SemPT: 用于视觉-语言模型的语义提示调优

Xiao Shi[1], **Yangjun Ou***[1], **and Zhenzhong Chen**[2]

[1]School of Computer Science and Artificial Intelligence, Wuhan Textile University
[2]School of Remote Sensing and Information Engineering, Wuhan University

Abstract

对于未见过的类别而言的视觉迁移学习是一个活跃的研究课题，同时也是一项具有挑战性的任务，因为在保持类别特异性表示和获取可迁移知识之间存在固有的冲突。预训练在大量图片-文本对上的视觉语言模型（VLMs）提供了一种可行的解决方案。然而，现有的提示微调方法依赖于稀疏的类别标签或由大型语言模型（LLM）生成的分散描述，这会导致知识表示的碎片化并阻碍可迁移能力。为了解决这一限制，我们引入了语义提示微调（SemPT），一个通过利用跨类别共享的属性级知识来应对泛化挑战的新框架。具体来说，SemPT 采用了双步骤提示策略来指导 LLM 提取共享的视觉属性并生成属性级描述，捕捉超越标签的可迁移语义线索，同时确保结构一致。然后，视觉引导加权应用于属性级描述的嵌入，以减少来自不相关属性的噪声并增强文本嵌入。此外，图像嵌入与标签和属性增强文本嵌入共同对齐，平衡对已见类别的判别力和对未见类别的可迁移能力。考虑到类别曝光的可用性，我们的推理在已见类别中动态选择标准的标签嵌入，而在未见类别中选择属性增强嵌入，以确保有效的适应性。在 15 个基准数据集上的大量实验表明，SemPT 在各种设置中实现了最先进的性能，包括从基础到新颖的泛化、跨数据集迁移、跨域迁移和少样本学习。

1 介绍

视觉语言模型（VLMs），如 CLIP [1]，通过对大规模图文对进行对比学习进行预训练，形成了用于迁移学习任务的基础知识库。然而，在迁移学习中，保留已见类别的类别特定表示和获取未见类别的可迁移知识之间存在基本冲突。为了增强 VLMs 的迁移能力，出现了两种主要方法，即全量微调 [2, 3, 4] 和提示微调 [5, 6, 7, 8]。其中，提示微调方法在少量可学习参数的情况下，可以实现与完全微调方法相媲美甚至更优的性能。

现有的提示调整方法主要采用两种策略来进行文本表示。基于标签的方法 [9, 10, 11, 12] 仅从类别名称中学习提示，能够高效地优化嵌入，但缺乏用于迁移学习的显式属性语义。而基于标签驱动的 LLM 增强方法 [13, 14, 15, 16] 则采用 LLM 生成更丰富的描述并增强语义覆盖，但忽略了跨类别共享属性的建模。由于属性分解不足，生成的描述多样却分散，导致文本嵌入碎片化。

这种碎片化本质上削弱了已学习知识的可迁移性。模型学习将图像嵌入与稀疏标签（例如，图 1 (a) 中的“睡莲”和“矮牵牛”）或不同描述（例如，图 1 (b) 中的“睡莲是一种水生开花植物，具有宽大的浮叶和鲜艳的花朵”）的文本嵌入对齐。然而，这种碎片化的文本嵌入无法在类别之间建立有意义的语义连接（例如，从“睡莲”到“杜鹃”），从而阻止模型在推理过程中对未见过的类别进行泛化。这一差距揭示了有效的迁移学习需要将类别之间共享的属性编码为语义桥梁，从而使模型能够将未见过的类别分解为以前学习的属性组合。

在本文中，我们通过利用共享的属性级知识构建从已见类别到未见类别的可转移语义路径。如图 1 (c) 所示，这种方法将一个分散的文本嵌入景观转化为一个基于共享视觉属性的语义连接空间。共享属性（例如，“叶结构”，“花瓣形状”）被提取出来，并用于为每个类别生成属性级描述（例如，“睡莲：漂浮的叶垫，星状花瓣形状”，“杜鹃花：星状花瓣形状，有光泽的叶面”）。共享属性“星状花瓣形状”在类别之间形成了一条明确的语义链接（例如，从“睡莲”到“杜鹃花”），从而实现了跨越已知-未知边界的知识转移，其中标签锚定已知类别，属性引导未知类别。

为实现这一方法，我们提出了语义提示调优（Semantic Prompt Tuning，简称 SemPT），它由四个关键部分组成。首先，采用两步提示策略，引导大型语言模型（LLM）构建属性级别的知识空间：LLM 最初提取所有类别的共享视觉属性，然后生成基于这些属性的属性级描述。其次，根据视觉相关性对属性级描述进行加权，然后与类别标签文本嵌入融合，形成属性增强文本嵌入。第三，采用双分支监督方案，将图像嵌入对齐到标签和属性增强文本嵌入上，平衡鉴别能力和迁移能力。最后，基于类别曝光进行推断：对已知类别应用标签文本嵌入，对未知类别则使用属性增强文本嵌入，以完成从训练到推断的过渡，实现稳健的识别。

为了验证我们的方法，我们在具有挑战性的评估环境中进行了广泛的实验，包括从基础到新的类别的泛化、跨数据集传输、跨域传输和小样本学习。这些多样化的场景全面测试了在不同类别、数据集、域和数据条件下的泛化能力。SemPT 在所有环境中始终如一地实现了最先进的性能，并表现出与各种 VLM 架构的广泛兼容性，确立了其作为一种通用增强策略的多功能性。

本文的主要贡献如下：

- 语义提示调整（Semantic Prompt Tuning，SemPT）旨在通过利用共享的属性级知识、解决碎片化的知识表示以及实现稳健的跨类别理解来应对迁移学习中的泛化挑战。
- 一种两步提示策略构建了属性级别的文本描述，这些描述为通过专用模块逐步增强文本嵌入构建了可转移的语义路径，从而显著增强了知识转移能力。
- 提出了一种统一的训练-推理适配机制，以定制文本表示来适应类别暴露，有效地平衡泛化和特异性。
- 我们在 15 个基准数据集上进行了大量实验，结果

* 通讯作者: 欧阳军, 电子邮件: yjou@wtu.edu.cn

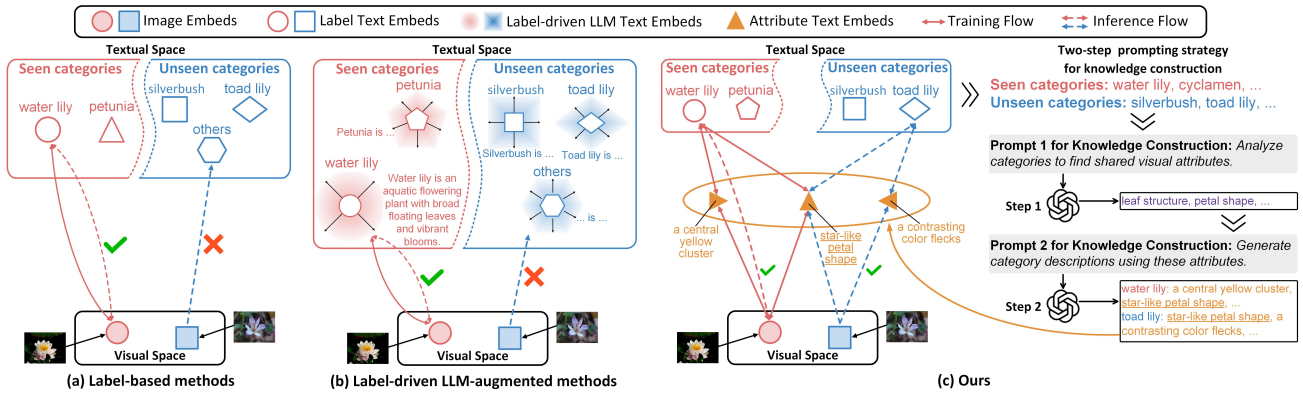

Figure 1: 现有方法与提出的 SemPT 框架的示意图。(a) 基于标签的方法依赖于稀疏类别标签，没有属性建模，阻碍了从已见类别向未见类别的知识转移。(b) 标签驱动的 LLM 增强方法生成描述以扩展语义，但由于属性分解不足导致了知识表达的碎片化。(c) 提出的 SemPT 框架通过两步提示策略明确建模共享的属性级别知识，显著提高了 VLM 的泛化能力。

表明，SemPT 在所有评估设置中均达到了最先进的结果，同时保持了与各种基线的出色兼容性。

本文的其余部分组织如下：第二部分回顾了为本研究奠定基础的相关工作。第三部分详细介绍了提出的 SemPT 框架，包括其四个核心模块。第四部分描述了实验设置，并提供了详细分析的广泛实验结果。第五部分总结了本文内容。

2 相关工作

2.1 视觉-语言模型

视觉-语言模型（VLMs）通过在一个共享的嵌入空间中对齐其表示形式，实现对视觉和文本模态的联合理解。大多数视觉-语言模型采用对比学习作为核心训练目标，在提高匹配的图像-文本对之间的相似性的同时，降低不匹配对之间的相似性。

CLIP [1] 是一项开创性工作，它使用双编码器框架在 4 亿对图像文本对上进行训练，将视觉和文本表示对齐，在各种视觉基准测试中实现了强大的零样本性能。ALIGN [17] 通过使用嘈杂的替代文本注释和更大的数据集扩展了这一范式，展示了简单的架构可以从大规模预训练中显著受益。BLIP [18] 引入了一个新方向，将视觉-语言理解与生成结合在一个统一的预训练管道中，使用自举学习机制在弱监督下提高鲁棒性。

