细粒度识别中的表现。

除了模型结构的演进，提示调优[22-23] 作为一种参数高效的迁移学习范式，也被广泛应用于 VLM 中。该方法通过优化可学习的文本提示，使预训练模型能够更好适应下游任务，无需更新大规模模型参数。尽管提示调优具备良好的可扩展性与效率优势，但其在泛化性、迁移性及跨任务一致性方面仍存在一定局限，且在开放世界场景中依赖于先验语义信息，难以满足无标签分类等更加灵活的需求。

1.2 零样本图像分类

零样本图像分类（zero-shot image classification, ZIC）任务旨在在训练阶段不涉及目标类别的前提下，实现对未见类别图像的自动识别，因而在开放世界场景中具有重要的研究价值和广泛的应用前景[24]。为弥合训练类别与测试类别之间的语义鸿沟，早期研究多依赖于类属性向量或基于文本描述的语义嵌入，将图像特征投影到一个共享的语义空间中，从而实现跨类别识别。然而，这类方法在高层语义表示能力、语义泛化性以及属性标注成本方面均存在显著局限。

为进一步提升 ZIC 模型对类别语义结构的建模能力，CHiLS[25] 提出在标签空间中引入层次结构，通过构建上位概念与下位概念的组织关系，增强类别标签的语义丰富性与可区分性，显著提升了模型在细粒度分类任务中的性能。另一方面，ZPE[26] 提出了一种提示语集合生成与加权评估机制，利用多个语义变体引导预训练视觉-语言模型进行图像识别，在不引入额外知识库的前提下提升了预测准确率。此外，为适应更复杂的开放世界环境，CaSED[10] 打破了传统 ZIC 对固定词表的依赖，引入外部语义数据库进行候选类别的扩展检索，从而实现词表无约束的图像分类。然而，此类方法的有效性高度依赖于外部数据库的覆盖范围与检索质量，且仍需借助外部资源构建显式的分类先验，难以彻底解决在“无标签、无数据库”设定下的图像识别问题。

与之相比，本文探索了更具挑战性的 GIC 任务。通过引导视觉-语言模型生成多尺度粒度标签，并构建了一种多模型协作机制，实现了在不依赖任何额外语义信息和训练过程的情况下完成图像分类任务。这一方法突破了传统方法对于预定义标签或外部知识库的依赖，为图像分类任务提供了更加灵活和开放的解决方案。通过利用 VLM 的多模态理解能力，不仅提升了分类的准确性，还增强了模型的泛化能力，使其能够适应更复杂和动态的应用场景。

2 方法论

在本节中，本文首先给出 GIC 任务的符号化定义。然后，介绍所提出的粒度引导多模型协作框架中的 3 个核心模块：粒度生成器、标签过滤器和标签扩展器。它们共同作用，通过综合图像描述与多个粒度标签调整标签细节，在嵌入空间中过滤无关标签，以及扩展多个潜在标签来优化分类过程。GMCF 的整体工作流程如图 2 所示。

2.1 问题定义

给定输入图像 $\boldsymbol{x}$，文本提示模板 t 以及 1 组预定义的标签集合 C，基于对比学习的 VLM，如 CLIP，通过计算图像与各候选标签的文本描述之间的语义相似度来实现分类。这一过程通常被建模为嵌入空间中的检索问题。VLM 可表示为一个从图像空间与文本空间到标签集合的映射 f_{VLM}，目标是选择出与图像最相似的标签 $c\in C$，其计算公式为：

$$f_{\mathrm{VLM}}(\boldsymbol{x},t,c)=\arg\max_{c\in C}\left\langle f^{\mathrm{v}}_{\mathrm{VLM}}(\boldsymbol{x}),f^{\mathrm{t}}_{\mathrm{VLM}}(\mathcal{I}(t,c))\right\rangle,\tag{1}$$

其中 $f^{\mathrm{v}}_{\mathrm{VLM}}(\cdot)$ 和 $f^{\mathrm{t}}_{\mathrm{VLM}}(\cdot)$ 分别表示 VLM 的视觉编码器和文本编码器，$\mathcal{I}(t,c)$ 表示将标签 c 插入提示模板 t 的字符串拼接操作，$\langle\cdot,\cdot\rangle$ 表示 2 个编码器输出间的余弦相

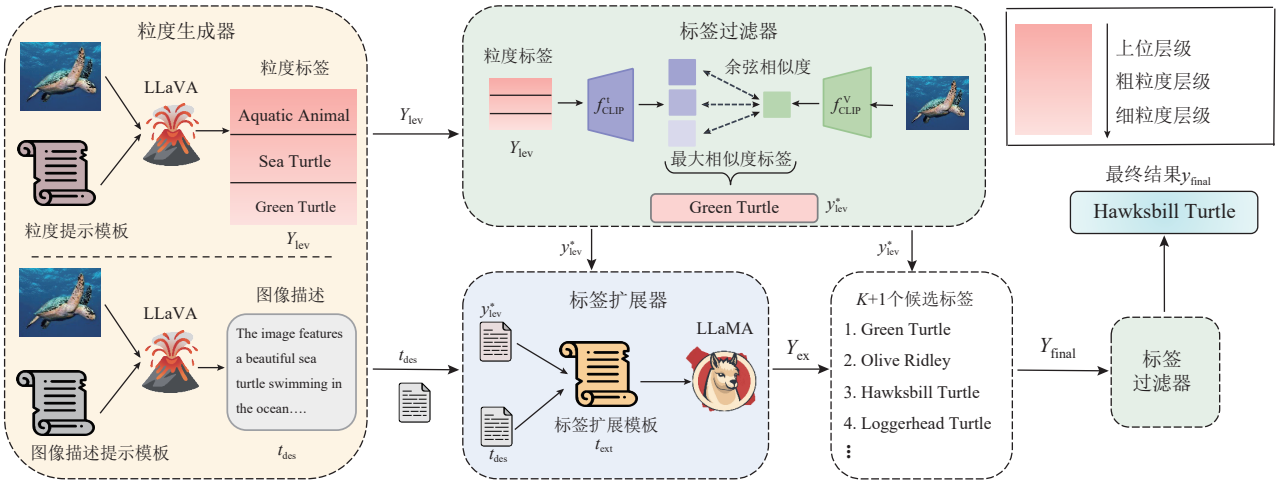


Fig. 2 Framework of granularity-guided multi-model collaboration

图 2 粒度引导的多模型协作框架

似度。

式(1)度量了图像特征向量与文本特征向量在共享多维空间中的相似度。余弦值越大(即夹角越小),表明图像包含的视觉内容与文本描述的语义概念的几何距离越近。因此,该分类范式的本质是在高维跨模态语义空间中,寻找一个与目标图像向量方向最一致、语义对齐度最高的文本特征。基于此,模型能够在共享的嵌入空间中,根据图文匹配度对候选标签进行排序和选择。

在 GIC 任务中,由于不存在预定义的标签集合C,目标标签必须从更广泛的概念空间S中推断,其中$C \subset S$。然而,概念空间S本质上是开放且潜在无限的,涵盖了大量语义概念。由于缺乏明确的预定义类别作为对齐锚点,模型在概念空间中缺乏检索边界的约束,这使得式(1)中的分类范式失效。因此从图像与文本提示构建映射成为一个具有挑战性的问题。

为解决该问题,本文引入了生成式 VLM,如 LLaVA 等,它们能够生成自由格式的文本标签。具体地,给定图像$\boldsymbol{x}$和文本提示模板t,模型生成标签$y = f_{\mathrm{VLM}}(\boldsymbol{x}, t)$,其中响应$y$是通过对条件概率分布$p(y|\boldsymbol{x}, t)$进行预测得到的。

2.2 粒度生成器

粒度生成器模块旨在针对输入图像生成具有不同语义粒度的标签。为实现该目标,本文利用生成式 VLM 的语义建模与生成能力,从图像中提取多层次的概念信息。具体而言,本文设计了粒度提示模板引导模型在输出中体现从粗粒度类别到细粒度实例的语义区分。该提示模板通过控制语义抽象程度,引导模型生成具有不同细化层级的标签序列。

具体地,遵循 Kim 等人[27]的设定,本文定义 3 种粒度等级如下:

上位层级表示语义层次中最抽象、最广泛的分类级别。该粒度下的标签用于描述具有高度概括性的广义类别,涵盖视觉或功能上差异显著的多类对象(如“物体”“动物”“交通工具”)。

