关注模态间隔：基于模态间隔保持和补偿的 CLIP 持续学习方法

黄林澜 [1], 曹续生 [1], 卢浩日 [1], 蒙弈帆 [1], 杨飞 [2,1], 刘夏雷 [2,1*]
[1]VCIP, CS, 南开大学　　[2]NKIARI, 深圳福田
{huanglinlan, caoxusheng, luhaori}@mail.nankai.edu.cn, {feiyang, xialei}@nankai.edu.cn

Abstract

持续学习旨在使模型能够从不断到来的数据中按顺序进行学习，同时保持在已学任务上的性能。随着对比语言-图像预训练模型（*CLIP*）在多种下游任务中展现出强大的能力，越来越多的研究开始关注如何在持续学习场景中利用 *CLIP*。然而，大多数现有方法忽视了 *CLIP* 中固有的模态间隔，而这一间隔是其泛化能力和适应性的重要因素。本文分析了视觉-语言预训练模型在微调过程中的模态间隔变化。我们的观察表明，模态间隔能够有效反映预训练知识的保留程度。基于这一发现，我们提出了一种简单而有效的方法——***MG-CLIP***，用于提升 *CLIP* 在类别增量学习中的表现。我们的方法通过保持模态间隔来缓解遗忘问题，并通过补偿模态间隔来提升对新数据的适应能力，从而为持续学习提供了一种基于模态间隔的新视角。在多个基准数据集上的广泛实验表明，我们的方法在无需额外重放数据的情况下，优于现有方法。代码已开源：*https://github.com/linlany/MindtheGap*.

1. 引言

持续学习的目标是使模型能够持续获取新知识，并适应不断变化的现实世界 [23, 34]。传统的持续学习方法通常从头开始训练模型，旨在减少对旧任务知识的灾难性遗忘，同时保持对新数据的适应能力 [19]，这通常被称为稳定性-可塑性权衡。然而，随着大规模预训练模型的快速发展，这些模型为持续学习提供了更强的稳定性与泛化能力 [16, 25, 39]。在预训练模型中，视觉-语言预训练模型（如 CLIP）在下游任务中表现出色，

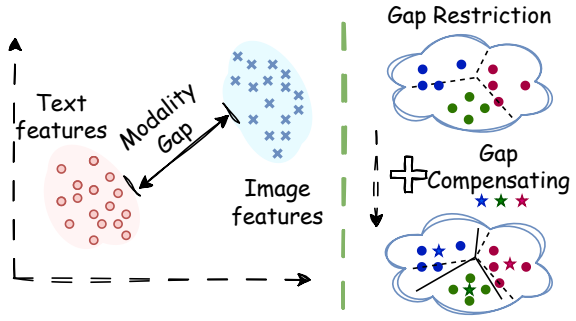


图 1. 左图：在持续学习中，我们通过保持模态间隔来保留 CLIP 的原始知识并减轻遗忘。右图：由于存在模态间隔，文本分类器的性能受到限制。我们通过引入模态内分类器进行补偿，以增强其适应性并优化决策边界。

展示了令人印象深刻的零样本能力 [17, 18, 28]。因此，基于 CLIP 的持续学习（CL）逐渐成为一个极具前景的新方向，正受到越来越多研究者的关注 [12, 13, 33, 43]。

现有基于 CLIP 的持续学习方法大致可分为两类：一类通过微调主干网络来修改特征表示 [22, 43]，另一类则冻结主干网络，并引入可学习模块以实现持续学习 [12, 13, 52]。这类方法通常将 CLIP 视为一个特征提取器，近似看作一种使用文本信息增强的视觉模型 [22]。在此基础上，它们或关注于更优的特征融合以提升性能 [13, 52]，或利用文本信息引导视觉特征的适应 [12]。然而，这些方法常常忽略了 CLIP 所特有的跨模态属性——模态间隔，这是其区别于单模态系统的重要特征。模态间隔在已有研究中已有观察 [20]。如图1（左）所示，模态间隔表现为模态内的特征距离较小，而模态间的特征距离较大。两种模态的特征分布分别位于两个明显分离的锥体中，彼此之间存在固有间隔。要充分利用 CLIP 的跨模态特性，关键在于探索保

*通信作者。

持其固有模态间隔的同时，解决其在持续学习中的局限性。

本文从模态间隔的视角出发，研究基于 CLIP 的类别增量学习。在现有关于多模态模型训练与下游任务适配的研究中，模态间隔常被视为导致性能次优的来源。研究者致力于在预训练多模态模型中缩小模态间隔 [6, 11, 26]。然而，在 CLIP 的持续学习背景下，我们的目标是在学习新数据的同时保持模型的强泛化能力。因此，如何在持续学习中处理模态间隔仍是一个开放问题。我们的工作建立在这样一个假设之上：模态间隔反映了预训练模型中固有的知识。