最近的努力集中在增强架构能力和改进模态之间的交互。例如，BLIP-2 [19] 引入了一个具有可学习查询的 Q-Former，以更好地提取相关的视觉特征，而 Flamingo [20] 采用感知重采样器来促进跨模态选择。ViLA [21] 探索更高效的预训练策略，以提高训练效率和下游迁移性能。在这些进展的基础上，MSGM [22] 提出了一种视觉-语言预训练方法，旨在增强具身推理任务中的零样本泛化能力。

这些进展极大地提升了 VLMs 的泛化能力，使得跨模态检索、小样本学习和组合推理等应用成为可能。然而，大多数现有的 VLMs 仍然依赖于全局监督，这限制了它们在新领域中的适应能力，尤其是在资源匮乏或领域转移的环境中。

2.2 基于标签的提示调整

提示调优最初在自然语言处理（NLP）中被提出，作为完全微调的一种参数高效的替代方法，使预训练语言模型能够通过可学习和任务特定的提示适应新任务 [23]。这个想法后来被扩展到视觉语言模型（VLM），其中 CoOp [5] 首次使用连续提示来替代手工制作模板用于零样本图像分类。虽然有效，但 CoOp 依赖于缺乏输入感知的静态提示，这限制了其在未见领域中的泛化能力。

为了应对 CoOp 有限的泛化能力，后续方法集中于动态和多层次的提示适应以及架构创新以增强图像-文本对齐。CoCoOp [6] 引入了基于每个输入图像的动态提示，提高了可迁移性，但仅限于图像基础的适应而没有文本适应。MaPLe [7] 通过联合学习视觉和文本分支的提示扩展了这一思想。它在变换器块中应用了多层提示微调。网络中强制施加跨模态耦合，促使超越最终输出层的更深层次交互。除提示设计外，MMA [10] 和 MMRL [24] 提出了架构改进：MMA 使用轻量级适配器来将特征聚合到共享空间中，促进有效的梯度流动，而 MMRL 构建了用于细粒度令牌级别交互的模态无关令牌空间。类似地，TIPPLE [25] 提出了测试时任务到实例提示学习，以动态适应视觉语言模型。同时，KIM [11] 探索了知识注入策略以增强模型的适应性和鲁棒性。

另一项工作重点在于提升提示微调在分布变化或低资源条件下的泛化能力。PromptSRC [8] 通过互相一致性约束和提示集成来对提示学习进行正则化，以提升泛化性和稳定性，同时使用多样性正则化减少过拟合。ADAPT [26] 将对抗训练与提示调优相结合，提出对抗性双重提示，以解决无监督领域适应问题，并提高对领域转换的鲁棒性。MetaPrompt [9] 通过元学习学习领域不变提示来解决领域转移，通过情节训练使提示能够在保留的领域中泛化。PromptKD [12] 采用一种传导性零样本学习框架，从教师模型中提炼知识到提示中，利用未标记的目标领域数据，使学生预测与教师输出对齐，以在无监督情况下识别未见类别。

这些方法为适应 VLMs 提供了高效的基础，但通常缺乏对共享语义结构的明确建模，这限制了它们的可解释性和泛化能力，尤其是在分布变化和监督稀缺的情况下。

2.3 大型语言模型增强提示微调

大型语言模型（LLMs）的迅速发展，例如 ChatGPT [27]、LLaMA [28]、Gemini [29] 和 Claude [30]，彻底改变了自然语言理解和生成。这些能力激发了将 LLMs 集成到 VLMs 中以改进提示调优的灵感，利用它们的语言知识和推理能力来实现更具信息性和可迁移性的提示。

最近，增强大语言模型（LLM）的提示调优方法通过各种策略提高提示质量和模型适应性。TCA [31] 学习增强了提示的上下文特征，以更好地建模弱监督下的视频异常检测场景。为了捕捉更丰富的语义结构，HPT [13] 提示 LLMs 生成编码多层次关系的层次化类别描述，增强对未知类别的泛化能力。CoPrompt [14] 通过在可训练提示和 LLM 生成特征之间强制一致性来缓解过拟合，使学习的提示与语言学先验一致。CoCoLe [16] 将 LLMs 的高级概念知识提炼到离散的代码本空间中，提供语义上有根据的标记，促进稳健的跨领域泛化。ArGue [15] 将重点转向细粒度属性级语义，将辨别性的描述注入提示中，以增强细微的视觉差异。ProText [32] 更进一步，通过 LLM 生成文字来纯粹监督提示学习，实现零样本迁移，无需标记图像。

这些包含大语言模型的方法通过利用外部语言知识丰富语义表示，增强在视觉语言模型中的适应性和泛化性，从而推进提示学习。然而，它们在建模共享属性层级语义方面的关注有限，导致文本嵌入的碎片化，从而阻碍知识转移，并在分布变化时降低稳健性。

3 方法论

为了解决由碎片化知识表示引起的泛化挑战，我们提出了用于多样化迁移场景的语义提示调优（SemPT）。所提出 SemPT 的整体架构如图 2 所示。

3.1 概述

给定一个由图像 x_j 及其对应标签 y_j 组成的数据集 $\mathcal{D} = \{(x_j, y_j)\}_{j=1}^{M}$，其中标签来自已见类别 $C_s = \{c_i\}_{i=1}^{N_s}$ 或未见类别 $C_u = \{c_i\}_{i=1}^{N_u}$，我们的目标是调整一个基于 CLIP 的视觉语言模型，使其在推理期间能够准确识别 C_s 和 C_u 中的内容。

为此，所提出的 SemPT 框架包括四个关键模块：知识构建（KC）、语义对齐（SA）、特征增强（FE）以及统一的训练-推理适应（UTIA）。

知识构建 (KC)。KC 模块提取共享的视觉属性，并通过使用 LLM 的双步骤提示策略为所有类别生成属性级文本描述。在第一步中，LLM 识别所有类别中的一组共享视觉属性，形成统一的语义词汇 $\mathcal{A} = \{a_m\}_{m=1}^{M_a}$。在第二步中，基于 $\mathcal{A}$，LLM 生成一组属性级描述：$\mathcal{T} = \{t_i^j \mid {}^{j=1,\ldots,S}_{i=1,\ldots,N}\}$，其中 $N = N_s + N_u$ 是类别的总数，S 是每个类别的描述数量，t_i^j 表示类别 $c_i \in C_s \cup C_u$ 的第 j 个属性级描述。

语义对齐（SA）。SA 模块为每个输入图像选择并聚合视觉上最相关的属性级描述。给定图像 x_j，我们通过图像编码器 E_V 获取其嵌入 $\mathrm{v} \in \mathbb{R}^d$，其中 d 是嵌入维度。我们计算 v 与通过文本编码器对属性级描述 $\mathcal{T}$ 进行编码得到的所有属性文本嵌入 $\mathrm{T}_{\text{atr}} \in \mathbb{R}^{N \times S \times d}$ 之间的余弦相似性。对于每个类别，选择前 K 个最相关的描述，并通过基于 softmax 的加权进行聚合，最终得到属性对齐的嵌入 $\hat{\mathrm{T}}_{\text{atr}} \in \mathbb{R}^{N \times d}$。

特征增强（FE）。FE 模块融合了标签文本嵌入和属性级语义。标签文本嵌入 $\mathrm{T}_{\text{lbl}} \in \mathbb{R}^{N\ N \times d}$ 是通过文本编码器对 C_s 和 C_u 的每个类别名称进行编码得到的。然后，这些嵌入通过投影层与 $\hat{\mathrm{T}}_{\text{atr}}$ 结合，随后进行残差连接。这产生了属性增强的文本嵌入 $\mathrm{T}_{\text{enh}} \in \mathbb{R}^{N \times d}$。

统一训练-推理适应（UTIA）。UTIA 模块采用双嵌入策略来弥合训练与推理之间的差距。在训练期间，T_{lbl} 和 T_{enh} 由单独的分类损失进行监督，并通过加权和共同优化。在推理期间，T_{lbl} 用于已知类别以保持判别能力，而 T_{enh} 则应用于未知类别，通过属性级别语义增强泛化能力。

3.2 知识构建

SemPT 框架的基础在于构建可转移的语义表示，这些表示连接了已知类别和未知类别。为此，我们提出了一种两步提示策略，使用 LLM 生成基于共享视觉属性的结构化和低冗余的文本描述。这些描述提供了一个统一和区分性的语义基础，有助于视觉和文本特征的对齐，并支持跨类别的有效知识转移，如图 3 所示。

步骤 1：跨类别属性发现。为了构建一个共享的语义基础，我们首先提示 LLM 识别一组广泛适用于已知（C_s）和未知（C_u）类别的视觉属性。具体来说，我们设计了一个提示模板，指导 LLM 提取涵盖多种语义维度和功能特征的视觉属性。该提示强调视觉维度间的多样性和正交性，以确保全面的语义覆盖并最小化属性冗余。为了提高提取属性的质量，我们在提示中加入了来自 C_s 和 C_u 的类别示例，使 LLM 能够识别在类别谱系上具有普遍性的属性。此过程生成的共享属性词汇 $\mathcal{A}$，作为引导后续描述生成的语义骨架。