粗粒度层级位于上位层级与细粒度层级之间,兼顾语义的泛化能力与区分能力。该层级的标签能够对超类中的显著子类进行基本划分(如“狗”“汽车”)。

细粒度层级表示分类层级中最具体的一层。此层级下的标签强调语义上的微小差异与视觉细节,对应于具体实例或专业化的子类(如“金毛猎犬”“轿车”)。

此外,为提升粒度生成器在指令遵循能力与生成语义质量方面的表现,本文在每一粒度级别定义之后,补充了 1 组示例内容。这些示例在提示模板中起到了双重作用:一方面有助于模型理解并强化各粒度级别的语义边界,另一方面提供了输出的结构性范式,从而在推理过程中有效降低生成模糊性。值得注意的是,这些示例不依赖于任何预定义标签先验,而是从常用词汇中随机抽样,以增强提示模板的通用性和跨数据集适应性。粒度提示模板的关键提示词如图 3 所示。

将图像与粒度提示一同输入生成式 VLM 后,模型输出包含 3 个语义粒度标签的元组,记作Y_{lev}。该结构化输出反映了图像中所蕴含的层次化语义信息,构成对图像的多层概念表示。除标签生成外,本文还引导 VLM 输出图像的详细自然语言描述t_{des},以此作为后续候选标签扩展阶段的语义参考。该描述不仅增强了上下文表达能力,也有助于在初始粒度标

Input:
Carefully observe the given image. Classify it according to the provided guideline
Guideline: This image will be used in classification task. Your task is to provide the following three levels of labels for this image. Their definitions are as follows:
Superordinate-level: …, such as animals, plants, furniture, tools.
Coarse-level: …, such as mammals, trees, chairs, kitchen tools.
Fine-level: …, such as Golden Retriever, oak tree, rocking chair, paring knife.
Provide your answer after corresponding level.
superordinate-level:
coarse-level:
fine-level:

请仔细观察给定图像。根据所提供的指南对其进行分类。
指南： 该图像将用于分类任务。你的任务是为该图像提供以下三个层级的标签。它们的定义如下：
上位层级：…，例如动物、植物、家具、工具。
粗粒度层级：…，例如哺乳动物、树木、椅子、厨房工具。
细粒度层级：…，例如金毛猎犬、橡树、摇椅、削皮刀。
请按照对应层级填写你的答案。
上位层级：
粗粒度层级：
细粒度层级：

Fig. 3 The key prompts of the granularity prompt template

图 3 粒度提示模板的关键提示词

签基础上挖掘更丰富、更相关的语义标签。

2.3 标签过滤器

根据粒度标签元组Y_{lev}，本文使用 CLIP 作为标签过滤器，用于根据标签与图像之间的语义相关性对各粒度标签进行评分与排序。具体而言，本文利用 CLIP 的视觉编码器$f_{\text{CLIP}}^{\text{v}}$和文本编码器$f_{\text{CLIP}}^{\text{t}}$分别对输入图像$\boldsymbol{x}$与粒度标签$y_{\text{lev}} \in Y_{\text{lev}}$进行特征提取，将它们映射至共享的多模态表示空间。随后通过计算图像与各粒度标签之间的余弦相似度来度量其语义匹配程度。得分最高的标签被选为与图像语义最一致的粒度标签，记作y_{lev}^{*}，其计算过程如式(2)所示。

$$y_{\text{lev}}^{*} = \arg\max_{y_{\text{lev}} \in Y_{\text{lev}}} \left\langle f_{\text{CLIP}}^{\text{v}}(\boldsymbol{x}), f_{\text{CLIP}}^{\text{t}}(y_{\text{lev}}) \right\rangle 。 \tag{2}$$

由于 CLIP 在大规模视觉-语言对齐数据上进行了充分训练，其多模态嵌入具有较强的语义表达能力，因此该筛选机制不仅确保了所选标签在嵌入空间中接近，也保证了其与图像内容之间的语义一致性。

这一双重约束使得标签过滤器能够有效剔除在嵌入空间中虽接近但语义上无关的噪声标签，从而提高粒度标签选择的准确性与上下文相关性，为后续的标签扩展提供了可靠基础。

2.4 标签扩展器

尽管粒度生成器与标签过滤器共同确保了标签的语义契合与粒度合理性，但仅依赖单一路径生成标签难以覆盖更丰富的语义空间，限制了候选标签的多样性与表达能力。为进一步提升候选标签集的语义广度与适应性，本文引入了第 3 个核心模块标签扩展器，该模块利用 LLM 强大的上下文理解与语义关联能力，扩展生成与图像内容相关的补充标签，增强整体分类系统的鲁棒性与泛化能力。

具体而言，标签扩展器以所选粒度标签y_{lev}^{*}与图像详细描述t_{des}为输入，结合扩展提示模板t_{ext}，通过字符串连接操作$\mathcal{I}(t_{\text{ext}}, t_{\text{des}}, y_{\text{lev}}^{*})$构造输入提示，从而引导 LLM 生成$K$个语义相似的扩展标签，记为扩展标签集合$Y_{\text{ex}}$，其过程表达式为

$$Y_{\text{ex}} = f_{\text{LLM}}(\mathcal{I}(t_{\text{ext}}, t_{\text{des}}, y_{\text{lev}}^{*}))。 \tag{3}$$

值得注意的是，标签扩展器被明确指令生成与所选粒度标签y_{lev}^{*}相同粒度层级的标签，从而确保输出范围的可控性与语义一致性。扩展提示模板t_{ext}中的关键提示词如图 4 所示。

Input:
Carefully analyze the given image caption **Caption**: '\{\}' and corresponding label **Label**: '\{\}' Your task is using your imagination to associate other categories at the same granularity level as **Label** according to the **Caption**, and the provided guideline.
Guideline: Firstly, understand the definition of granularity level. There are three kinds of granularity levels, which are defined as follows:
Superordinate-level: …, such as animals, plants, furniture, tools.
Coarse-level: …, such as mammals, trees, chairs, kitchen tools.
Fine-level: …, such as Golden Retriever, oak tree, rocking chair, paring knife.
Secondly, associate \{\} different categories at the same granularity level as **Label**.
Provide your answer after "Answer:".

请仔细分析给定图像的描述（Caption）：\{\} 和对应标签（Label）：\{\}。
你的任务是根据Caption 和所提供的指南，发挥想象力，联想到与 Label 相同粒度层级的其他类别。
指南：首先，理解粒度层级的定义。分类粒度层级分为3种，定义如下：
上位层级（Superordinate-level）：…，例如动物、植物、家具、工具。
粗粒度层级（Coarse-level）：…，例如哺乳动物、树木、椅子、厨房工具。
细粒度层级（Fine-level）：…，例如金毛猎犬、橡树、摇椅、削皮刀。
其次，根据 Caption 的内容，联想出与 Label 属于相同粒度层级的其他 {K} 个类别。
请在 Answer: 后填写你的答案。

Fig. 4 The key prompts of the label expander prompt template

图 4 标签扩展提示模板的关键提示词

此外，考虑到真实场景中图像语义复杂度与数据集的差异性，预定义生成扩展标签的数量K无法满足动态变化的实际部署需求。本文提出了一种基于计算预算与置信度引导的自适应机制，旨在动态调整候选标签数量以平衡系统推理效率与准确率。该机制的核心思想是在考虑推理效率的条件下利用筛选出的粒度标签的置信度来计算后续扩展标签的数量。

具体地，标签过滤器初步筛选出最优粒度标签y_{lev}^{*}后，GMCF 通过 Softmax 函数将式(2)中y_{lev}^{*}的余弦相似度得分转化为模型对初步分类结果的置信度$Conf$。

$$Conf = \frac{\exp(\left\langle f_{\text{CLIP}}^{\text{v}}(\boldsymbol{x}), f_{\text{CLIP}}^{\text{t}}(y_{\text{lev}}^{*}) \right\rangle)}{\sum_{y_{\text{lev}} \in Y_{\text{lev}}} \exp(\left\langle f_{\text{CLIP}}^{\text{v}}(\boldsymbol{x}), f_{\text{CLIP}}^{\text{t}}(y_{\text{lev}}) \right\rangle)} 。 \tag{4}$$

置信度越高，表明该结果更可靠，因此可以适当减少扩展标签的数量，以提高计算效率。相反，置信度越低，模型对初步分类结果的不确定性越高，需要更多的扩展标签来丰富候选标签池。基于此，本文将扩展标签的数量K定义为关于模型对y_{lev}^{*}预测置信度的线性负相关函数：

$$K = \max(K_{\min}, \lceil K_{\max} \cdot (1 - Conf) \rceil), \tag{5}$$

其中$K_{\min}$是保证多样性条件下的扩展标签生成的最小数量，$K_{\max}$是考虑计算效率的扩展标签生成的最大数量，$\lceil\cdot\rceil$表示向上取整。通过这一机制，当初始置信度较高时，GMCF会自动收缩K，避免无效的语义发散并节约计算资源；反之，当置信度较低时，GMCF将释放更多预算以促使大语言模型进行深层语义挖掘。这种自适应策略使得GMCF框架能够在不增加额外调参负担的前提下，实现资源在不同识别难度的样本间的优化分配，提高其在开放场景下的适应性。