下游数据集通常远小于预训练数据集，因而不足以对模型进行充分训练。因此，若依据下游数据来修改预训练模型的基本属性（如模态间隔），可能会破坏原有的预训练知识，从而削弱持续学习的效果。在当前任务的训练过程中，交叉熵优化目标倾向于进一步拉大这一模态间隔；该现象在数据逐步增量的过程中将愈发显著。此外，若直接对两种模态进行对齐，也将显著改变预训练知识。因此，我们保持模态间隔的相对稳定性，以维护模型的整体稳定性。我们分析了训练过程中模态间隔的变化，并提出了一种模态间隔感知的调整策略。通过跟踪模态间隔的变化，我们在训练中进行调节，以维持稳定的模态间隔，从而实现预训练知识的保留，并减轻遗忘问题。

此外，我们还分析了保持模态间隔所带来的影响。如图1（右）所示，当使用文本特征作为分类器时，模态间隔可能限制模型在持续学习中对新数据的学习能力，降低模型的适应性与可塑性。为补偿模态间隔，我们提出在视觉空间中构建一个分类器，在该空间中不受模态间隔的限制。通过将该视觉分类器的输出与文本分类器的输出进行融合，我们实现了对模态间隔的补偿，提升了 CLIP 在持续学习中的学习能力。

本文的主要贡献如下：

- 我们提出了一种保持模态间隔的方法，用于在持续学习中保留预训练知识并缓解遗忘。
- 我们通过引入一种互补机制来补偿模态间隔的局限性，从而增强模型的可塑性。
- 我们的方法在多个数据集上无需重放或复杂结构即可达到当前最优性能。

2. 相关工作

2.1. 类别增量学习

类别增量学习方法通常被归为三类 [3]。正则化方法通过限制模型的变化来实现持续学习。一些方法使用蒸馏来减少模型特征的偏移，或通过惩罚模型参数的变动来限制遗忘 [2, 14, 19, 47, 50]。动态网络方法允许模型结构在引入新任务时发生演化 [5, 35, 40, 41]。重放方法保留原始样本或相关信息。在学习新任务时，这些方法会重放原始样本 [21, 29, 32]，或根据保留的信息恢复旧样本 [12, 36, 44, 46]，后者通常采用如内存压缩或特征重放等技术。动态网络与特征重放方法往往会增加参数数量和内存存储的开销。随着预训练模型的广泛应用，面向视觉预训练模型的增量学习方法也不断出现。参数高效的微调方法在任务逐步增加的过程中，仅扩展少量参数 [31, 37, 38]。Aper [51] 通过仅在第一个任务上训练模型以保持其稳定性。SLCA [46] 则采用小学习率微调主干网络以实现持续学习。这些研究表明，保持视觉预训练模型的稳定性有助于提升其泛化能力，并增强其在下游任务中的类别增量学习表现。

2.2. 基于 CLIP 的类别增量学习

CLIP 在使用下游数据进行类别增量学习任务中表现出色 [33]。基于 CLIP 的持续学习正受到越来越多的关注。一些方法在原始 CLIP 特征上添加可学习模块，以更好地适应新任务。例如，PROOF [52] 和 CLAP [13] 在 CLIP 的输出特征上添加可学习模块，以促进跨模态交互。RAPF [12] 则在 CLIP 的视觉编码器后引入线性层用于下游任务的适配，并利用文本模态信息引导特征重放。另一些方法则通过微调 CLIP 以改变其输出特征。ZSCL [49] 在训练过程中引入额外数据集进行蒸馏；MOE4CL [43] 采用专家混合结构微调 CLIP，引入选择机制以有选择地使用原始 CLIP 模型；Magmax [22] 顺序微调整个 CLIP 模型，并在推理阶段使用任务向量算法合并模型；LGVLM [48] 为每个任务训练独立的 LoRA 模块。这些方法主要集中于利用自然语言的先验知识来辅助持续学习，但在强调顺序任务性能的同时，常常忽略了对零样本能力的保留。相比之下，我们从 CLIP 的固有属性出发，在增强其持续学习能力的同时，注重保留模型原有的能力。

2.3. 模态间隔

Liang 等人 [20] 指出了在预训练视觉-语言模型中存在一种现象，称为模态间隔。在对比式视觉-语言模型中，模态间隔体现为文本特征与图像特征分布于两个狭窄且彼此分离的锥形区域内。不同模态的特征呈现出明显的分离，而同一模态内部的特征则更趋于聚集。现有研究已探讨模态间隔对不同任务的影响，如领域适应 [11]、小样本分类 [42]、模型预训练 [7] 和检索任务 [26]。在这些任务中，缩小模态间隔有助于提升模型性能。然而，在持续学习任务中，预训练知识的稳定性对长期下游任务而言更为关键。因此，我们提出将模态间隔作为衡量预训练模型特征空间变化的指标，目标是在持续任务中保持模态间隔。