步骤 2：属性引导的描述生成。基于 $\mathcal{A}$，我们随后提示 LLM 为每个类别 $c_i \in C_s \cup C_u$ 生成多个简明描述。对于每个类别 c_i，我们提示 LLM 创建 S 个独特描述，其中每个描述突出显示来自 $\mathcal{A}$ 的相关属性的不同组合。每个描述引用一组共享属性，以确保语义一致性的同时捕捉类别特定的视觉特征。为了保持描述的多样性，我们采用不同的提示形式，并明确指示 LLM 避免在同一类别的描述中出现重复措辞。此外，我们限制每个描述的简明性，以便在保留语义丰富性的同时促进高效处理。最终形成一个结构化的属性级别描述集 $\mathcal{T}$。

通过将描述基于共享词汇 $\mathcal{A}$，我们确保了各类别间的语义一致性，同时保持了各自的独特性。这些低冗余的、属性感知的描述 $\mathcal{T}$ 作为可转移的语义锚点，促进了与下游模块中视觉表示的对齐，从而实现从已见类别到未见类别的有效知识转移。

3.3 语义对齐

尽管属性层级的描述 $\mathcal{T}$ 提供了丰富的语义知识，但并不是所有的描述都与视觉内容完全一致。在每个类别的多种描述中，一些具有很强的视觉基础，而另一些可能是多余的、过于抽象的，或者与视觉表现弱相关。为了减轻语义噪声并提高文本监督的相关性，我们提出了一种动态语义对齐机制，能够自适应地强调每个类别中最具视觉对齐力的描述。

每个描述 $t_i^j \in \mathcal{T}$ 使用文本编码器 E_T 进行编码，生成标准化的属性文本嵌入 T_{atr}：

$$\mathrm{T}_{\text{atr}} = \{E_T(t_i^j)\}_{i=1,j=1}^{(N,S)} \in \mathbb{R}^{N \times S \times d} \tag{1}$$

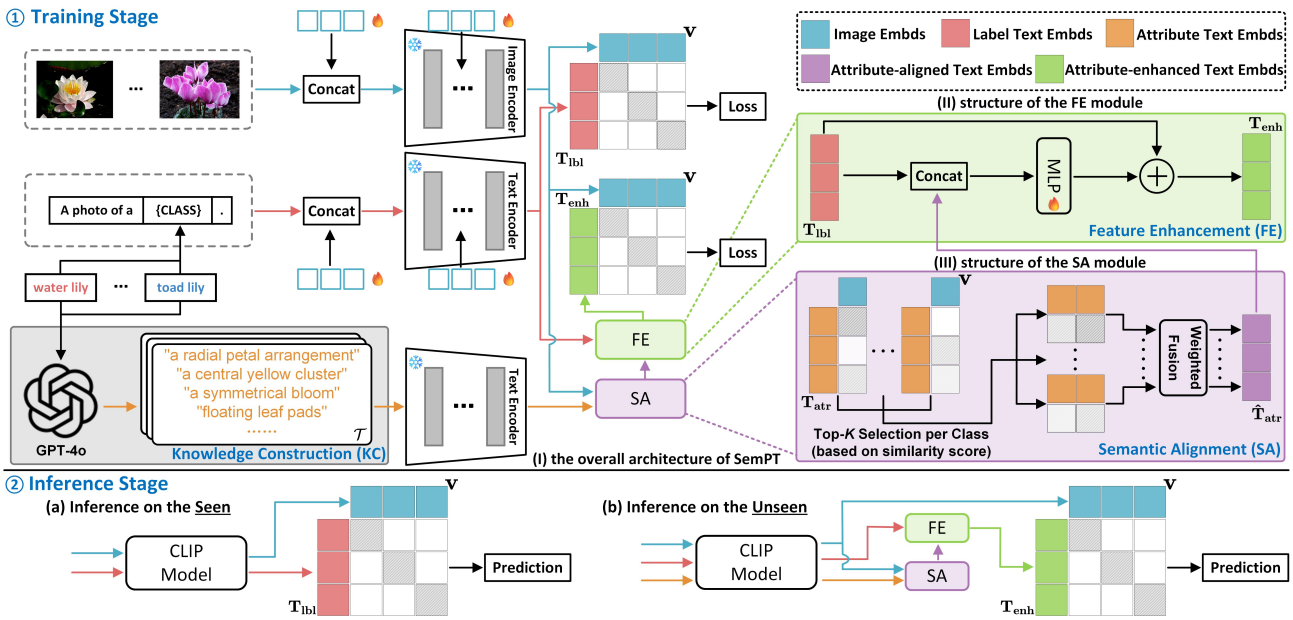

Figure 2: SemPT 框架概述。训练阶段：在 KC 模块中，使用 GPT-4o 通过两步提示策略来提取共享视觉属性，并生成所有类别的属性级描述 $\mathcal{T}$ 。SA 模块根据图像-文本相似性选择每个类别最相关的顶部 K 描述，接着进行 softmax 加权平均以获得属性对齐的文本嵌入 $\hat{\mathrm{T}}_{\mathrm{atr}}^{(i)}$ 。FE 模块通过连接、线性投影和残差连接来整合标签文本嵌入 $\mathrm{T}_{\mathrm{lbl}}$ 和属性对齐文本嵌入 $\hat{\mathrm{T}}_{\mathrm{atr}}$ ，生成属性增强的文本嵌入 $\mathrm{T}_{\mathrm{enh}}$ 。$\mathrm{T}_{\mathrm{lbl}}$ 和 $\mathrm{T}_{\mathrm{enh}}$ 都受到监督以促进稳健识别。推理阶段：标签文本嵌入 $\mathrm{T}_{\mathrm{lbl}}$ 用于 seen 类别，属性增强文本嵌入 $\mathrm{T}_{\mathrm{enh}}$ 用于 unseen 类别。

这一组属性文本嵌入作为类别层次上的语义储库，捕捉多样的语言属性，提供了超越原始类别名称的补充信号。

给定一个输入图像 x_j ，图像编码器 E_V ，与可学习的视觉提示 P_V 结合后，生成一个标准化的图像嵌入 v：

$$\boldsymbol{v} = E_V(x_j \oplus P_V) \in \mathbb{R}^d \tag{2}$$

，其中 $\oplus$ 表示提示集成操作符，该操作符将可学习提示标记附加到输入之前。

为了评估与类别 c_i 相关的每个文本描述 t_i^j 的视觉相关性，我们计算图像嵌入 v 与相应的属性文本嵌入 $\mathrm{T}_{\mathrm{atr}}^{(i,j)}$ 之间的余弦相似度：

$$s_{i,j} = \langle \mathrm{v}, \mathrm{T}_{\mathrm{atr}}^{(i,j)} \rangle, \quad j = 1, 2, \ldots, S \tag{3}$$

，其中 $\langle \cdot, \cdot \rangle$ 表示余弦相似度函数。

基于相似度分数，我们为每个类别 c_i 识别出前 K 个最相关的描述。令 $\Omega_{c_i} = \{j_1, j_2, \ldots, j_K\}$ 表示对应于类别 c_i 的 K 个最高相似度分数的索引集，其中 $K < S$ 。为了进一步反映每个选定描述的相对重要性，我们使用一个温度调节的 softmax 函数计算在 Ω_{c_i} 上的注意力权重：

$$w_{i,j} = \frac{\exp(s_{i,j}/\tau)}{\sum_{j' \in \Omega_{c_i}} \exp(s_{i,j'}/\tau)}, \quad j \in \Omega_{c_i} \tag{4}$$

，其中 τ 是控制注意力分布锐度的温度参数。

对于类别 c_i ，最终的语义对齐文本表示通过所选属性文本嵌入的加权组合获得：

$$\hat{\mathrm{T}}_{\mathrm{atr}}^{(i)} = \sum_{j \in \Omega_{c_i}} w_{i,j} \cdot \mathrm{T}_{\mathrm{atr}}^{(i,j)} \tag{5}$$

这个过程针对所有类别重复，以构建属性对齐的文本嵌入 $\hat{\mathrm{T}}_{\mathrm{atr}}$ ：

$$\hat{\mathrm{T}}_{\mathrm{atr}} = [\hat{\mathrm{T}}_{\mathrm{atr}}^{(1)}, \hat{\mathrm{T}}_{\mathrm{atr}}^{(2)}, \ldots, \hat{\mathrm{T}}_{\mathrm{atr}}^{(N)}] \in \mathbb{R}^{N \times d} \tag{6}$$

这一组属性对齐的嵌入作为每个类别的最终语义表示，整合了从各种候选描述中提炼出的视觉基础和语义信息丰富的信号。

通过结合前 K 选择和基于注意力的加权，所提出的对齐机制确保了最终的属性对齐文本嵌入 $\hat{\mathrm{T}}_{\mathrm{atr}}$ 在语义上信息丰富且视觉上有依据。这有助于更强健的视觉-语义对齐，特别是在开放集场景中，共享属性在识别未见类别中起着至关重要的作用。此外，使用自适应对齐减轻了对任何单一描述的过度依赖，促进了泛化，并减少了由于属性级描述 $\mathcal{T}$ 的多样性和不一致性导致的语义偏差。

3.4 特征增强

为了进一步增强类别表示，我们在语义属性对齐的文本嵌入的基础上引入了一种特征增强机制，以构建更具表现力和泛化能力的类别表示。虽然属性级别描述 $\mathcal{T}$ 包含细粒度、视觉基础的语义，有助于跨类别的泛化，但从类别名称得出的标签文本嵌入则提供了简洁且语义一致的锚点，反映了每个类别的规范身份。为了利用这两种表示的互补优势，我们提出了一种自适应整合策略，通过将细粒度的语义属性与来自属性级别描述 $\mathcal{T}$ 的标签文本嵌入整合来丰富标签文本嵌入。