对候选标签池的扩展完成后，记最终的候选标签池为$Y_{\text{final}}=Y_{\text{ex}}\cup\{y^{*}_{\text{lev}}\}$。为从中选择最终标签$y_{\text{final}}$，本文复用标签过滤器模块，对其$K+1$个候选标签在共享嵌入空间中进行相似度评分，选取与图像语义最契合的标签作为最终输出：

$$y_{\text{final}}=\arg\max_{y'\in Y_{\text{final}}}\left\langle f^{\text{v}}_{\text{CLIP}}(\boldsymbol{x}),f^{\text{t}}_{\text{CLIP}}(y')\right\rangle。\tag{6}$$

GMCF框架的具体流程如算法1所示。

算法 1. 粒度引导的多模型协作GIC。

输入：输入图像$\boldsymbol{x}$，粒度提示模板t_{gra}，扩展提示模板t_{ext}；

输出：最终分类标签y_{final}。

① $Y_{\text{lev}},t_{\text{des}}\leftarrow VLM(\boldsymbol{x},t_{\text{gra}})$; /*粒度生成器输出候选粒度标签及图像描述*/

② 初始化最高匹配度$d_{\max}=0$;

③ for y in Y_{lev} do /*遍历候选粒度标签集*/

④ if $\left\langle f^{\text{v}}_{\text{CLIP}}(\boldsymbol{x}),f^{\text{t}}_{\text{CLIP}}(y_{\text{lev}})\right\rangle\geqslant d_{\max}$ then

⑤ $d_{\max}\leftarrow\left\langle f^{\text{v}}_{\text{CLIP}}(\boldsymbol{x}),f^{\text{t}}_{\text{CLIP}}(y_{\text{lev}})\right\rangle$;

⑥ $y^{*}_{\text{lev}}\leftarrow y_{\text{lev}}$; /*筛选最匹配的粒度标签*/

⑦ end if

⑧ end for

⑨ 根据式(4)与式(5)计算扩展标签数量K;

⑩ $Y_{\text{ex}}\leftarrow LLM(t_{\text{ext}},t_{\text{des}},y^{*}_{\text{lev}},K)$; /*标签扩展器生成$K$个扩展标签*/

⑪ $Y_{\text{final}}\leftarrow Y_{\text{ex}}\cup\{y^{*}_{\text{lev}}\}$; /*构建最终候选标签池*/

⑫ $d_{\max}\leftarrow 0$; /*重置最高匹配度*/

⑬ for y' in Y_{final} do /*2次筛选*/

⑭ if $\left\langle f^{\text{v}}_{\text{CLIP}}(\boldsymbol{x}),f^{\text{t}}_{\text{CLIP}}(y')\right\rangle\geqslant d_{\max}$ then

⑮ $d_{\max}\leftarrow\left\langle f^{\text{v}}_{\text{CLIP}}(\boldsymbol{x}),f^{\text{t}}_{\text{CLIP}}(y')\right\rangle$;

⑯ $y_{\text{final}}\leftarrow y'$;

⑰ end if

⑱ end for

⑲ return y_{final}。

在算法1中，GMCF框架严格遵循GIC任务的零样本设定，全流程未引入任何人工先验或特定数据集知识，且所有模型参数保持冻结状态。这种设计不仅提升了系统的通用性，也保证了其在跨数据集、跨领域任务中的迁移能力与实用性。

2.5 多模型协作机制分析

针对GMCF框架中异构模型(VLM、LLM与对比式VLM)的协作逻辑，本节对其多模态耦合机制进行分析。

在传统的多模型级联中，特征对齐通常发生在底层的隐式高维向量空间内，不仅计算复杂且极易引发维度与分布的不匹配。为了解决这一特征对齐难题，GMCF采用自然语言作为跨模型的通用符号化表示接口。自然语言将连续且包含冗余的视觉特征投影为离散、高层级的语义符号。具体而言，粒度生成器调用视觉语言模型将视觉感知映射为文本概念；而标签扩展器则调用LLM，在文本模态中进行逻辑发散，生成候选标签。这种“符号耦合”确保了不同异构模型能够在统一的语义基准下进行协作交互，避免了底层特征空间的对齐难题。

然而，在多模型协作过程中，由于LLM在进行标签扩展时缺乏视觉信号的直接约束，极易产生偏离图像真实内容的“语义漂移”，从而导致最后生成的标签与图像不匹配。为了弥合这一鸿沟，GMCF在生成粒度标签的同时使用视觉语言模型生成了图像的细致描述，为标签扩展器提供了重要参考。此外，GMCF采用了2次筛选策略，使用预训练的CLIP作为标签过滤器，利用其在预训练阶段已建立的稳定的图文联合嵌入空间，强制要求最终筛选出的标签必须在多模态共享空间中与图像特征向量的夹角最小。这一机制有效剔除了语言逻辑上合理但视觉特征上不匹配的噪声标签，建立了视觉与文本匹配的约束关系。

总体而言，GMCF通过自然语言符号化耦合与视觉文本对齐约束校正的双重机制，实现了异构模型间高效协同与语义一致性的统一。

3 实验分析

3.1 实验数据

本文在5个多类别图像分类公开数据集上进行实验，包括Caltech-101[28]，Food-101[29]，Describable Textures Dataset[30]，Oxford Pets，Flowers-102[31]。由于所提出的方法不需要进行额外的训练，本文只使用各个数据集的测试集样本对GMCF的效果进行测试。表1给出了数据集所对应的详细信息。

Table 1 Experimental Datasets

表 1 实验数据集

数据集	缩写	样本数	类别数
Caltech-101	C101	6 085	101
Food-101	F101	25 250	101
Oxford Pets	Pets	3 669	37
Flowers-102	Flwr	6 149	102
Describable Textures Dataset	DTD	1880	47

Caltech-101[28] 数据集由加州理工学院发布，包含 102 个类别（101 个对象类别与 1 个背景类别），总计约 9 146 张图像，其中测试集图像 6 085 张。每个类别的样本数量不一，通常在 40~800 张之间。图像主要以中央对齐的单个物体为主，适用于评估模型在中等粒度、多样性较高类别条件下的分类能力。

Food-101[29] 是一个专注于菜品识别的大规模数据集，共包含 101 个食品类别，每类约 1 000 张图像，总计 101 000 张图像，其中测试集图像 25 250 张。该数据集图像源自真实世界网络图像，包含大量背景噪声和拍摄风格的变化。菜品之间的视觉相似度较高，如“cheese cake”和“carrot cake”常在颜色与形状上相近，因此对细粒度识别能力提出较高要求。该数据集广泛用于研究细粒度分类以及多模态学习等问题。

Describable Textures Dataset[30] 是一组专门用于纹理分析的数据集，由伦敦大学学院提出，包含 47 个基于人类可描述语义的纹理类别，每类包含约 120 张图像，总规模为 5 640 张，其中测试集数据 1880 张。该数据集的标签来自自然语言描述，具有人类认知语义特征，适合用于研究视觉-语言对齐、视觉属性建模和主观概念识别等任务。

Oxford Pets 数据集由牛津大学发布，涵盖 37 种宠物品种（包括 25 种狗和 12 种猫），总计 7 349 张图像，其中测试集图像 3 669 张。每张图像配有类别标签和像素级分割掩码，适用于图像分类与目标分割等多任务研究。由于许多宠物品种在颜色、体型、姿态等方面存在细微差异，Oxford Pets 是一个典型的细粒度分类数据集，也常用于多模态语义推理任务。

Flowers-102[31] 由牛津大学发布，包含 102 个花卉品种，图像总数为 8 189 张，其中测试集图像 6 149 张。每个类别的图像数量介于 40~258 之间，图像采集自野外或网络，具有光照、姿态、焦距等自然变化。该数据集挑战性在于大量花卉类别之间外观相似，需要模型具备细节识别与局部特征建模能力。Flowers-102 是评估细粒度分类、生成模型以及视觉语言联动机制性能的标准数据集之一。

3.2 评价指标与基准模型

3.2.1 评价指标

为适应 GIC 任务的设定，由于不存在预定义的类别集合与语义先验，传统依赖于真实标签的准确率类指标，如 Top-k 准确率、平均精确率平均值（mean average precision, mAP）等，难以全面衡量模型性能。因此，本文从标签合理性与视觉相关性 2 个互补的维度对模型生成标签的质量进行评估。

1）标签合理性

为衡量生成标签是否具备语义一致性与判别有效性，本文引入视觉语言模型 GLM-4v-plus[32] 作为评估代理模型来计算 GLM-4v 分数。GLM-4v-plus 是一个在多个学术基准上超越 GPT-4 的先进模型，本文通过其官方 API 访问该模型并设置统一提示词模板，引导模型对生成标签的合理性进行评分。评分标准包括：生成标签与图像内容的一致性、对多义性或模糊情况的处理能力等，关键提示词如图 5 所示。

Input:
You are a professional label quality evaluator tasked with evaluating the reasonableness of image annotations. Given an image and its corresponding label, please assess the annotation based on the following criteria:
1. **Label consistency with image content**: Does the label accurately reflect the actual content and context of the image, including the type of objects, attributes, and scene information?
2. **Handling of ambiguous cases**: For samples that might be ambiguously situated between two categories, has the label made a clear and reasonable decision?
Based on this assessment, provide a quality score for the annotation on a scale from 0 to 1, where:
0 indicates that you have the Poor agreement with this annotation.
0.010-0.200 indicates you have the Slight agreement with this annotation.
…
0.810-1.000 indicates you have the Almost perfect agreement with this annotation.