3. 预备知识

类别增量学习定义。 我们考虑一种基于预训练 CLIP 模型 M 的类别增量学习设置。目标是对模型进行顺序训练，使其完成一系列分类任务。每个任务 t 包含一组类别 C_t，且不同任务之间的类别集合互不重叠，即 $\forall i \neq j, C_i \cap C_j = \emptyset$。在训练第 t 个任务时，模型无法访问与前 $t-1$ 个任务相关的信息。在完成第 t 个任务训练后，模型 M_t 需要在不使用任务标识符的情况下，能够正确分类之前所有学习过的类别 $C_1 \cup C_2 \cup \cdots \cup C_t$。

模态间隔度量。 在分类任务中，给定 N 张图像和 K 个类别名称的文本，我们将图像与文本特征之间的平均余弦相似度作为跨模态相似度的度量方式：$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{j=1}^{K}\cos(\mathbf{x}_i, \mathbf{t}_j)$，其中 $\mathbf{x}_i$ 表示第 i 张图像的特征，$\mathbf{t}_j$ 表示第 j 个文本的特征。该度量反映了图像模态与文本模态在特征空间中的整体相似程度，从而体现出模态间隔的大小。

为了更细粒度地分析图像与文本之间相似度的影响，我们定义正样本图文对的平均相似度为：

$$pos = \frac{1}{N}\sum_{i}^{N}\cos(\mathbf{x}_i, \mathbf{t}_{yi}), \tag{1}$$

其中 $\mathbf{t}_{yi}$ 表示与图像 $\mathbf{x}_i$ 对应的类别文本特征。

类似地，我们将负样本图文对的平均相似度定义为：

$$neg = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K-1}\sum_{j=1}^{K-1}\cos(\mathbf{x}_i, \mathbf{t}_j^{neg}), \tag{2}$$

其中 $\mathbf{t}_j^{neg}$ 表示与图像 $\mathbf{x}_i$ 不匹配的类别文本特征。

4. 方法

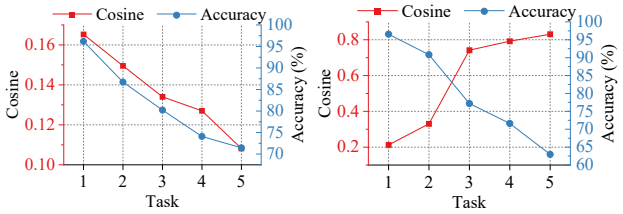
我们的方法从模态间隔的角度出发研究基于 CLIP 的持续学习，认识到在持续学习过程中保持 CLIP 的泛化能力时，模态间隔起到了关键作用。我们发现，若训练过程不加控制，可能会破坏这种固有的模态间隔，从而导致下游增量学习中的性能下降。为此，我们提出了一个两阶段策略：(1) 保持模态间隔以维持模型的稳定性。我们通过调控训练过程来确保在各个任务之间模态间隔保持稳定；(2) 补偿模态间隔以增强模型的可塑性。由于保持模态间隔可能会限制模型的适应能力，因此我们引入一个补充分类器，在不改变已保持模态间隔的前提下提升任务性能。在推理阶段，我们融合两个输出，以平衡模型的稳定性与任务可塑性。接下来，我们将讨论方法的动机（第 4.1节）以及方法的详细内容（第 4.2节和第 4.3节）。伪代码见附录第 5 节。

4.1. 模态间隔在持续学习中的影响

本节我们首先分析持续学习过程中模态间隔的演化。

交叉熵损失会扩大模态间隔。 交叉熵的优化目标与原始 CLIP 中的模态间隔不一致。在使用交叉熵损失进行优化时，训练目标是使匹配的图文对的余弦相似度趋近于 1，而非匹配对趋近于 -1。然而，这一优化方式与原始 CLIP 模型中适度的相似度分布（通常约为 0 到 0.3）不符，反映了 CLIP 原有的模态间隔。因此，该过程会导致训练后的模型模态间隔扩大。此外，下游分类任务引入了 CLIP 预训练中所没有的结构性不平衡。在预训练阶段，图文配对是对称的，而在分类任务中，多个图像对应于同一个类别文本嵌入。在一个 C 类数据集中，每个文本仅与 $\frac{1}{C}$ 的图像构成正样本，而与其余 $\frac{C-1}{C}$ 的图像构成负样本。这种不平衡使得优化过程主要受到非匹配样本排斥项的主导，进一步加剧了模态间隔的扩大。