令 $C = \{c_i\}_{i=1}^N$ 表示 N 类别名称的集合。每个类别标签 c_i 使用一个可学习的文本提示 P_T 进行编码，并结合基于模板的描述，从而生成归一化的标签文本嵌入 $\mathrm{T}_{\mathrm{lbl}}^{(i)}$ ：

$$\mathrm{T}_{\mathrm{lbl}}^{(i)} = E_T(P_T \oplus \text{“a photo of a } \{ c_i \} \text{.”}) \tag{7}$$

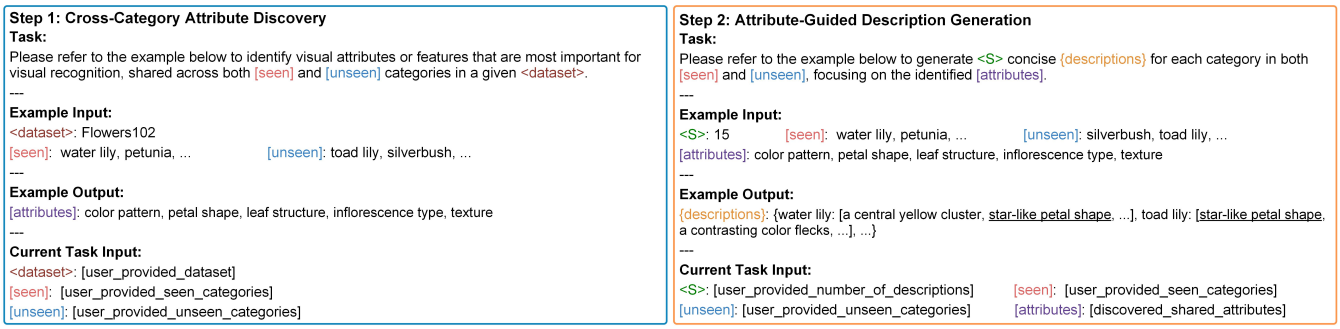

Figure 3: 利用 LLM 进行知识构建的两步提示策略。第一步识别已知类和未知类中共享的多样且互补的视觉属性。第二步为每个类别生成低冗余的、属性引导的描述，为后续选择与视觉表现一致的信息文本特征提供全面的语义基础。

完整的标签文本嵌入集 $\mathrm{T}_{\mathrm{lbl}}$ 因此定义为：

$$\mathrm{T}_{\mathrm{lbl}} = [\mathrm{T}_{\mathrm{lbl}}^{(1)}, \mathrm{T}_{\mathrm{lbl}}^{(2)}, \ldots, \mathrm{T}_{\mathrm{lbl}}^{(N)}] \in \mathbb{R}^{N \times d} \tag{8}$$

为了引入互补的语义信号，我们通过将每个 $\mathrm{T}_{\mathrm{lbl}}^{(i)}$ 与其对应的、由语义对齐模块获得的语义属性对齐文本嵌入 $\hat{\mathrm{T}}_{\mathrm{atr}}^{(i)}$ 相结合来增强其功能。具体而言，这两个嵌入被串联并通过多层感知器（MLP）传递，以学习一个联合投影：

$$\tilde{\mathrm{Z}}^{(i)} = \mathrm{MLP}([\mathrm{T}_{\mathrm{lbl}}^{(i)}; \hat{\mathrm{T}}_{\mathrm{atr}}^{(i)}]) \in \mathbb{R}^{d} \tag{9}$$

，其中 MLP 用于将类别特定的身份和跨类别的语义非线性地融合到统一的嵌入空间中，$[\cdot\,;\,\cdot]$ 表示在嵌入维度上的向量连接。

然而，直接使用 $\tilde{\mathrm{Z}}^{(i)}$ 可能会稀释原始标签表示中编码的核心身份。为了在保留基本类别语义的同时仍然受益于辅助属性，我们引入了一种残差连接：

$$\mathrm{T}_{\mathrm{enh}}^{(i)} = (1-\alpha) \cdot \mathrm{T}_{\mathrm{lbl}}^{(i)} + \alpha \cdot \tilde{\mathrm{Z}}^{(i)} \tag{10}$$

，其中 $\alpha \in [0,1]$ 是一个超参数，控制标签文本嵌入 ${\mathrm{T}_{\mathrm{lbl}}}^{(i)}$ 和丰富表示 $\tilde{\mathrm{Z}}^{(i)}$ 之间的平衡。该公式确保增强的嵌入在灵活结合来自文本描述的辅助知识时，保持与原始标签语义的强连接。

通过将这一细化过程应用于所有 N 类别，我们得到了属性增强文本嵌入集 $\mathrm{T}_{\mathrm{enh}}$ ：

$$\mathrm{T}_{\mathrm{enh}} = [\mathrm{T}_{\mathrm{enh}}^{(1)}, \mathrm{T}_{\mathrm{enh}}^{(2)}, \ldots, \mathrm{T}_{\mathrm{enh}}^{(N)}] \in \mathbb{R}^{N \times d} \tag{11}$$

这通过创建丰富的表示来实现有效的知识转移，这些表示在保留类别特定语义的同时结合了共享的视觉属性。残差公式确保了训练的稳定性，并防止了基本类别信息的稀释，使得 $\mathrm{T}_{\mathrm{enh}}$ 在迁移学习中特别有效，其中属性级别的知识对于泛化到未见过的类别至关重要。

3.5 统一训练-推理适应

为了调和类别特定的区分和跨类别的泛化这两个目标，我们引入了一个双嵌入训练和推理框架。该框架通过嵌入类型分离训练目标并对齐推理决策：标签文本嵌入 $\mathrm{T}_{\mathrm{lbl}}$ 被优化用于已知类别识别，而增强属性的文本嵌入 $\mathrm{T}_{\mathrm{enh}}$ 则编码用于未知类别泛化的可转移语义。这个统一的框架确保在训练期间任务对齐的优化，并在推理期间进行自适应嵌入选择，支持跨类别边界的稳健和灵活的知识转移。

3.5.1 训练阶段

泛化的核心挑战在于学习能够同时在区分已见类别和为未见类别编码可转移知识方面表现出色的表示。传统的单嵌入方法面临固有的权衡：为已见类别区分优化的嵌入可能缺乏泛化所需的语义丰富性。相反，具有属性级语义丰富的嵌入可能会损害区分精度。为了解决这一难题，我们提出了一种双嵌入训练策略，将这些竞争目标解耦为互补的优化路径。

我们通过分别的分类损失共同优化两种互补的文本表示，使模型能够捕捉到精确的类别区分和丰富的跨类别语义知识。

具体来说，第一项损失作用于标签文本嵌入 $\mathrm{T}_{\mathrm{lbl}}$ ，鼓励与真实类别对齐的清晰和显著的决策边界：

$$\mathcal{L}_{\mathrm{lbl}} = -\log \frac{\exp(\langle \mathrm{v}, \mathrm{T}_{\mathrm{lbl}}^{(y)} \rangle / \tau)}{\sum_{j=1}^{N} \exp(\langle \mathrm{v}, \mathrm{T}_{\mathrm{lbl}}^{(j)} \rangle / \tau)}, \tag{12}$$

，其中 y 是真实标签，$\langle \cdot, \cdot \rangle$ 表示余弦相似度，τ 是温度参数。

同时，为了结合跨越单个类别的更丰富的语义关系，我们优化了增强属性的文本嵌入 $\mathrm{T}_{\mathrm{enh}}$ ，它编码了跨类别共享和可转移的语义属性：

$$\mathcal{L}_{\mathrm{enh}} = -\log \frac{\exp(\langle \mathrm{v}, \mathrm{T}_{\mathrm{enh}}^{(y)} \rangle / \tau)}{\sum_{j=1}^{N} \exp(\langle \mathrm{v}, \mathrm{T}_{\mathrm{enh}}^{(j)} \rangle / \tau)}. \tag{13}$$

这种双重目标设计确保 $\mathrm{T}_{\mathrm{lbl}}$ 专注于最大化类别间的分离，以便精确分类已知类别，而 $\mathrm{T}_{\mathrm{enh}}$ 学习捕捉共享的语义属性，从而促进知识向未见类别的转移。独立的优化路径防止了区分性和可转移性目标之间的相互干扰。

总体训练目标结合了两种损失，采用自适应加权：

$$\mathcal{L} = (1-\beta) \cdot \mathcal{L}_{\mathrm{lbl}} + \beta \cdot \mathcal{L}_{\mathrm{enh}} \tag{14}$$

，其中 $\beta \in [0,1]$ 控制辨别精度和可转移语义之间的平衡。

在训练过程中，我们仅优化可学习的参数，同时保持骨干 CLIP 编码器不变。具体而言，可学习的参数包括视觉提示 P_V 、文本提示 P_T ，以及特征增强模块中的线性投影层 MLP 。冻结的骨干编码器（E_V 和 E_T ）保留了预训练的跨模态对齐知识。

3.5.2 推理阶段

推理中的核心挑战在于进行类别预测时，既能保持已见类别的辨别准确性，又能实现对未见类别的可靠泛化。传统

方法通常依赖于对所有类别使用单一嵌入空间，这些方法要么偏重于已见类别的辨别，要么过度强调语义泛化，导致表现不理想。为了克服这一局限性，我们提出了一种双嵌入推理策略，该策略根据类别的曝光情况，自适应地在基于标签的嵌入和增强属性的嵌入之间进行选择，从而使决策机制与每个类别的性质保持一致。