你是一位专业的标签质量评估师，任务是评估图像标注的合理性。系统将提供一张图像及其对应的标签，请根据以下标准对该标签进行综合判断：
1. 标签与图像内容的一致性：标签是否准确反映图像中的实际内容与语境，包括物体种类、属性及场景等信息？
2. 对模糊样本的处理能力：对于语义上可能处于两个类别之间的模糊样本，该标签是否做出了清晰且合理的判定？
请基于上述标准，对该标签的质量进行评分，分值范围为 0 到 1，具体含义如下：
•0 表示该标签与图像内容严重不符（完全不一致）
•0.010~0.200 表示标签与图像内容略有一致（轻微一致）
…
•0.810~1.00表示标签与图像内容高度吻合（近乎完全一致）0

Fig. 5 The key prompts of GLM-4v score

图 5 GLM-4v 分数的关键提示词

GLM-4v-plus 从 2 个维度对生成的标签质量进行判断：首先，标签是否与图像内容保持一致，准确反映图像中实体、属性及场景等语义信息；其次，对于图像语义边界模糊或具有多重解释的样本，标签是否展现出清晰、合理的判断能力。在评分机制上，模型输出一个区间为 [0, 1] 的连续数值作为标注合理性分数。为增强可解释性，该分值对应 6 个评价等级，从“完全不一致”到“高度一致”，用于量化标签质量的主观评判。

2）视觉相关性

为从感知角度度量生成标签与图像的对应关系，本文引入视觉相似度指标，衡量生成标签在嵌入空间中是否比真实标签更贴近输入图像。具体而言，

设数据集$D=\{(\boldsymbol{x}_i,y_i,\bar{y}_i)|1\leqslant i\leqslant N\}$, 其中$\boldsymbol{x}_i$为第$i$个图像样本, y_i为生成标签, $\bar{y}_i$为真实标签, N为总样本数。计算图像与生成标签之间的余弦相似度, 并与图像和真实标签的相似度进行比较, 定义视觉相似度:

$$Visual_Similarity=\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[\langle \boldsymbol{x}_i,y_i\rangle>\langle \boldsymbol{x}_i,\bar{y}_i\rangle\right], \quad (7)$$

其中$\mathbb{I}[\cdot]$为指示函数, 当括号内条件成立时取值为1, 否则取0。该指标反映生成标签在视觉语义空间中对图像的“贴合度”。

3.2.2 评估基线设置

为全面评估所提出方法的有效性, 本文选取了2类具有代表性的基线方法进行比较。

1)直接查询式生成(NaivePrompt)。该方法以最简提示词使用VLM进行图像分类。具体地, 对于每一张图像, 仅输入1个自然语言查询作为提示词, 例如:“图中物体的名称是什么?”, 让模型在无任何先验信息或上下文约束的条件下直接生成标签。该方法反映了VLM在零先验、零结构设定下的基本分类能力。

2)引入数据集级弱语义先验的约束生成(ContextPrompt)。为提升生成标签的判别性, 引入了数据集级的语义先验信息, 以缩小模型生成空间中的潜在概念集合。具体实现方式为, 在提示语中加入该图像所属的上位语义类别。例如在Flowers-102数据集中, 将提示语设置为:“图中的花的品种是什么?”, 通过“flower”一词对VLM生成结果的语义空间加以限定。该方法不依赖具体的标签枚举列表, 而是借助自然语言中的语义引导能力, 通过语义先验约束的方式引导VLM进行更精细化的类别预测。

3.3 实验方法

本文所有实验在2块NVIDIA vGPU(32 GB显存)环境下完成。在模型选择上, 采用LLaVA-v1.5-7B作为粒度生成器, 该模型具备良好的图文对齐能力和细粒度理解能力, 用于生成图像在上位类、粗粒度类及细粒度类3个层次的语义标签。同时, LLaVA还作为粒度生成器具备良好的图文对齐能力和细粒度理解能力, 用于生成图像在上位类、粗粒度类及细粒度类3个层次的语义标签。同时, LLaVA还用于自动生成图像描述, 为后续语义扩展提供基础语义。为实现对标签语义表示与图像特征的一致性度量, 使用CLIP-ViT-L/14@336px作为标签过滤器, 通过对图像与文本在共享嵌入空间中的余弦相似度进行计算, 从多个粒度标签中筛选最具代表性的结果。此外, 结合大语言模型LLaMA3-8B-Instruct, 本文设计标签扩展模块, 用于根据图像内容描述及初步标签信息进一步生成语义相近但不完全重复的补充标签集合, 以丰富分类候选集并提升标签多样性和覆盖率。

在主实验中, 标签扩展模块生成的扩展标签的最小数量$K_{\min}=10$, 最大数量$K_{\max}=50$, 该参数的选择以及GMCF对K的敏感性分析在3.7节进行充分讨论。

3.4 GMCF与基线方法的比较

为了全面验证所提出的方法在GIC任务中的有效性, 本文在5个具有代表性的细粒度图像分类数据集上进行了系统评估, 并与2个基线方法进行对比。评估指标包括由大型多模态模型GLM-4v-plus所提供的标签合理性评分以及基于嵌入空间相似度的视觉相关性指标, 涵盖了生成标签在语义一致性与视觉贴合度2个维度的表现。

表2展示了实验结果。GMCF在2个指标上的综合表现均显著优于基线方法。对于标签合理性指标, GMCF在C101, F101, DTD, Pets这4个数据集上均取得最优结果, 得分分别为90.71%, 91.37%, 90.54%, 93.25%, 相较于ContextPrompt方法分别提升了3.20个百分点、3.47个百分点、9.02个百分点、6.62个百分点, 展现出其对图像内容的强表征与语义理解能力。在Flwr数据集上, GMCF得分为93.75%, 略低于NaivePrompt的94.87%, 该现象可以归因于该数据集本身具有显著的类别先验, 使得NaivePrompt所引入的限定词能够直接收敛类别空间, 从而获得更高的匹配率。然而, 在该类任务中GMCF仍具备无先验、无训练的优势, 表现出更强的泛化能力和稳定性。综合5个数据集结果, GMCF在标签合理性的平均得分为91.92%, 相比NaivePrompt和ContextPrompt方法分别提升了2.78个百分点与10.32个百分点, 充分体现出其在无需语义先验约束下仍能稳定生成高质量标签。

而对于视觉相关性指标, GMCF的性能提升更加显著。该指标反映了模型生成标签在嵌入空间中与图像特征的一致程度, 能够衡量语义生成与视觉感知的贴合性。在所有5个数据集上, GMCF均优于2个基线方法。在C101数据集上, GMCF达到80.48%, 较ContextPrompt提升27.58个百分点; 在DTD上则达到86.46%, 较ContextPrompt提高了29.73个百分点; 即便在图像类别相对固定的Flwr数据集上, GMCF也取得了35.11%的得分, 较ContextPrompt有21.51个百分点的性能提升。这说明GMCF所生成的标签不仅语义准确, 而且与图像特征在嵌入空间