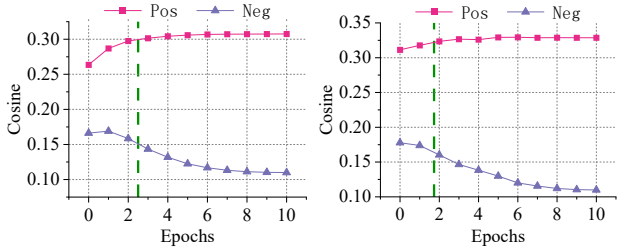
模态间隔在某种程度上反映了预训练模型的知识，模态间隔的扩大可能导致已有知识的遗忘。如图 2a 所示，实验结果验证了这一现象：随着任务的推进，图像与文本表示之间的余弦相似度稳步下降，表明模态间隔不断扩大。这一过程伴随着准确率的持续下降，突显



(a) 朴素微调会降低余弦相似度，扩大模态间隔，并导致严重的遗忘。

(b) 对齐损失增加余弦相似度并缩小模态间隔，但会破坏预训练知识，导致遗忘。

图 2. 持续学习过程中，文本与图像特征之间的平均余弦相似度及 ImageNet-R 上的准确率。



(a) ImageNet-R　　(b) ImageNet-100

图 3. 训练过程中任务内正负余弦相似度均值的变化。绿虚线前后的余弦变化模式不同。

了模态间隔扩大对 CLIP 预训练知识和类别增量学习整体性能的不利影响。

直接对齐损失减少模态间隔但破坏预训练知识。 如图 2b 所示，在下游任务中引入对齐损失 [7]，通过最小化匹配的图像与文本特征之间的欧氏距离，可以减小模态间隔。关键问题在于子空间的不匹配：下游数据集仅覆盖了原始 CLIP 特征空间的一小部分区域，这使得直接对齐存在过度专门化的风险。直接对齐这两种模态可能显著改变预训练表示，导致知识遗忘。

维持 CLIP 的终身学习能力。 先前分析揭示了一个现象：朴素微调倾向于扩大模态间隔，而直接对齐则倾向于缩小模态间隔。随着类别增量学习的推进，这一问题日益严重，逐渐侵蚀 CLIP 的预训练能力。由于 CLIP 的预训练知识在为下游任务提供稳定性方面起着关键作用，因此我们需要保持相对稳定的模态间隔。确保这种稳定性对于使 CLIP 在不损害其基本视觉-语言对应关系的前提下，顺利吸收新知识至关重要。

4.2. 自适应模态间隔保持

本节中，我们描述了持续学习中模态间隔的不对称演变，并提出了一种自适应保持策略以增强模型的稳定性。

训练过程中模态间隔变化的不对称现象。 如图 3所示，我们观察到在单任务训练过程中，文本与图像特征之间的余弦相似度变化表现出不对称性。“pos”和“neg”分别按照公式 1和公式 2计算。训练开始时，即第 0 个轮次，“pos”已大于“neg”，表明模型具备零样本能力。即使图像与其对应类别文本的相似度仍然较低，也反映了模态间隔的存在。训练早期，输出变化主要源于正样本对（图像-文本）的相似度提升，表明模型学到了新知识并对真实类别更加自信。这对应图 3中绿色虚线之前的部分。只要正样本输出大于负样本输出，即使存在损失，模型仍能做出正确预测。然而，训练后期，正样本输出趋于稳定，负样本输出开始下降，说明图像与大多数非匹配文本特征的距离增大。因此，存在一个最优训练阶段，在保持跨模态距离相对稳定的同时，保留已学知识。

自适应训练以保持模态间隔。 基于上述观察，我们通过监控负样本输出的均值来确定训练轮次。如图 4所示，使用数据集的第一任务数据估计后续任务所需训练轮数。具体而言，首先使用原始 CLIP 模型计算第一任务负样本输出的均值 neg^0（见公式 2）。随后用 LoRA 训练模型，在第 e 个轮次后，计算所有数据的负样本均值 neg^e。计算当前负样本均值与初始负样本均值的相对差异：

$$\Delta = \frac{|neg^e - neg^0|}{neg^0}. \tag{3}$$

当差异 Δ 超过预设阈值 $\boldsymbol{\alpha}$ 时，记录最后一个 Δ 仍小于 α 的训练轮次 e。最终，对于下游数据集中的所有任务，模型训练轮次设为 $\max(e, 1)$。