在推理过程中，我们采用自适应嵌入选择策略，根据类别在训练中是否出现来动态分配类别特定的文本嵌入。具体来说，用于类别 c_i 的推理的文本嵌入 $\mathrm{T}_{\mathrm{inf}}^{(i)}$ 定义为：

$$\mathrm{T}_{\mathrm{inf}}^{(i)} = \begin{cases} \mathrm{T}_{\mathrm{lbl}}^{(i)}, & \text{if } c_i \in C_s, \\ \mathrm{T}_{\mathrm{enh}}^{(i)}, & \text{if } c_i \in C_u. \end{cases} \tag{15}$$

，其中 $\mathrm{T}_{\mathrm{inf}}{}^{(i)}$ 表示用于推理的类别特定文本嵌入。

因此，预测类别 $\hat{y}$ 由以下公式确定：

$$\hat{y} = \arg \max_{c_i \in C_s \cup C_u} \langle \mathrm{v}, \mathrm{T}_{\mathrm{inf}}^{(i)} \rangle \tag{16}$$

这种自适应推理策略利用了两种嵌入类型的互补优势，以支持准确和具有普遍适用性的预测。对于已知类别，通过直接视觉监督优化的标签文本嵌入 $\mathrm{T}_{\mathrm{lbl}}$ 维持了明确的决策边界和强大的辨识能力。相比之下，对于未知类别，属性增强嵌入 $\mathrm{T}_{\mathrm{enh}}$ 通过捕捉跨类别的共享语义结构实现有效的知识转移，从而弥补了直接监督的缺失。

4 实验

4.1 数据集

我们在 15 个不同的数据集上评估我们的方法，这些数据集根据它们在评估中的角色分为两个功能组：标准基准组和 ImageNet 系列。标准基准组由 11 个广泛使用的数据集组成，涵盖广泛的识别任务和视觉领域。这些包括一般的对象分类（例如，ImageNet [33]，Caltech101 [34]），细粒度识别（例如，StanfordCars [35]，OxfordPets [36]，FGVCAircraft [37]，Flowers102 [38]，Food101 [39]），场景理解（SUN397 [40]），纹理分类（DTD [41]），遥感（EuroSAT [42]），以及人体动作识别（UCF101 [43]）。这些数据集是用于在标准条件下评估跨各种视觉领域和任务类型的泛化性的重要测试平台。ImageNet 系列包括 ImageNet 及其四个具有挑战性的变体：ImageNet-V2 [44]，ImageNet-Sketch [45]，ImageNet-A [46]，和 ImageNet-R [47]。这些数据集具有与 ImageNet 相同的标签空间，但引入了显著的分布变化。这个组专门用于在受控扰动和真实世界分布变化下评估模型的鲁棒性和领域可迁移性。

4.2 评估设置

我们在四个不同的环境中评估我们的方法，每个环境都旨在考察泛化的不同方面。选择这些环境是为了在不同的监督和分布变化水平下全面评估模型的性能。

在基础到新颖泛化设置中，每个数据集中的类别被均匀地分为基础和新颖子集。模型仅在基础类别上进行训练，并在没有任何微调的情况下评估基础和新颖类别。这个评估衡量了模型在之前见过的类别上保留辨识能力的同时，将其泛化到以前未见过的类别的能力。这反映了模型将学习到的表示转移到识别新概念上的能力，从而在专业化和泛化之间达到平衡。

在跨数据集迁移设置中，模型在一个数据集上进行训练，并直接在一组完全不同的目标数据集上进行评估，无需任何适应或微调。该评估严格测量模型学习领域无关和可迁移图像表示的能力，这些表示可以有效地跨具有多样数据分布、标签空间和视觉领域的数据集进行泛化。它为模型在以前未见的环境中部署时的鲁棒性和适应性提供了关键见解。

在跨域迁移设置中，模型在原始数据集上进行训练，并在多个变体上进行评估，而无需进一步的适应。这些变体与 ImageNet 共享相同的标签空间，但在风格、纹理或构图上引入了显著的视觉变化。此评估关注模型对自然分布变化的弹性，同时保持类别识别的一致性，这对于在视觉条件变化的现实世界部署中至关重要。它展示了模型在遇到常见于实际应用中的视觉变化时保持稳健性能的能力。

4.2.1 少样本学习

在小样本学习设定中，模型仅通过每个类别有限数量的已标记样本进行训练，并在整个测试集上进行评估。为了模拟不同程度的数据稀缺性，我们考虑了五种样本配置：每个类别 1、2、4、8 和 16 个样本。这种评估直接考察了模型的样本效率及其从最小化监督中泛化的能力。它揭示了模型在训练数据极度有限时，如何有效利用先验知识和语义理解来实现出色的表现。

我们在标准基准上进行基础到新颖的泛化、跨数据集传输和小样本学习设置，而跨域传输设置则在 ImageNet 系列上执行。

4.3 实现细节

所有实验均在 NVIDIA RTX 3090 GPU 上使用 PyTorch [48] 框架实现。对于我们的方法，我们设置关键超参数如下：基于相似性选择的顶级属性文本嵌入的数量在所有实验中固定为 $K = 2$。控制我们选择性文本特征集成贡献的方程式 10 中的平衡系数 α 被设置为 0.2。此外，调节所提出的特征对齐目标影响的方程式 14 中的权重因子 β 设置为 0.4。对于基线集成，我们在基础到新颖泛化、跨数据集传输和跨域传输实验中使用 PromptKD [12] 和 MMRL [24] 与我们的 SemPT 结合，而小样本学习实验使用 MMRL 与 SemPT。为公平比较，所有基线方法都通过严格遵循原始参数配置忠实重现，包括训练轮数、批量大小、学习率、优化器选择等。这确保了性能差异可以归因于我们提出的方法，而非实现差异。

4.4 实验结果

在本节中，我们展示并分析了我们提出的方法在四种不同评估任务中的实验结果。我们将我们的方法与最先进的基线进行比较，并提供详细的性能分析。

我们针对两类基线方法评估我们的方法。第一类包括传统的提示调整方法：CoOp [5]、CoCoOp [6]、MaPLe [7]、PromptSRC [8]、MetaPrompt [9]、MMA [10]、KIM [11]、PromptKD [12] 和 MMRL [24]。第二类包括结合大型语言模型（LLM）的方法：HPT [13]、CoPrompt [14]、ArGue [15]、CoCoLe [16] 和 ProText [32]。为了确保与这些基于 LLM 的基线方法进行公平比较，我们的方法通过利用 LLM 生成的描述来遵循相同的实验设置，适用于所有类别。

Table 1: 在基础到新颖的泛化设置下，我们的方法与其他现有方法的比较。

Method	Average			ImageNet			Caltech101			OxfordPets		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CoOp (IJCV2022)	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoCoOp (CVPR2022)	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43
MaPLe(CVPR2023)	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58
PromptSRC(ICCV2023)	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30
MetaPrompt(TIP2024)	83.65	75.48	79.09	77.52	70.83	74.02	98.13	94.58	96.32	95.53	97.00	96.26
MMA(CVPR2024)	83.20	76.80	79.87	77.31	71.00	74.02	98.40	94.00	96.15	95.40	98.07	96.72
KIM(TIP2024)	84.70	76.10	80.20	77.90	69.00	73.20	98.40	92.60	95.40	94.70	97.40	96.00
PromptKD(CVPR2024)	86.96	80.73	83.73	80.83	74.66	77.62	98.91	96.65	97.77	96.30	98.01	97.15
MMRL(CVPR2025)	85.68	77.16	81.20	77.90	71.30	74.45	98.97	94.50	96.68	95.90	97.60	96.74
Using LLM:												
HPT(AAAI2024)	84.32	76.86	80.23	77.95	70.74	74.17	98.37	94.98	96.65	95.78	97.65	96.71
CoPrompt(ICLR2024)	84.00	77.23	80.48	77.67	71.27	74.33	98.27	94.90	96.55	95.67	98.10	96.87
ArGue(CVPR2024)	83.77	78.74	81.18	76.95	71.86	74.32	98.63	94.70	96.63	96.23	98.59	97.40
CoCoLe(ECCV2024)	85.22	80.31	82.70	79.25	74.58	76.84	98.17	95.67	96.90	96.21	98.55	97.37
ProText(AAAI2025)	72.95	76.98	74.91	75.00	71.38	73.14	98.06	95.63	96.83	94.95	98.00	96.45
PromptKD w /SemPT(Ours)	87.08	81.21	84.04	80.71	74.82	77.65	98.77	96.77	97.76	96.38	98.37	97.36
MMRL w /SemPT(Ours)	85.96	77.54	81.53	77.74	71.62	74.55	98.91	94.75	96.79	95.84	97.72	96.77