Table 2 Performance Comparison Between GMCF and Baseline Methods

表 2 GMCF 与基线方法的性能比较

评价指标	方法	数据集					平均值
		C101	F101	DTD	Pets	Flwr	
标签合理性	NaivePrompt/%	87.51	90.71	84.49	88.10	**94.87**	89.14
	ContextPrompt/%	87.51	87.90	81.52	86.63	64.46	81.60
	GMCF (本文)/%	**90.71**	**91.37**	**90.54**	**93.25**	93.75	**91.92**
	Δ	+3.20	+0.66	+6.05	+5.15	−1.12	+2.78
视觉相关性	NaivePrompt/%	52.90	6.86	54.13	4.98	9.70	25.71
	ContextPrompt/%	52.90	17.22	56.73	25.18	13.60	33.13
	GMCF (本文)/%	**80.48**	**51.18**	**86.46**	**32.95**	**35.11**	**57.23**
	Δ	+27.58	+33.96	+29.73	+7.77	+21.51	+24.10

注：黑体突出显示相同条件下的最优值，下划线突出显示了相同条件下的次优值，Δ显示了 GMCF 与最优基线的性能变化百分点。

中具有更高的结构对齐度，表明其具备良好的图文一致性建模能力。

此外，NaivePrompt 基线方法在多个数据集上得分偏低，在 Pets 和 Flwr 数据集上的视觉相关性分别仅为 4.98% 和 9.70%，说明直接使用通用提示词无法充分激活视觉-语言模型对细粒度语义的表达能力。而 ContextPrompt 虽然引入了数据集级别的语义限定，在一定程度上缩小了预测空间，但整体仍不具备足够的语义多样性与鲁棒性，尤其在 DTD 与 C101 等开放语义场景中表现不佳。

3.5 计算效率分析

为了系统评估 GMCF 框架中不同规模与结构配置的多模态模型组合在推理效率与性能之间的权衡关系，本文在 DTD 数据集上构建了标准组合、轻量化组合以及蒸馏压缩组合 3 类设定，并测试其推理时间与标签合理性、视觉相关性。

实验结果如表 3 所示。从整体实验结果看，不同规模的 LLM 在推理效率与性能之间呈现出反比关系。在标准设定下，LLM 组合（LLaVA(7B)＋LLaMA3(8B)＋ViT-L/14(0.4B)）取得了最优性能，其标签合理性与视觉相关性分别达到 90.54% 和 86.46%，同时整体推理时间也相对较高. 这表明大参数模型在多模态语义对齐与细粒度视觉理解方面具有显著优势，但其计算开销亦不可忽视。

在轻量化设定中，通过引入较小规模的语言模型 Phi-3-mini[33] 及视觉编码器 ViT-B/16 和 RN50，推理效率得到明显提升。其中，LLaVA (7B) + Phi-3-mini (3.8B) + ViT-B/16 (0.15B) 在保持较高标签合理性（89.74%）的同时，将 LLM 与 CLIP 推理时间分别降低至 0.17 s 和 0.09 s，仅带来约 0.8% 与 2.5% 的性能下降。在使用 BLIP-2[34] 替代 LLaVA 后，整体推理时间进一步下降，但性能出现更明显退化（标签合理性降至 84.35%），说明 VLM 主体能力对最终表现具有关键影响。

本文还探索了知识蒸馏对加速 GMCF 推理的有效性。MobileVLM (1.7B)[35]+Distil-Llama (1B)+ MobileCLIP (0.05B)[36] 通过蒸馏 LLaVA, LLaMA, CLIP 将推理时间压缩至最低（0.14 s, 0.09 s, 0.05 s），显著提升了系统的实时性与部署友好性。尽管其标签合理性（85.66%）与视觉相关性（80.23%）较标准模型存在一定差距，但仍保持在可接受范围内，体现出知识蒸馏策略在效率与性能之间取得的有效折中。

Table 3 The Inference Efficiency of GMCF Under Different Model Configurations

表 3 不同模型配置下 GMCF 的推理效率

设定	模型（参数量）			推理时间/s				性能指标/%	
	VLM	LLM	CLIP	T_{VLM}	T_{LLM}	T_{CLIP}	T_{total}	标签合理性	视觉相关性
标准	LLaVA (7B)	LLaMA3 (8B)	ViT-L/14 (0.4B)	0.56	0.27	0.15	0.98	90.54	86.46
轻量	LLaVA (7B)	Phi-3-mini (3.8B)	ViT-B/16 (0.15B)	0.56	0.17	0.09	0.82	89.74	83.92
	BLIP-2 (2.7B)	Phi-3-mini (3.8B)	RN50 (0.1B)	0.27	0.17	0.08	0.52	84.35	79.81
蒸馏	MobileVLM (1.7B)	Distil-Llama (1B)	MobileCLIP (0.05B)	0.14	0.09	0.05	0.28	85.66	80.23

注：B 为模型参数量“10 亿（Billion）”的缩写。

综上, 模型规模缩减能够显著提升推理效率, 但会带来不同程度的性能损失。其中, VLM 的能力对整体性能影响最为显著, 一旦对其进行替换或压缩, 性能下降显著。而 LLM 与视觉编码器的轻量化提供了更灵活的效率优化空间, 这为 GMCF 实际应用中根据资源约束选择合适模型组合提供了有价值参考。

3.6 消融实验

3.6.1 模块有效性分析

为了深入分析 GMCF 框架中各模块的有效性, 本文在 DTD 数据集上设计并开展了一系列消融实验, 重点考察 GMCF 中粒度生成器、标签扩展器与标签过滤器 3 个核心子模块对整体性能的影响。实验中, 本文分别启用或禁用某一模块, 评估其对标签合理性指标和视觉相关性指标的贡献。表 4 展示了实验结果。

Table 4 Ablation Experimental Results of Different Modules of GMCF

表 4 GMCF 不同模块的消融实验结果

模块			评价指标	
粒度生成器	标签扩展器	标签过滤器	标签合理性/%（Δ）	视觉相关性/%（Δ）
×	×	×	81.52	56.73
√	×	×	88.98 (+7.46)	62.23 (+5.50)
√	√	×	83.57 (+2.05)	52.88 (−3.85)
√	×	√	90.33 (+8.81)	80.20 (+23.47)
√	√	√	**90.54 (+9.02)**	**86.46 (+29.73)**

注: 黑体突出当前指标下的最优值, Δ显示了在当前指标下启用部分模块与禁用所有模块的性能变化百分点; “√”表示具备, “×”表示不具备。

首先, 当仅启用粒度生成器时, 模型性能即获得明显提升: 视觉相关性提高了 5.5 个百分点, 标签合理性提升了 7.46 个百分点, 这表明通过 VLM 自动生成具有语义层次的细粒度提示, 能够有效缩小标签候选空间, 并增强标签与图像内容之间的一致性, 从而显著提升标签的合理性与视觉相关性。然而, 当在此基础上引入标签扩展器模块后, 模型性能反而出现一定程度的下降。这一现象表明, 该模块虽然旨在利用 LLM 的语言推理能力扩展语义候选标签, 从而增加语义覆盖面, 但由于其主要依据图像描述进行标签生成, 缺乏直接的视觉语义对齐, 容易引入噪声标签。尤其在未启用标签过滤器的情况下, 系统采用随机选择策略决定最终标签, 进一步放大了噪声标签的负面影响, 造成整体性能下滑。

为验证该模块的必要性, 本文进一步测试了在仅启用粒度生成器和标签过滤器但不使用标签扩展器的情形。实验结果显示该组合虽优于完全无筛选机制的设置, 但仍未达到完整 GMCF 系统的最优性能, 进一步验证了标签扩展器在语义多样性建构方面的积极贡献。最后, 在完整配置中引入标签过滤器后, 系统能够在语义候选集中剔除与图像视觉信息不一致的噪声标签, 仅保留最具图文一致性的结果, 从而保障了标签输出的高质量。

综上, GMCF 各子模块之间具有良好的互补性与协同效应: 粒度生成器提供语义聚焦, 标签扩展器带来语义多样性, 而标签过滤器实现视觉对齐与语义筛选。三者协同运行, 共同推动图像标签生成结果在合理性与视觉一致性方面达到最佳水平。

3.6.2 粒度划分策略分析

为了验证本文所提出的层级粒度设计在 GIC 任务中的优势, 本文在保持模型结构与实验参数设置一致的前提下, 仅替换粒度划分方式, 将本文采用的层级粒度划分(上位层级→粗粒度层级→细粒度层级)分别替换为感知粒度、功能粒度与语义粒度 3 种策略, 从而分析不同划分方式对模型生成质量的影响。

其中, 层级粒度(本文方法)基于知识本体中的上下位包含关系, 将标签划分为上位层级、粗粒度层级与细粒度层级 3 个层次, 强调由大类到小类的逐级细化过程; 感知粒度则从视觉感知角度出发, 将图像划分为整体场景、区域结构与局部细节 3 个层面, 更关注视觉信息的空间分布; 功能粒度侧重于对象或场景的用途表达, 将标签组织为整体功能、通用功能与精确功能, 强调语义中的功能属性; 语义粒度则直接作用于语言表达本身, 将标签划分为模糊语义、细致描述与精确类别, 体现语义表述的精确程度递进。