4.3. 模态间隔的模态内补偿

本节我们解释模态间隔如何限制模型能力，并在上一节中已说明该间隔需要保持。因此，我们构建一个模态内分类器以补偿其限制。

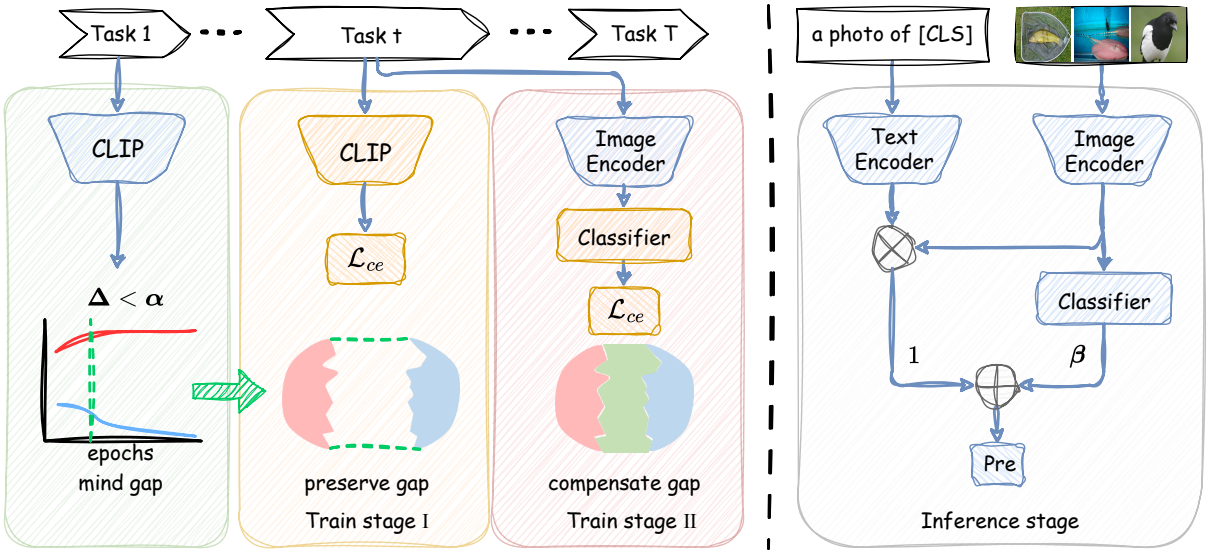


图 4. 我们的方法通过双重机制解决基于 CLIP 的持续学习问题：(1) **模态间隔保持**，当模态间隔偏差超过稳定性阈值时停止训练，防止跨模态知识扭曲；(2) **间隔补偿**，在冻结特征上训练视觉空间分类器，增强任务特定的可塑性，同时保持已保存的间隔。推理阶段，我们结合两个不同分类器子空间进行预测。

模态间隔对文本分类器能力的限制。 为了实现最小的交叉熵损失，最优分类器必须存在于图像特征空间中。具体而言，对于任何最小化分类误差的分类器，均存在一个等效形式 $\mathbf{W}_{\text{opt}}$ 完全包含于图像特征的张成空间中。这是因为任何分类器都可分解为与图像空间平行和正交的两个分量，其中正交分量对分类无贡献。该结论的详细证明见补充材料 2.1。

在此基础上，我们分析模态间隔对文本分类器的影响。文本特征矩阵 $\mathbf{T}$ 可分解为：$\mathbf{T} = \mathbf{T}_{\parallel} + \mathbf{T}_{\perp}$ 其中 $\mathbf{T}_{\parallel}$ 位于视觉特征 $\mathbf{X}$ 的张成子空间内，$\mathbf{T}_{\perp}$ 与其正交。由于文本特征通常无法完全覆盖图像特征空间，最佳的文本分类器被限制在一个低秩子空间中，导致对齐误差。从 $\mathbf{T}_{\parallel}$ 到最优图像空间分类器的距离下界由超出文本子空间的奇异值 s^2 决定：

$$\|\mathbf{T}_{\parallel} - \mathbf{W}_{\text{opt}}\|_F^2 \geq \sum_{i=r+1}^{r'} s_i^2, \tag{4}$$

其中 r' 是 $\mathbf{W}_{\text{opt}}$ 的秩，r 是 $\mathbf{T}_{\parallel}$ 的秩。该界的形式化推导见补充材料 2.2。

该结果揭示了一个内在的限制：除非文本特征空间具有足够的容量表示最优分类器，否则无法实现完美对齐。由于模态间隔的存在，文本分类器往往工作于一个低秩子空间中，限制了其分类效果。

通过模态内分类器补偿模态间隔。 为补偿模态间隔，我们在视觉空间中引入一个辅助分类器。微调后的 CLIP 模型 $f_{\text{clip}}(\cdot)$ 及旧类别的分类器权重被冻结。对于当前任务中新引入的类别，我们在余弦分类器 $\mathbf{W}_v$ 中使用类别原型初始化其分类器权重，并使用图像特征进行训练，不依赖文本。由于该分类器的梯度仍位于输入空间，即视觉空间中，因此该分类器的操作范围保持在视觉子空间内。

如图 4所示，在模型推理阶段，我们结合文本分类器与视觉分类器的预测。最终预测分数计算如下：

$$pre(\mathbf{x}) = f_{clip}(\mathbf{x}, \mathbf{t}) + \boldsymbol{\beta} \cdot softmax(\mathbf{W}_v^{\top}\mathbf{x}), \tag{5}$$