Method	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CoOp(IJCV2022)	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
CoCoOp(CVPR2022)	70.49	73.59	72.01	94.87	71.75	81.71	90.70	91.29	90.99	33.41	23.71	27.74
MaPLe(CVPR2023)	72.94	74.00	73.47	95.92	72.46	82.56	90.71	92.05	91.38	37.44	35.61	36.50
PromptSRC(ICCV2023)	78.27	74.97	76.58	98.07	76.50	85.95	90.67	91.53	91.10	42.73	37.87	40.15
MetaPrompt(TIP2024)	76.34	75.01	75.48	97.66	74.49	84.52	90.74	91.85	91.29	40.14	36.51	38.24
MMA(CVPR2024)	78.50	73.10	75.70	97.77	75.93	85.48	90.13	91.30	90.71	40.57	36.33	38.33
KIM(TIP2024)	81.00	73.40	77.00	97.80	76.30	85.70	90.50	91.10	90.80	42.80	37.70	30.10
PromptKD(CVPR2024)	82.80	83.37	83.13	99.42	82.62	90.24	92.43	93.68	93.05	49.12	41.81	45.17
MMRL(CVPR2025)	81.30	75.07	78.06	98.97	77.27	86.78	90.57	91.50	91.03	46.30	37.03	41.15
Using LLM:												
HPT(AAAI2024)	76.95	74.23	75.57	98.17	78.37	87.16	90.46	91.57	91.01	42.68	38.13	40.28
CoPrompt(ICLR2024)	76.97	74.40	75.66	97.27	76.60	85.71	90.73	92.07	91.40	40.20	39.33	39.76
ArGue(CVPR2024)	75.06	74.18	74.62	98.62	77.96	87.08	91.42	92.40	91.91	41.29	38.80	40.41
CoCoLe(ECCV2024)	80.32	78.84	79.57	97.72	81.04	88.60	92.23	94.28	93.24	43.86	42.65	43.25
ProText(AAAI2025)	64.54	76.08	68.84	74.36	78.44	76.35	90.20	91.98	91.08	30.91	34.13	32.44
PromptKD w /SemPT(Ours)	82.73	84.30	83.51	99.52	82.75	90.36	92.50	93.74	93.12	49.40	42.35	45.60
MMRL w /SemPT(Ours)	81.89	75.48	78.55	98.86	76.84	86.47	90.88	91.76	91.32	46.23	38.31	41.90

Method	SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CoOp(IJCV2022)	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
CoCoOp(CVPR2022)	79.74	76.86	78.27	77.01	56.00	64.85	87.49	60.04	71.21	82.33	73.45	77.64
MaPLe(CVPR2023)	80.82	78.70	79.75	80.36	59.18	68.16	94.07	73.23	82.35	83.00	78.66	80.77
PromptSRC(ICCV2023)	82.67	78.47	80.52	83.37	62.97	71.75	92.90	73.90	82.32	87.10	78.80	82.74
MetaPrompt(TIP2024)	82.26	79.04	80.62	83.10	58.05	68.35	93.53	75.21	83.38	85.33	77.72	81.35
MMA(CVPR2024)	82.27	78.57	80.38	83.20	65.63	73.38	85.46	82.34	83.87	86.23	80.03	82.20
KIM(TIP2024)	83.10	78.50	80.7	84.10	66.20	74.10	94.80	79.00	86.20	86.30	75.90	80.80
PromptKD(CVPR2024)	83.69	81.54	82.60	85.84	71.37	77.94	97.54	82.08	89.14	89.71	82.27	86.10
MMRL(CVPR25)	83.20	79.30	81.20	85.67	65.00	73.82	95.60	80.17	87.21	88.10	80.07	83.89
Using LLM:												
HPT(AAAI2024)	82.57	79.26	80.88	83.84	63.33	72.16	94.24	77.12	84.82	86.52	80.06	83.16
CoPrompt(ICLR2024)	82.63	80.03	81.31	83.13	64.73	72.79	94.60	78.57	85.84	86.90	79.57	83.07
ArGue(CVPR2024)	81.89	80.48	81.18	80.33	67.03	73.08	95.10	90.68	92.84	86.00	79.43	82.58
CoCoLe(ECCV2024)	83.97	82.24	83.10	82.46	68.38	74.76	95.03	84.17	89.27	88.30	83.05	85.60
ProText(AAAI2025)	76.14	79.14	77.61	63.08	61.59	62.33	59.71	80.97	68.73	75.54	79.50	77.47
PromptKD w /SemPT(Ours)	83.81	81.69	82.74	87.03	71.62	78.58	97.40	84.46	90.47	89.62	82.47	85.90
MMRL w /SemPT(Ours)	83.27	79.69	81.44	86.69	65.22	74.44	97.29	80.77	88.26	87.94	80.82	84.23

Table 2: 我们的方法与其他现有方法在跨数据集迁移设置下的比较。

Method	Source	Target										
	ImNet	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	EuroSAT	UCF	Avg.
CoOp (IJCV2022)	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CoCoOp (CVPR2022)	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
MaPLe (CVPR2023)	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
PromptSRC (ICCV2023)	71.27	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75	65.81
MMA (CVPR2024)	71.00	93.80	90.30	66.13	72.07	86.12	25.33	68.17	46.57	49.24	68.32	66.61
KIM (TIP2024)	73.52	94.05	90.65	64.20	71.10	85.60	25.20	67.10	45.25	51.45	68.60	66.32
PromptKD (CVPR2024)	72.42	93.61	91.59	73.93	75.33	88.84	26.24	68.57	55.08	63.74	76.39	71.33
MMRL (CVPR2025)	72.03	94.67	91.43	66.10	72.77	86.40	26.30	67.57	45.90	53.10	68.27	67.25
Using LLM:												
HPT(AAAI2024)	71.72	94.20	92.63	66.33	74.84	86.21	25.68	68.75	50.87	47.36	70.50	67.74
CoPrompt(ICLR2024)	70.80	94.50	90.73	65.97	72.30	86.43	24.00	67.57	47.07	51.90	69.73	67.00
CoCoLe(ECCV2024)	73.88	95.88	91.93	67.79	74.17	87.97	28.83	68.75	49.26	51.75	72.78	68.91
ProText(AAAI2025)	69.80	94.81	91.01	66.60	72.35	86.66	24.72	67.34	47.93	51.86	69.60	67.23
PromptKD w /SemPT (Ours)	72.56	93.92	92.17	75.24	75.52	89.31	28.26	68.76	57.03	65.23	75.64	72.11
MMRL w /SemPT (Ours)	72.21	93.97	91.94	66.83	73.32	86.64	27.96	67.67	46.91	54.45	69.04	67.87

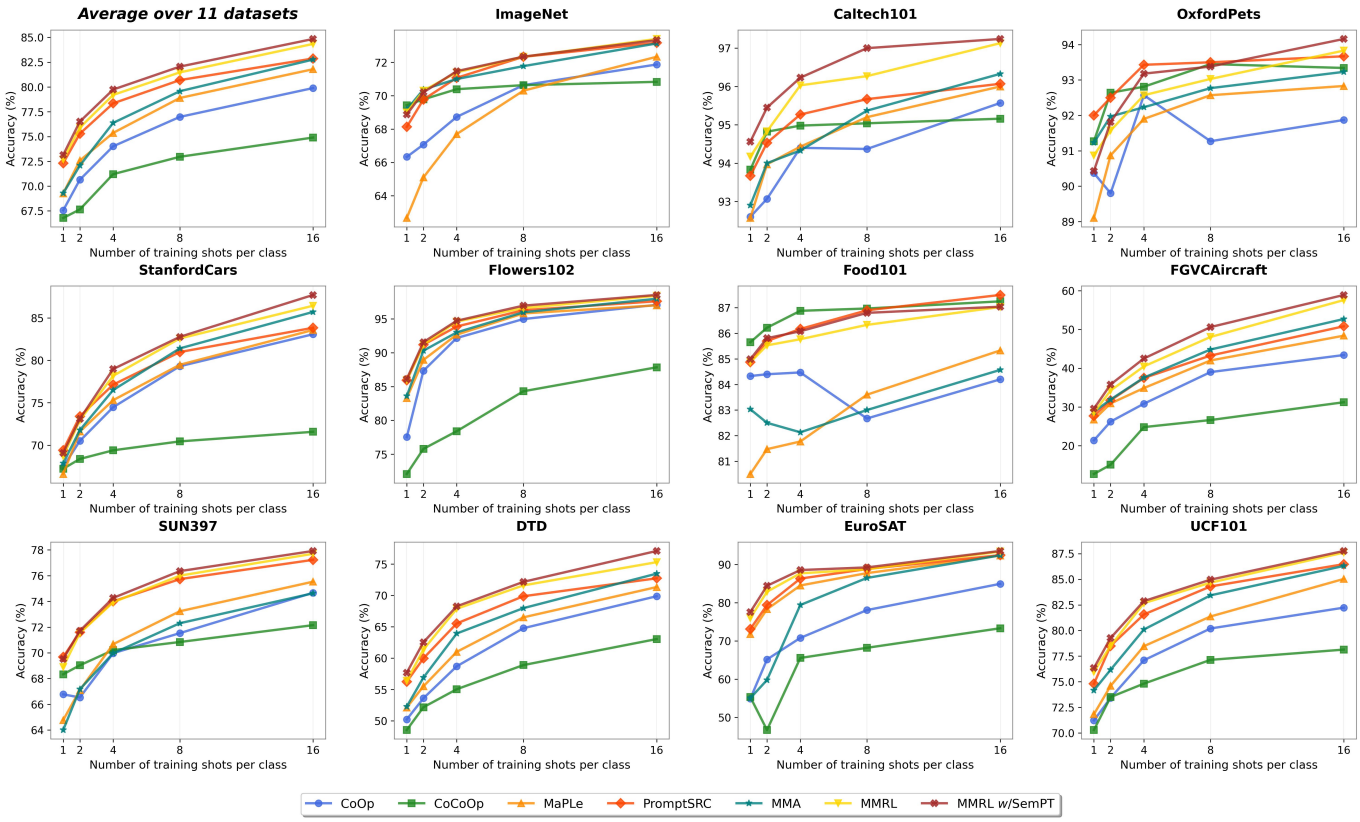

Figure 4: 在小样本学习设定下我们的方法与其他现有方法的比较。

4.4.1 基础到新颖的泛化结果

表 1 展示了在基础到新颖的泛化设置下，我们的 SemPT 方法与现有最先进方法在 11 个数据集上的综合结果。我们的方法在结合强基线方法时表现出一致的改进。具体来说，PromptKD w /SemPT 以平均调和均值（HM）84.04 % 达到最佳整体性能，相比原始的 PromptKD 方法显著提高了 0.31 % 。这种增强尤其显著，因为 PromptKD 已经是该设置中最强的基线之一。同样，MMRL w /SemPT 在大多数数据集上显示出改进，取得了有竞争力的结果，同时保持了基础方法的稳健性。