实验结果如表 5 所示, 层级粒度策略在标签合理性(90.54%)与视觉相关性(86.46%)两项指标上均取得最优表现, 说明基于上下位语义关系的结构化划分能够有效保持与挖掘多个 LLM 协作过程中的语义一致性与跨模态对齐能力。语义粒度策略表现与之接近, 表明单纯从语言精细化角度进行约束同样具有较强有效性, 但在结构一致性上略逊一筹。相比之下, 功能粒度策略取得中等性能, 说明功能属性在一定程度上能够辅助语义表达, 但其与视觉内容的直接对应关系相对较弱。感知粒度策略则表现最差, 尤其在视觉相关性指标上下降明显, 反映出仅依赖视觉空间划分难以支撑高质量的语义生成, 容易导致标签偏向局部特征而忽略整体语义一致性。

综上, GMCF 对粒度划分策略较为敏感, 其中基于语义层级关系的结构化方法能够使其在保证表达

Table 5 Experimental Results of Different Granularity Partitioning Strategies

表 5 不同粒度划分策略的实验结果 %

粒度划分策略	粒度结构	评价指标	
		标签合理性	视觉相关性
层级粒度（本文）	上位层级→粗粒度层级→细粒度层级	**90.54**	**86.46**
感知粒度	整体场景→区域结构→局部细节	86.92	72.57
功能粒度	整体功能→通用功能→精确功能	88.13	77.42
语义粒度	模糊语义→细致描述→精确类别	90.47	85.88

注：黑体数值表示最优值。

灵活性的同时提供更强的全局约束，从而在多模态标签生成任务中取得更优的综合性能。这也进一步说明，引入符合人类认知方式的层级化语义建模，对于提升生成式 VLM 的可控性与稳定性具有重要作用。

3.6.3 VLM，LLM，CLIP 的贡献分析

通过模块移除，本文分析了 VLM, LLM, CLIP 在 GMCF 框架中的相对贡献。具体而言，以完整模型（LLaVA + LLaMA + CLIP）作为基准，对 3 种模型分别进行替换或移除，观察到的性能变化如表 6 所示。

Table 6 Performance of GMCF After Removing VLM, LLM and CLIP

表 6 移除 VLM，LLM，CLIP 后 GMCF 的性能

操作	模型			评价指标	
				标签合理性/%（Δ）	视觉相关性/%（Δ）
GMCF	LLaVA	LLaMA	CLIP	90.54	86.46
移除 VLM	ResNet	LLaMA	CLIP	83.29 (−7.25)	77.45 (−9.01)
移除 LLM	LLaVA	WordNet	CLIP	84.80 (−5.74)	80.74 (−5.72)
移除 CLIP	LLaVA	LLaMA	ResNet+BERT	82.66 (−7.88)	51.13 (−35.33)

注：Δ表示移除 VLM，LLM，CLIP 后的 GMCF 性能与移除前的 GMCF 性能的变化百分点。

从整体结果来看，完整使用 3 种模型在标签合理性与视觉相关性上均取得最优表现，表明三者协同作用能够实现较好的语义生成与视觉对齐能力。当移除 VLM 以 ResNet 替代时，标签合理性下降 7.25 个百分点，视觉相关性下降 9.01 个百分点，表明 VLM 在跨模态语义建模中起到核心作用，其端到端的视觉-语言对齐能力是高质量标签生成的基础。相比传统视觉编码器，VLM 能够直接建模图像与文本之间的语义关联，因此其缺失会同时影响语义合理性与视觉一致性。

移除 LLM 后，本文通过在 WordNet[37] 中查找同义词作为扩展标签，标签合理性与视觉相关性分别下降 5.70 个百分点与 5.72 个百分点，这表明 LLM 主要负责语言生成的流畅性与语义组织能力，在标签表达的规范性与一致性方面具有重要作用。尽管其对视觉信息的直接建模能力有限，但其提供的语言先验对于提升整体语义质量仍具有显著贡献。

最后，本文移除 CLIP，用 ResNet 提取图像特征，用 BERT[38] 提取 LLM 候选词的文本特征，计算二者余弦相似度进行标签筛选。该操作性能退化最为显著，尤其是在视觉相关性指标上下降高达 35.33 个百分点，同时标签合理性也下降 7.88 个百分点，这表明 CLIP 在跨模态对齐过程中起到了关键的“桥梁”作用，其对比学习机制能够有效约束生成标签与图像内容之间的一致性。一旦缺失该模块，模型虽然仍具备一定的语言生成能力，但难以保证生成内容与视觉输入的匹配程度，从而导致严重的语义偏移问题。

综上，3 类模块在 GMCF 框架中具有明显的功能分工：VLM 决定多模态语义建模的上限，LLM 增强语言表达能力与语义组织，而 CLIP 则在视觉-文本对齐中发挥关键约束作用。其中，CLIP 对视觉相关性的影响最为直接且显著，而 VLM 对整体性能的支撑作用最为基础，三者缺一不可，相互协作，共同支撑 GMCF 在开放环境下生成高质量的标签。

3.7 扩展标签数量 K 的影响分析

本文进一步围绕 GMCF 框架中的关键超参数 K，即标签扩展器模块生成的扩展标签数量，展开实验分析，以评估该参数对整体模型性能的影响及其敏感性。该实验在 DTD 数据集上进行。

本文首先在固定 K 的条件下分析了 GMCF 对 K 的敏感性，并确定扩展标签最小生成数量 $K_{\min}$ 与最大生成数量 $K_{\max}$ 的值。然后，本文可视化了 DTD 上所有样本对应的动态 K 值，分析其分布与影响。

固定 K 条件下的实验结果如图 6 所示，K 的变化范围是 $\{1,5,10,20,50\}$。随着 K 的增加，模型在 2 个指标上均表现出一定程度的稳定性，表明 GMCF 对扩展标签数量的敏感性不高，尤其在 $K>10$ 的情况下，性能波动趋于平缓。具体而言，在视觉相关性指标下，当 K 从 1 增加至 10 时，模型性能显著上升，最大增幅可达 1.78 个百分点，表明适当增加候选标签数量有助于提高模型在嵌入空间中的表达匹配能力。然而，当 $K>10$ 时，性能增长趋于稳定，甚至略有波动，最大差异不超过 0.59 个百分点，表明候选标签进一步扩展的边际效应逐渐减弱。此外，在标签合理性评分方面，整体变化范围更小，从 $K=1$ 到 $K=50$ 的最大性能波动仅为 0.44 个百分点，这说明，在标签语义

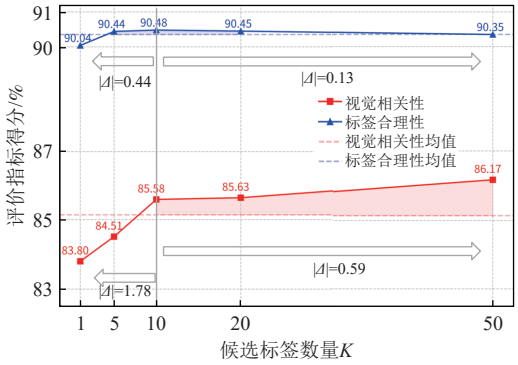


Fig. 6 Effect of the number of candidate labels on performance of GMCF

图 6 生成的候选标签数量对 GMCF 性能的影响

合理性维度上, GMCF 对 K 的变化具有更强的稳定性, 体现出框架中多模块协作机制的调节与缓冲能力。

尽管 GMCF 对 K 的敏感性较低, 自适应 K 值仍然取得了更好的结果。根据图 6 的实验结果, 考虑到模型性能与计算资源的权衡, 本文将 $K_{\min}$ 的值设置为 10, 以保持 GMCF 生成拓展标签的多样性, 将 $K_{\max}$ 的值设置为 50, 以减少模型的推理时间。随后, 本文统计了测试集上自适应 K 的分布情况, 如图 7 所示。实验数据显示, 样本的平均扩展标签数量 $K \approx 15.77$, 其分布呈现显著的置信度引导特征。

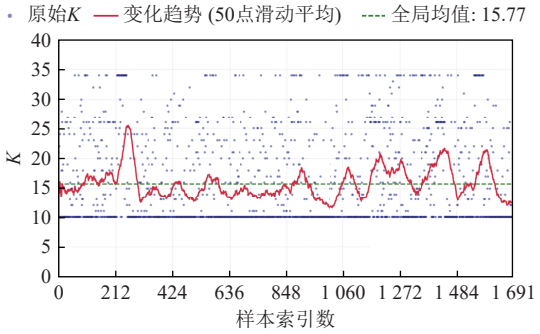


Fig. 7 Visualization of the number of expanded labels dynamically generated by GMCF