其中 $\boldsymbol{\beta}$ 是一个常数超参数。

5. 实验

5.1. 实验设置

数据集 我们在五个基准数据集上评估我们的方法：CIFAR-100 [15]、ImageNet-R [9]、ImageNet-100 [4]、ImageNet-1K [4] 和 VTAB [45]。除 VTAB 外，其余所有数据集均被平均划分为 10 个连续任务。参考先前的工作 [51]，我们从 VTAB 中提取了一个子集，包含 5 个任务，每个任务含有十个类别。更多实验细节请参考补充材料第 4 节。

Method	Exemplar	CIFAR-100		ImageNet-R		ImageNet100		ImageNet-1K		VTAB	
		Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last
PROOF [52]	DR	84.88	76.29	82.83	77.05	84.71	72.48	76.23	65.26	89.09	83.97
CLAP [13]		86.13	78.21	85.77	79.98	87.76	79.16	81.72	73.19	91.37	89.67
SLCA [46]	FR	80.53	67.58	75.92	70.37	78.63	59.92	79.10	68.27	84.25	82.54
RAPF [12]		86.19	79.04	85.58	80.28	87.51	80.23	81.73	72.58	90.88	82.31
L2P++ [38]	NR	81.90	73.08	81.67	75.98	80.51	67.22	79.30	69.60	63.23	38.37
DualPrompt [37]		81.45	72.51	82.01	75.77	80.65	67.38	79.39	69.79	61.89	37.58
CODA [31]		76.98	62.25	78.00	67.52	64.13	34.76	76.99	66.96	62.51	38.25
Continual-CLIP [33]		75.15	66.68	79.12	72.00	84.98	75.40	72.96	64.44	53.64	31.50
Aper-Adapter [51]		75.76	65.50	78.65	71.35	85.84	76.40	76.60	68.74	80.75	71.21
MOE4CL [43]		85.36	78.37	85.28	80.77	86.39	76.66	81.29	72.73	68.49	61.70
CLAP* [13]		74.19	63.45	81.22	75.80	81.07	72.00	75.85	67.36	82.11	80.11
MagMax [22]		85.63	79.00	87.13	80.85	86.33	75.92	80.74	71.31	64.63	53.90
MG-CLIP (Ours)	NR	87.00	80.57	87.58	82.67	87.31	78.38	81.88	73.68	94.67	91.53

表 1. 不同方法性能的对比。DR 表示使用真实数据进行回放，FR 表示生成旧类别特征进行回放，NR 表示不进行回放。结果主要来自参考文献 [12, 13]，并使用其公开代码复现。我们的方法是在三种不同类别顺序下的平均性能。除 VTAB 被划分为 5 个任务外，其他数据集均被划分为 10 个任务。CLAP* 表示 CLAP 论文中提供的不使用回放的版本。

对比方法 我们的实验比较了两类方法：（1）仅视觉方法，包括 L2P++ [38]、DualPrompt [37]、CODA [31]、SLCA [46] 和 Aper-Adapter [51]；（2）基于 CLIP 的方法，包括 PROOF [52]、CLAP [13]、RAPF [12]、MOE4CL [43]、MagMax [22] 以及零样本 CLIP 基线 Continual-CLIP [33]。所有方法默认采用 OpenAI 的 ViT-B/16 权重。尽管原始 MagMax 实现采用了增强的数据增强策略和优化的文本模板，我们遵循先前工作 [12, 43, 43]，使用基础图像增强和固定的 prompt 模板以保证公平对比。

评估指标 在训练第 t 个任务后，对第 1 到 t 个任务的测试数据的平均准确率记为 A_t。'Avg' 表示所有任务准确率的平均值，即 $\frac{1}{t}\sum_{i=1}^{t} A_i$。'Last' 表示完成最后一个任务 T 后的平均准确率，即 A_T。

实现细节 我们在一张 A40 GPU 上使用 PyTorch 实现了我们的方法。除非另有说明，我们使用 OpenAI 的预训练 CLIP 模型，具体为 ViT-B/16 版本。其它版本的 CLIP 结果可见补充材料。主干训练过程中，我们采用 LoRA [10] 进行模型调整，默认 rank 设置为 8。由于本文的重点不在微调过程本身，我们仅将 LoRA 应用于注意力模块的 key 和 value 部分以简化实现。不同的微调模型实现留作后续工作探索。我们使用 Adam 优化器和 cosine 学习率调度器，初始学习率为 0.001。训练轮数由方法的第一阶段决定，其中阈值 $\boldsymbol{\alpha}$ 设置为 10%。图像空间分类器训练阶段使用 3 个轮次，初始学习率为 0.0005，分类器采用 cosine classifier。在推理阶段，用于整合两个分类器结果的超参数 $\boldsymbol{\beta}$ 默认设为 4。超参数的影响分析见补充材料。