实验结果揭示了我们方法有效性的一些重要见解。首先，我们的语义提示调整方法在基础类别和新类别中始终提高了性能，这证实了一个假设，即由 LLM 生成的语义描述提供了语义丰富的补充信息，从而增强了少样本学习的能力。其次，性能提升在像 StanfordCars 和 OxfordPets 这样的细粒度分类数据集上尤为明显，这些数据集中，细致的语义描述有效地捕捉了微妙的类别间视觉差异，这是准确分类所必需的。第三，与其他 LLM 增强方法（包括 HPT、

Table 3: 在跨域迁移设置下，我们的方法与其他现有方法的比较。

Method	Source	Target				
	ImageNet	-V2	-S	-A	-R	Avg.
CoOp (IJCV2022)	71.51	64.20	47.99	49.71	75.21	59.28
CoCoOp (CVPR2022)	71.02	64.07	48.75	50.63	76.18	59.91
MaPLe (CVPR2023)	70.72	64.07	49.15	50.90	76.98	60.27
PromptSRC (ICCV2023)	71.27	64.35	49.55	50.90	77.80	60.65
MMA (CVPR2024)	71.00	64.33	49.13	51.12	77.32	60.48
KIM (TIP2024)	73.52	64.76	48.84	49.80	76.72	60.03
PromptKD (CVPR2024)	72.42	69.77	58.72	70.36	87.01	71.47
MMRL (CVPR2025)	72.03	64.47	49.17	51.20	77.53	60.59
Using LLM:						
HPT(AAAI2024)	71.72	65.25	49.36	50.85	77.38	60.71
CoPrompt(ICLR2024)	70.80	64.25	49.43	50.50	77.51	60.42
ArGue(CVPR2024)	71.84	65.02	49.25	51.47	76.96	60.68
CoCoLe(ECCV2024)	73.88	65.86	50.89	51.75	78.89	61.85
ProText(AAAI2025)	70.22	63.54	49.45	51.47	77.35	60.45
PromptKD w /SemPT (Ours)	72.56	70.46	59.48	70.83	87.42	72.05
MMRL w /SemPT (Ours)	72.21	65.14	49.97	51.87	77.94	61.23

CoPrompt 和 CoCoLe）相比，我们的方法在评估的基准上表现出更优越的经验性能。在异构数据集上观察到的一致改进证实了我们的语义提示调整范式的广泛性和稳健性，能够有效减少基础和新识别任务之间的领域差距。

4.4.2 跨数据集迁移结果

表 2 展示了我们 SemPT 方法在跨数据集迁移范式下的综合评估，其中模型仅在 ImageNet 上进行训练，随后在 11 个异构目标数据集上进行评估，而无需任何特定领域的适应。我们的方法在所评估的基准上展示了令人信服的性能提升，其中 PromptKD 结合 SemPT 设立了新的平均准确性状态-of-the-art，达到 72.11 %，这比原始 PromptKD 框架有实质性提高，提升了 0.78 %。鉴于跨数据集泛化的内在复杂性，尤其是在明显的领域差距和分布差异对模型可迁移性构成巨大挑战的情况下，这一进步尤为显著。

实证研究揭示了我们语义提示调优范式在跨域泛化能力方面的几个关键方面。值得注意的是，我们的方法在具有挑战性的细粒度识别任务中展现出特别显著的改进，例如在 StanfordCars（75.24 % vs 73.93 %）和 DTD（57.03 % vs 55.08 %）上，在这些任务中，语义丰富描述的引入对于区分跨不同行域的视觉相似类别起到了关键作用。此外，在包括自然场景识别（SUN397）、航空影像分类（EuroSAT）和专门的航空数据集（FGVCAircraft）在内的各类目标领域中观察到的一致性能提升，证实了我们方法的稳健性和多样性。从属性级描述中提炼出的语义知识似乎提供了可以有效弥合源和目标分布之间语义差距的领域无关表示，从而增强了跨域辨别能力，同时缓解了领域偏移的不利影响。

4.4.3 跨领域转移结果

表格 3 评估了我们的 SemPT 方法在跨领域迁移设定下的鲁棒性，其中在 ImageNet 上训练的模型被直接转移到四个具有挑战性的 ImageNet 变体：ImageNet-V2、ImageNet-Sketch（-S）、ImageNet-Adversarial（-A）和 ImageNet-Rendition（-R）。这些变体展示了不同的领域转移，包括分布变化、风格变化、对抗性扰动和艺术再现，从而构成了评估跨领域泛化能力的严格测试平台。我们的方法在所有目标领域中展现了稳定的改进，其中 PromptKD 结合 SemPT 达到了最高平均表现 72.05 %，较原始的 PromptKD 基线有了 0.58 % 的显著提升。

实验结果强调了语义提示调试在缓解跨不同视觉分布的领域转移挑战中的有效性。尤其值得注意的是在 ImageNet-Sketch（59.48 % 对比 58.72 % ）和 ImageNet-Adversarial（70.83 % 对比 70.36 % ）上的表现，其中丰富的语义描述的融入在保持辨别能力方面发挥了关键作用，尽管存在显著的风格和对抗性扰动。在所有 ImageNet 变体中观察到的一致改进，每种呈现独特的分布挑战，证明了 LLM 生成的语义知识提供了稳健的领域不变表示，有效地弥合了自然和合成视觉领域之间的鸿沟。此外，MMRL w /SemPT（61.23 % 对比 60.59 % ）实现的相对较小但一致的增益证实了我们的语义增强策略在不同基准架构上的普遍适用性。这些发现证实了这样一个假设：语义增强提示促进了更稳健的特征表示，这种表示在跨域转移中保持辨别能力，从而增强了模型的跨域泛化能力。

4.4.4 小样本学习结果

图 4 展示了在 11 个独立数据集上的不同 shot 数（1, 2, 4, 8, 16）下的 few-shot 学习性能及其平均结果。我们的 SemPT 方法在不同的 shot 设置中始终优于基线方法。值得注意的是，我们的方法在 16-shot 设定中实现了 84.84 % 的最高平均准确率，超过了最强基线 MMRL 0.5 %。性能提升特别明显的是在具有挑战性的数据集，例如 FGVCAircraft 和 DTD，在这些数据集中，LLM 生成的丰富语义描述为细粒度识别任务提供了关键的辨别信息。在所有 shot 数上的持续改进表明，我们的语义增强策略有效利用了有限的训练数据，使其在标注数据稀少的实际 few-shot 场景中特别有价值。

4.5 消融研究

实验在基础到新颖的泛化设置下进行，以评估每个组件的贡献。我们的 SemPT 方法作为一个插件组件，所有实验都是通过将 SemPT 与 PromptKD 结合作为基础框架来进行的。

文本嵌入选择的效果。表格 4 (a) 比较了针对不同类别类型的三种文本嵌入策略。我们采用的混合策略（对于已知类别使用标签文本嵌入 T^{lbl}，对于未知类别使用属性增强文本嵌入 T^{enh}）取得了最佳表现：基础部分为 87.08 %，新颖部分为 81.21 %，调和平均为 84.04 %。对于所有类别使用属性增强的文本嵌入可提高新颖性能，但会略微影响基础准确度（86.78 % 对比 87.08 %）。结果表明，标签文本嵌入对已知类别更有效，因为它们在训练期间已经通过直接视觉监督进行针对性优化对齐。相比之下，未知类别显著受益于增强的嵌入，这些嵌入编码了可转移的语义属性，可以在缺乏视觉训练数据的情况下促进跨类别知识转移。这些发现支持我们自适应嵌入选择策略，表明根据专门的训练目标，不同的嵌入类型在不同的推理场景中是最佳的。

知识构建策略的效果。表格 4 (b) 比较了三种知识构建方法。两步提示策略取得了最佳结果（84.04 % HM），优于一步提示策略（83.75 % HM）和仅使用类别名称（83.83 % HM）。两步方法的优越表现展示了结构化知识获取的有效性。第一步使大型语言模型能够建立对每个类别的一般信息的全面理解，而第二步促进了更有针对性和区分性的描述生成。这一渐进的完善过程确保了生成的描述在语义上丰富且类别特定，减轻了单步生成可能产生的潜在歧义。这些结果强调了结构化知识获取的重要性，其中两步完善过程被证明比单步生成和单纯使用类别名称更为有效。

α 的影响。表格 4 (c) 显示了融合权重 α 的影响。最佳性能出现在 $\alpha = 0.2$ (84.04 % HM)，随着 α 增加到 0.4 (83.63