图 7 GMCF 动态生成扩展标签数量的可视化

大部分样本的扩展预算收敛于预设最小值 $K_{\min} = 10$。这表明对于特征显著的视觉目标, GMCF 能够通过高置信度反馈快速锁定语义区间, 避免了不必要的计算资源浪费, 保证了系统的推理效率。其次, 分布中最大扩展规模达到 $K_{\max} = 34$。这些高 K 值样本通常对应于细粒度辨析度极低的困难样本。此时, 自适应机制通过调高搜索预算, 引导 LLM 探索更广阔的语义候选空间, 从而弥补了视觉感知的初始不确定性。

综上, 自适应 K 值机制实现了“简单样本快速通过, 困难样本深度搜索”的计算分配逻辑。这种按需分配的策略不仅证明了 GMCF 对语义噪声的控制能力, 也揭示了其在维持全局高精度的同时, 实现计算开销与识别性能的平衡。

3.8 提示模板的鲁棒性分析

为了验证 GMCF 框架在实际应用中对提示模板变化的敏感度, 本文设计了提示词扰动实验。本文在粒度生成器与标签扩展器提示模板的基础上构建 3 种具有显著差异风格但核心含义的提示变体: 简略模板去除所有修饰语, 仅保留核心任务动词以及粒度定义; 详细模板在原有模板的基础上进一步模拟人类冗长的表达, 加入大量背景说明和补充性描述; 改写模板通过同义词替换/表述方式替换在不改变原有模板含义的条件下进行改写。所有模板在 DTD 测试集上分别进行 10 次独立运行, 实验结果如图 8 所示。

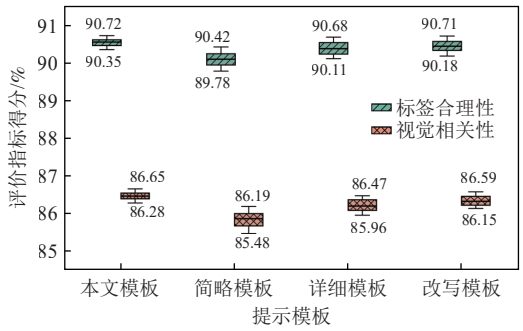


Fig. 8 Robustness analysis of the prompt templates

图 8 提示模板的鲁棒性分析

标签合理性和视觉相关性指标的波动表明, 4 种提示变体展现出了极高的稳定性: 本文模板取得了最优的综合性能(90.54%, 86.46%), 而即使是表现略低的简略模板, 其 2 项指标也分别维持在 90.09% 和 85.84%, 各变体间的最大性能级差仅为 0.45 个百分点(标签合理性)和 0.62 个百分点(视觉相关性)。此外, 所有实验组的标准差均保持在 0.25 以下。其中, 本文模板的方 差最低(0.118, 0.115), 证明了系统在固定指令下的极高确定性; 而在提示词表述大幅改变的情况下, 方差依然控制在 0.17 左右。

综上, GMCF 框架的有效性源于其“感知—扩展—校验”的多模型协作, 提示模板的修改并不能导致性能的大幅度波动。这种对提示模板的鲁棒性使其在开放场景下的实际部署具有更高的可靠性。

3.9 案例分析

3.9.1 有效案例分析

如图 9 所示, 本文比较了各个方法在 5 个样例上的可视化结果。首先, 从整体趋势来看, NaivePrompt 与 ContextPrompt 均表现出明显局限性。NaivePrompt

往往仅停留在粗粒度类别的识别，例如在 C101 样例中仅输出“Brush”，缺乏具体语境信息；ContextPrompt 尽管 NaivePrompt 有所改进，如能够输出“Applying”，但仍难以覆盖图像中的关键场景特征，语义表述较为片面。相比之下，在场景化描述上，GMCF 能够输出与 Ground-Truth 高度契合的标签，展现出对任务语境的深层理解；在细粒度类别识别上，GMCF 成功区分出了具体犬种“Great Pyrenees”，而 ContextPrompt 仅停留在“Husky”这一宽泛类别；在属性识别上，GMCF 能够输出更为全面的纹理与颜色信息。

数据集	输入图像	NaivePrompt	ContextPrompt	GMCF	Ground-Truth
C101		Brush	Applying	Woman applying makeup to her face	Apply Eye Makeup
DTD		Blanket	Striped	Red, green, and white stripes	Banded
Pets		Fence	Husky	Great Pyrenees	Great Pyrenees
F101		Plate	Lasagna	Lasagna and garlic bread	Lasagna
Flwr		Flower	Daisy	Marigold	Marigold

Fig. 9 Visualization of effective cases of GMCF

图 9 GMCF 的有效案例可视化

此外，在 F101 样例中，GMCF 输出整体保持了高度相关性且在生成过程中更倾向于完整描述图像内容，当存在信息扩展时，不仅可以保持对目标类别的正确覆盖，还能识别出图像中的其他信息。

综上，GMCF 不仅具备更强的泛化能力，也在实际应用中展现出更高的可靠性与实用价值。

3.9.2 错误案例分析

为了探索 GMCF 的局限性以提供未来改进方向，本文对数据集中 GMCF 的典型错误案例进行了分析。通过定性分析，可以发现 GMCF 框架的失效受限于缺少局部特征的关注以及复杂场景下的空间竞争。

如图 10 所示，在第 1 个样例中，其真实标签为西番莲花(passion flower)，而模型过度关注于花的颜色与其所处的环境背景，误导 LLM 将这些颜色、背景过度关联至具有相似视觉属性的语义类别，生成如紫色的花、植物特写等扩展标签，丢失了对整体类别的概括，从而导致最终生成的类别没有精确到物体的具体种类。针对这一局限，未来的改进方向可以聚焦于建立全局-局部的语义一致性校验机制，通过增强局部特征在全局上下文中的表现，防止模型因丢失局部细节而产生的语义偏移。

另一方面，在多目标共存的复杂场景中，GMCF

输入图像	真实标签	粒度标签	扩展标签	最终标签
	passion flower	Purple and white flower with green leaves	["Flower", "Purple Flower", "Blooming Plant", "Ornamental Plant", "Tropical Flora", "Garden Plant", "Wildflower", "Botanical Close-up", "Nature Scene", "Leafy Plant", "Fresh Vegetation", "Floral Macro", "Vibrant Colors", "Exotic Flower", "Decorative Flora"]	Purple and white flower with green leaves
	Abyssinian	Cat with yellow eyes and brown fur on a wooden table	["Orange cate", "Cat on table", "Fluffy kitten", "Small cat", "Indoor cat", "Cat with books", "Cat and bowl", "Pet on table", "Home cat", "Kitten looking", "Cat among books", "Dining table cat", "Pet indoors", "Cat with phone", "Kitten near bag", "Home scene cat", "Playful kitten", "Tabletop cat"]	Dining table cat

Fig. 10 Visualization of error cases of GMCF

图 10 GMCF 的错误案例可视化

表现出了明显的空间语义竞争。当画面中同时存在多个显著目标且空间位置高度重叠时，VLM 的特征提取过程往往难以实现精确的目标定位，导致生成的候选标签集包含了背景主体与核心目标的混合语义。由于当前的框架缺乏明确的空间约束，CLIP 在执行图文匹配时容易被视觉面积更大或特征更易对齐的非目标物体所干扰，从而导致预测结果偏离真实目标。未来的研究可以尝试集成空间注意力机制或引入轻量级的目标定位约束，通过在视觉空间上进行精细化的权重分配，强化模型对核心主体的空间聚焦能力，从而在复杂环境下实现精准的语义定位。

4 总 结

本文提出了一种开放且更具挑战性的图像分类任务：生成式图像分类(GIC)。该任务旨在借助大型视觉语言模型，在完全不依赖语义先验信息或外部知识库的前提下，为图像自动生成具有语义区分性的分类标签。针对该任务的关键挑战，本文设计了一种基于粒度引导的多模型协同框架 GMCF。该框架集成了生成式 VLM、对比式 VLM 以及 LLM 三种异构模型，通过粒度引导策略完成对图像语义标签的生成、扩展与筛选，从而在开放标签空间中构建更具信息密度和判别力的分类结果。具体而言，GMCF 首先利用生成式 VLM 从图像中提取多尺度粒度的信息，用以生成初步标签；接着，借助语言模型扩展候选标签以增强语义表达能力；最终通过对比式 VLM 基于视觉语义相关性进行标签筛选与优化，确保标签与图像内容的一致性与判别性。本文在多个标准图像分类数据集上开展了系统性的实证研究，实验结果充分验证了所提方法在无语义先验设定下的分类有效性与标签合理性。同时，相较于依赖预定义类别或外部知识库的传统方法，GMCF 展现出更强的泛化能力和语义扩展潜力，为开放世界图像