5.2. 比较结果

表 1 展示了我们方法与其他方法的比较。在大多数数据集上，我们的方法优于所有其他方法，包括依赖重放的那些。具体来说，在 CIFAR-100 上，我们的方法在 Last 准确率上至少超过所有竞争方法 1.53%。在 ImageNet-R 上，我们在 Last 准确率上至少提高了 1.82%。在 ImageNet-100 上，尽管我们的方法略逊于 RAPF 和 CLAP（这两种方法都使用了重放），但我们的方法并不依赖重放，且在 Last 准确率上仍然至少优于所有非重放方法 1.72%。此外，在 ImageNet-100 上，CLAP* 的表现显著差于 CLAP，表明 CLAP 的性能

提升主要来自数据重放。在更大规模的 ImageNet-1K 数据集上，我们的方法达到或略微优于最优替代方法的表现。

VTAB 数据集对 CLIP 构成了重大挑战，从 Continual_CLIP 的零样本准确率仅为 31.5% 可见。在这一具有挑战性的基准上，大多数方法的性能显著下降。我们的方法在 Last 准确率上超过了最优重放方法 CLAP 1.86%，且远优于其非重放版本，显示出我们方法在这一具有挑战性的跨领域场景中的明显优势。

5.3. 零样本能力的比较

持续学习的目标是使模型能够在保留已有能力的同时逐步学习新知识。与传统的预训练视觉模型相比，CLIP 模型展现出更强的零样本泛化能力。因此，基于 CLIP 的持续学习不仅应尽量减少在新下游任务上的遗忘，还应保留其原有的零样本能力。这确保了 CLIP 持续学习不只是简单的任务初始化方法，而后者易导致过拟合，是以往方法的局限性。这些方法往往沿用传统评估协议，忽视了对模型内在能力的保留。为了解决这个问题，我们提出在持续学习微调后，在独立基准上评估 CLIP 模型的零样本泛化能力。

如表 2 所示，我们在 CIFAR-100 的持续学习之后，评估了在三个标准数据集上的零样本性能。我们的方法在 Food101 [1] 和 Oxford Pets [27] 上略优于原始 CLIP，这两个数据集与 CIFAR-100 差异明显，而所有基线方法在这些数据集上的性能均有所下降。在更相似的 ImageNet-1K 数据集上，有三种方法优于原始 CLIP，其中我们的方法取得了最显著的提升（提高了 2.85%）。这些结果表明，我们的方法在有效整合新信息的同时，成功地保留了预训练知识。

值得注意的是，基于重放的方法（PROOF、RAPF 和 CLAP）在零样本任务中性能下降更为明显，相较于非重放方法。这表明重放机制可能导致对下游任务的过拟合，而非重放方法必须保留原始表示以防止遗忘。这进一步突出了重放策略对预训练模型可能产生的负面影响。

5.4. 训练成本分析

以 CIFAR-100 为例，我们通过比较本方法与其他基线方法引入的额外可学习参数，分析不同方法的参数开销，如图 5 所示。我们的方法仅引入了 0.54M 个额外的可训练参数。尽管 RAPF 方法需要更少的可学习参数，但它仍需为每个类别存储一个协方差矩阵，导致额外的存储开销与特征维度的平方成正比，即 nd^2。例如，在 100 个类别的情况下，这将带来超过 26M 参数的额外存储开销。相比之下，我们的方法完全避免了重放，保持了极小的存储消耗。

	Food101	Pets	ImageNet-1K	Avg
CLIP	85.14	87.6	64.44	79.06
PROOF	9.75	22.81	11.18	14.58
RAPF	17.18	28.56	15.2	20.31
CLAP	80.82	74.68	56.04	70.51
MOE4CL	82.85	84.06	66.02	77.64
MagMax	81.67	85.96	66.44	78.02
Ours	85.70	88.17	67.29	80.39

表 2. 在完成 CIFAR-100 上的所有类别增量学习任务后，模型在不同下游数据集上的零样本性能。CLIP 指的是未经任何下游任务特定训练的原始 CLIP 预训练模型。

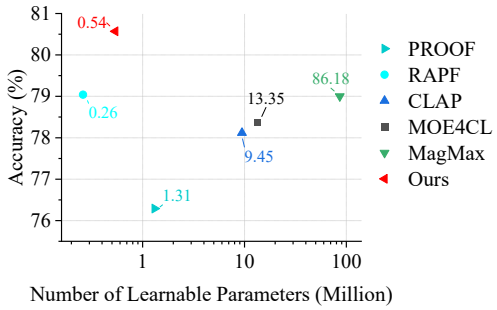


图 5. 不同方法在准确率与可学习参数数量方面的比较。

在训练过程中，我们方法唯一的额外开销出现在第一个任务的每个轮次结束后，我们对该任务的数据进行一次前向传播以评估模态间隔的变化。然而，这一步仅对第一个任务执行一次，不涉及反向传播，与整体训练成本相比几乎可以忽略不计。