Table 4: 各种组件的消融研究

(a) Text Embedding Choice

Method	Base	Novel	HM
T^{lbl} for all categories	87.08	80.56	83.69
T^{enh} for all categories	86.78	81.21	83.91
T^{lbl} for seen categories T^{enh} for unseen categories	87.08	81.21	84.04

(b) Knowledge Construction Strategy

Strategy	Base	Novel	HM
Using category names only	86.92	80.95	83.83
One-step prompting strategy	86.86	80.85	83.75
Two-step prompting strategy	87.08	81.21	84.04

(c) Hyperparameter α

α	Base	Novel	HM
0.2	87.08	81.21	84.04
0.4	86.53	80.92	83.63
0.6	86.75	80.75	83.64
0.8	86.72	80.96	83.74

(d) Hyperparameter β

β	Base	Novel	HM
0.2	86.92	80.95	83.83
0.4	86.86	80.85	83.75
0.6	87.08	81.21	84.04
0.8	86.81	80.89	83.75

(e) Hyperparameter K

K	Base	Novel	HM
1	87.01	80.74	83.76
2	87.08	81.21	84.04
3	86.59	80.93	83.66

% HM)、0.6 (83.64 % HM) 和 0.8 (83.74 % HM)，性能下降。低的最佳 α 值揭示了在基础到新颖迁移学习中文件融合的一个关键洞见。通过广泛的预训练和微调过程学习到的原始文本嵌入，包含稳定可靠的语义表示，应在最终特征空间中保持主导地位。LLM 生成的文本嵌入虽然能提供有价值的补充信息，但应作为细微的增强，而非替代。当 α 变得过大时，增强的文本嵌入可能会引入噪声或与已建立良好的原始表示不一致的语义信号，导致性能下降。

β 的效果。表 4 (d) 考察了损失平衡权重 β 。最佳的 β = 0.6 达到了 84.04 % HM，优于较低的取值（0.2: 83.83 % HM，0.4: 83.75 % HM）和较高的取值（0.8: 83.75 % HM）。β = 0.6 优于较低值的事实表明，模型需要更强的监督信号来学习如何有效地利用 LLM 增强的表示。在训练过程中，增强损失作为一个关键的指导机制，教会模型如何正确利用由属性级描述提供的丰富语义空间。然而，当 β 显得过高（0.8）时，模型可能会过度强调增强的嵌入，从而牺牲对已见类别保持良好表现，表明在传统和增强学习目标之间需要平衡的监督。

K 的影响。表 4 (e) 报告了改变前 K 选择策略的结果。设置 K = 2 实现了最佳性能（84.04 % HM），优于 K = 1（83.76 % HM）和 K = 3（83.66 % HM）。该结果突出了在选择属性级描述时语义多样性与噪声敏感性之间的权衡。当 K = 1 时，模型仅依赖于最相似的描述，这可能导致语义覆盖不足，限制其捕捉互补属性的能力。相反，将 K 增加到 3 会加入不太相关的描述，可能引入语义噪声，从而削弱与视觉特征的一致性。最佳的 K = 2 选择通过整合多样但相关的语义提供了平衡的表示，增强了类别级理解，同时保持了区分性的关注。

5 结论

在本文中，我们介绍了语义提示微调（SemPT）框架，该框架系统地结合了来自大型语言模型的属性级知识，以构建语义对齐和可转移的文本嵌入进行迁移学习。首先，SemPT 采用两步提示策略，引导 LLM 提取共享视觉属性并生成属性级描述，捕捉标签以外的可转移语义线索，同时确保结构的一致性。其次，视觉引导的加权被应用于这些属性级描述的嵌入，减少来自不相关属性的噪声，并将它们与标签文本嵌入整合形成属性增强的文本嵌入。第三，图像嵌入与标签和属性增强的文本嵌入共同对齐，平衡已见类别的辨别能力和未见类别的可转移性。最后，在推断过程中，模型动态选择已见类别的标准标签嵌入和未见类别的属性增强嵌入之间，从而实现基于类别曝光的有效适应。在 15 个基准数据集上进行的大量实验证实了 SemPT 的有效性和多功能性，在不同的评估协议下实现了一致的最先进结果，并展示了与各种视觉语言主干的强兼容性。

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and G. Krueger. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[2] H. Rasheed, M. U. Khattak, M. Maaz, S. Khan, and F. S. Khan. Fine-tuned CLIP models are efficient video learners. In *CVPR*, pages 6545–6554, 2023.

[3] Y. Wei, H. Hu, Z. Xie, Z. Liu, Z. Zhang, Y. Cao, J. Bao, D. Chen, and B. Guo. Improving CLIP fine-tuning performance. In *ICCV*, pages 5439–5449, 2023.

[4] Z. Duan, H. Cheng, D. Xu, X. Wu, X. Zhang, X. Ye, and Z. Xie. Cityllava: Efficient fine-tuning for VLMs in city scenario. In *CVPR*, pages 7180–7189, 2024.

[5] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022.

[6] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022.

[7] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan. MaPLe: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023.

[8] M. U. Khattak, S. T. Wasim, M. Naseer, S. Khan, M. H. Yang, and F. S. Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*, pages 15190–15200, 2023.

[9] C. Zhao, Y. Wang, X. Jiang, Y. Shen, K. Song, D. Li, and D. Miao. Learning domain invariant prompt for vision-language models. *IEEE Trans. Image Process.*, 33:1348–1360, 2024.

[10] L. Yang, R.-Y. Zhang, Y. Wang, and X. Xie. MMA: Multi-modal adapter for vision-language models. In *CVPR*, pages 23826–23837, 2024.

[11] S. Xuan, M. Yang, and S. Zhang. Adapting vision-language models via learning to inject knowledge. *IEEE Trans. Image Process.*, 33:5798–5809, 2024.

[12] Z. Li, X. Li, X. Fu, X. Zhang, W. Wang, S. Chen, and J. Yang. PromptKD: Unsupervised prompt distillation for vision-language models. In *CVPR*, pages 26617–26626, 2024.

[13] Y. Wang, X. Jiang, D. Cheng, D. Li, and C. Zhao. Learning hierarchical prompt with structured linguistic knowledge for vision-language models. In *AAAI*, pages 5749–5757, 2024.

[14] S. Roy and A. Etemad. Consistency-guided prompt learning for vision-language models. In *ICLR*, 2024.

[15] X. Tian, S. Zou, Z. Yang, and J. Zhang. ARGUE: Attribute-guided prompt tuning for vision-language models. In *CVPR*, pages 28578–28587, 2024.

[16] Y. Zhang, K. Yu, S. Wu, and Z. He. Conceptual codebook learning for vision-language models. In *ECCV*, pages 235–251, 2024.

[17] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021.

[18] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022.

[19] J. Li, D. Li, C. Xiong, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[20] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, et al. Flamingo: A visual language model for few-shot learning. In *NeurIPS*, pages 23716–23736, 2022.

[21] J. Lin, H. Yin, W. Ping, P. Molchanov, M. Shoeybi, and S. Han. Vila: On pre-training for visual language models. In *CVPR*, pages 26689–26699, 2024.

[22] K. Su, X. Zhang, S. Zhang, J. Zhu, and B. Zhang. To boost zero-shot generalization for embodied reasoning with vision-language pre-training. *IEEE Trans. Image Process.*, 33:5370–5381, 2024.

[23] B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, pages 3045–3059, 2021.

[24] Y. Guo and X. Gu. MMRL: Multi-modal representation learning for vision-language models. *arXiv:2503.08497*, 2025.

[25] Z. Lu, J. Bai, X. Li, Z. Xiao, and X. Wang. Task-to-instance prompt learning for vision-language models at test time. *IEEE Trans. Image Process.*, 34:1908–1920, 2025.

[26] C. Cui, Z. Liu, S. Gong, L. Zhu, C. Zhang, and H. Liu. When adversarial training meets prompt tuning: Adversarial dual prompt tuning for unsupervised domain adaptation. *IEEE Trans. Image Process.*, 34:1427–1440, 2025.

[27] OpenAI. ChatGPT. https://www.openai.com/chatgpt, 2023. [Online; accessed 2023].

[28] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, and A. Rodriguez. LLaMA: Open and efficient foundation language models. *arXiv:2302.13971*, 2023.

[29] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, and D. Silver. Gemini: A family of highly capable multimodal models. *arXiv:2312.11805*, 2023.

[30] Anthropic. Claude. https://www.anthropic.com/product, 2024. [Online; accessed 2024].

[31] Y. Pu, X. Wu, L. Yang, and S. Wang. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Trans. Image Process.*, 33:4923–4936, 2024.

[32] M. U. Khattak, M. F. Naeem, M. Naseer, L. Van Gool, and F. Tombari. Learning to prompt with text only supervision for vision-language models. In *AAAI*, pages 4230–4238, 2025.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[34] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, pages 178–178, 2004.

[35] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, pages 554–561, 2013.

[36] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505, 2012.

[37] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *arXiv:1306.5151*, 2013.

[38] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729, 2008.

[39] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101: Mining discriminative components with random forests. In *ECCV*, pages 446–461, 2014.

[40] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010.

[41] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014.

[42] P. Helber, B. Bischke, A. Dengel, and D. Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 12(7):2217–2226, 2019.

[43] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012.

[44] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, pages 5389–5400, 2019.

[45] H. Wang, S. Ge, Z. Lipton, and E. P. Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.

[46] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021.

[47] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, and D. Song. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8340–8349, 2021.

[48] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019.