理解任务提供了新的研究范式。

作者贡献声明：营鹏、李仲年负责方案整体设计、实验并撰写论文；袁宏广、魏孟负责部分算法思路和实验方法；许新征提出指导意见并修改论文。

参 考 文 献

[1] Addepalli S, Asokan A R, Sharma L, et al. Leveraging vision-language models for improving domain generalization in image classification [C] //Proc of the 41st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 23922−23932

[2] Jia Binbin, Liu Junying, Hang Junyi, et al. Learning label-specific features for decomposition-based multi-class classification[J]. Frontiers of Computer Science, 2023, 17(6): 176348

[3] Dharmasiri A, Naseer M, Khan S, et al. Cross-modal self-training: Aligning images and pointclouds to learn classification without labels [C] //Proc of the 41st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 708−717

[4] Murthy V N, Singh V, Chen T, et al. Deep decision network for multi-class image classification [C] //Proc of the 33rd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2240−2248

[5] Cao Zhi, Chen Enhong, Huang Ye, et al. Learning from crowds with annotation reliability [C] //Proc of the 46th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2023: 2103−2107

[6] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [C] //Proc of the 38th Int Conf on Machine Learning. New York: PMLR, 2021: 8748−8763

[7] Zhang Youcai, Huang Xinyu, Ma Jinyu, et al. Recognize anything: A strong image tagging model [C] //Proc of the 41st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 1724−1732

[8] Lin Yuqi, Chen Minghao, Wang Wenxiao, et al. CLIP is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation [C] //Proc of the 40th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 15305−15314

[9] Lai Zhengfeng, Zhang Haotian, Zhang Bowen, et al. VeCLIP: Improving CLIP training via visual-enriched captions [C] //Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2024: 111−127

[10] Conti A, Fini E, Mancini M, et al. Vocabulary-free image classification [C] // Proc of the 36th Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2023: 30662−30680

[11] Wei Meng, Li Zhongnian, Ying Peng, et al. Seeing the undefined: Chain-of-action for generative semantic labels [C] //Proc of the 33rd ACM Int Conf on Multimedia. New York: ACM, 2025: 3702−3711

[12] Chen Weize, Su Yusheng, Zuo Jingwei, et al. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents [J]. arXiv preprint, arXiv: 2308.10848, 2023

[13] Wang Luqiao, Zhou Yangtao, Li Qingshan, et al. Multi-agent collaborative code reviewer recommendation based on large language model[J]. Journal of Software, 2025, 36(6): 2558−2575 (in Chinese)
(王路桥, 周洋涛, 李青山, 等. 基于大语言模型的多智能体协作代码评审人推荐[J]. 软件学报, 2025, 36(6): 2558−2575)

[14] Liu Haotian, Li Chunyuan, Wu Qingyang, et al. Visual instruction tuning [C] //Proc of the 36th Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2023: 34892−34916

[15] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models [J]. arXiv preprint, arXiv: 2302.13971, 2023

[16] Jiang Zetao, Zhu Wencai, Jin Xin, et al. An image captioning method based on DSC-Net[J]. Journal of Computer Research and Development, 2024, 61(11): 2897−2908 (in Chinese)
(江泽涛, 朱文才, 金鑫, 等. 一种基于双重语义协作网络的图像描述方法[J]. 计算机研究与发展, 2024, 61(11): 2897−2908)

[17] Dai Wenliang, Li Junnan, Li Dongxu, et al. InstructBLIP: Towards general-purpose vision-language models with instruction tuning [C] // Proc of the 36th Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2023: 49250−49267

[18] Chen Jiuhai, Yang Jianwei, Wu Haiping, et al. Florence-VL: Enhancing vision-language models with generative vision encoder and depth-breadth fusion[C] //Proc of the 42nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2025: 24928−24938

[19] Wang Peng, Bai Shuai, Tan Sinan, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution [J]. arXiv preprint, arXiv: 2409.12191, 2024

[20] Ye Qinghao, Xu Haiyang, Ye Jiabo, et al. mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration [C] //Proc of the 41st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 13040−13051

[21] Sun Zeyi, Fang Ye, Wu Tong, et al. Alpha-CLIP: A CLIP model focusing on wherever you want [C] //Proc of the 41st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 13019−13029

[22] Zhou Kaiyang, Yang Jingkang, Loy C C, et al. Learning to prompt for vision-language models[J]. International Journal of Computer Vision, 2022, 130(9): 2337−2348

[23] Sun Ximeng, Hu Ping, Saenko K. DualCoOp: Fast adaptation to multi-label recognition with limited annotations [C] //Proc of the 35th Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2022: 30569−30582

[24] Zhao Peng, Wang Chunyan, Zhang Siying, et al. A zero-shot image classification method based on subspace learning with the fusion of reconstruction[J]. Chinese Journal of Computers, 2021, 44(2):

409−421 (in Chinese)
(赵鹏, 汪纯燕, 张思颖, 等. 一种基于融合重构的子空间学习的零样本图像分类方法[J]. 计算机学报, 2021, 44(2): 409−421)

[25] Novack Z, McAuley J, Lipton Z C, et al. CHiLS: Zero-shot image classification with hierarchical label sets [C] //Proc of the 40th Int Conf on Machine Learning. New York: PMLR, 2023: 26342−26362

[26] Allingham J U, Ren Jie, Dusenberry M W, et al. A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models [C] //Proc of the 40th Int Conf on Machine Learning. New York: PMLR, 2023: 547−568

[27] Kim J, Ji Heng. FINER: Investigating and enhancing fine-grained visual concept recognition in large vision language models [C] //Proc of the 29th Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2024: 6187−6207

[28] Li Feifei, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories [C] //Proc of the 21st IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2004: 178−178

[29] Bossard L, Guillaumin M, Van Gool L. Food-101: Mining discriminative components with random forests [C] // Proc of the 12th European Conf on Computer Vision. Berlin: Springer, 2014: 446−461

[30] Cimpoi M, Maji S, Kokkinos I, et al. Describing textures in the wild [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 3606−3613

[31] Nilsback M E, Zisserman A. Automated flower classification over a large number of classes [C] //Proc of the 6th Indian Conf on Computer Vision, Graphics and Image Processing. Piscataway, NJ: IEEE, 2008: 722−729

[32] GLM Team. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools [J]. arXiv preprint, arXiv: 2406.12793, 2024

[33] Abdin M, Jacobs S A, Awan A A, et al. Phi-3 technical report: A highly capable language model locally on your phone [J]. arXiv preprint, arXiv: 2404.14219, 2024

[34] Li Junnan, Li Dongxu, Savarese S, et al. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models [C] //Proc of the 40th Int Conf on Machine Learning. New York: PMLR, 2023: 19730−19742

[35] Wu Qinzhuo, Xu Weikai, Liu Wei, et al. MobileVLM: A vision-language model for better intra- and inter-UI understanding [C] //Proc of the 29th Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2024: 10231−10251

[36] Vasu P K A, Pouransari H, Faghri F, et al. MobileCLIP: Fast image-text models through multi-modal reinforced training [C] //Proc of the 41st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 15963−15974

[37] Miller G A. WordNet: A lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39−41

[38] Koroteev M V. BERT: A review of applications in natural language processing and understanding [J]. arXiv preprint, arXiv: 2103.11943, 2021



Ying Peng, born in 2001. PhD candidate. Student member of CCF. His main research interest includes machine learning.

营　鹏, 2001 年生。博士研究生。CCF 学生会员。主要研究方向为机器学习。



Li Zhongnian, born in 1990. PhD, lecturer. Member of CCF. His main research interest includes machine learning.

李仲年, 1990 年生。博士, 讲师。CCF 会员。主要研究方向为机器学习。



Yuan Hongguang, born in 2003. Master candidate. His main research interests include machine vision and image recognition.

袁宏广, 2003 年生。硕士研究生。主要研究方向为机器视觉、图像识别。



Wei Meng, born in 1997. PhD. Student member of CCF. His main research interests include machine learning and image recognition.

魏　孟, 1997 年生。博士。CCF 学生会员。主要研究方向为机器学习、图像识别。



Xu Xinzheng, born in 1980. PhD, professor. Senior member of CCF. His main research interests include machine learning and medical image processing.

许新征, 1980 年生, 博士, 教授。CCF 高级会员。主要研究方向为机器学习、医学图像处理。