5.5. 消融实验

模块消融分析 表 3 展示了我们所提出模块的消融研究。基线设置是对每个任务简单地对 CLIP 模型进行微调。按照以往工作的设定，我们对每个任务微调 10 个轮次 [22]，并采用广泛使用的交叉熵损失函数以及旧类输出屏蔽策略 [43]。MGC 表示模态间隔补偿（Modality Gap Compensation），MGP 表示模态间隔

MGP	MGC	Avg	Last
✗	✗	84.30	72.74
✗	✓	85.10	74.58
✓	✗	86.73	76.86
✓	✓	**87.31**	**78.38**

表 3. 在 ImageNet-100 上对各模块的消融实验结果。MGP 表示我们自适应的模态间隔保持，MGC 表示模态间隔补偿。

保持（Modality Gap Preservation）。

如表 3 所示，我们方法的两个组件在单独使用时均能提升性能，其中 MGP 带来的提升更为显著。当两个组件联合使用时，性能达到最佳。

当仅在基线设置中使用 MGC 时，Last 准确率提升了 1.84%，这一提升高于将 MGC 添加到 MGP 后的提升（1.52%）。在 'Avg' 准确率上也观察到了类似趋势。这表明，在基线的简单微调过程中模态间隔被进一步扩大，因此 MGC 在补偿该间隔时的作用更加突出。

图像特征空间与分类器空间。 我们研究了图像特征空间、基于文本的分类器空间以及学成的模态间隔补偿分类器空间之间的线性子空间关系。为了量化它们的差异，我们对各空间进行矩阵分解，提取其正交规范基。详细计算过程见补充材料第 3 节。$\mathbf{B}_i$：图像特征空间的正交规范基；$\mathbf{B}_t$：文本分类器空间的正交规范基；$\mathbf{B}_{vc}$：模态间隔补偿分类器空间的正交规范基；$\mathbf{B}_{t+vc}$：由文本分类器和补偿分类器共同张成空间的正交规范基。

为了衡量文本分类器空间 $\mathbf{B}_t$ 对图像特征空间 $\mathbf{B}_i$ 的覆盖程度，我们计算 $\mathbf{B}_i$ 关于 $\mathbf{B}_t$ 的正交部分的平均范数：

$$d(\mathbf{B}_i, \mathbf{B}_t) = \frac{1}{|\mathbf{B}_i|} \sum_{\mathbf{x} \in \mathbf{B}_i} \|\mathbf{x} - \mathbf{B}_t \mathbf{B}_t^\top \mathbf{x}\|. \tag{6}$$

该指标反映了有多少 $\mathbf{B}_i$ 位于 $\mathbf{B}_t$ 之外：若两个空间正交则值为 1，若一个是另一个的子空间则为 0。同理，我们计算 $d(\mathbf{B}_i, \mathbf{B}_{vc})$ 和 $d(\mathbf{B}_i, \mathbf{B}_{t+vc})$。

如表 4 所示，由于模态间隔，文本分类器空间与图像特征空间存在显著偏离。相比之下，补偿模态间隔的分类器空间与图像空间更为契合。将两个分类器空间合并后，覆盖度进一步提升，体现了它们的互补作用。

分析我们方法对模态间隔的影响。 我们分析了我们的方法对模态间隔的影响。表 5 展示了不同实验设置下最终任务的正负样本平均余弦相似度以及最终准确率（Last Acc）。'Base' 指的是简单微调的结果。可以观察到，负样本相似度均值相比原始 CLIP 明显降低，表明模态间隔被扩大，导致最终准确率下降。'Distill' 代表传统的蒸馏方法，该方法显式限制模型输出幅度，抑制模态间隔的变化，具有一定的积极效果。然而，完全限制模态间隔明显降低了模型的学习能力，导致性能不理想。可以看到其性能接近于原始 CLIP。相比之下，我们的方法保持了相对稳定的模态间隔，使得正样本相似度适度提升，负样本相似度适度下降。这种平衡保证模型能够适当地学习新知识，同时保留预训练知识，最终实现最优性能。

	CIFAR-100	ImageNet-R	ImageNet100
$d(\mathbf{B}_i, \mathbf{B}_t)$	0.7732	0.7160	0.7664
$d(\mathbf{B}_i, \mathbf{B}_{vc})$	0.6076	0.5843	0.5414
$d(\mathbf{B}_i, \mathbf{B}_{t+vc})$	**0.4917**	**0.3284**	**0.4431**

表 4. 不同数据集上图像子空间与不同分类器子空间的差异度量。

	CLIP	Base	Distill	Ours
pos	0.3157	0.3037	0.3168	0.3251
neg	0.1719	0.0579	0.1718	0.1520
Last Acc	75.40	72.74	75.85	**78.38**

表 5. 在 ImageNet-100 上不同实验设置下的余弦相似度与最终准确率（Last Acc）比较。

6. 结论

本文研究了模态间隔对视觉语言预训练模型在类别增量学习中性能的影响。我们发现，保持相对稳定的模态间隔有助于保留预训练知识并防止其退化。在模态间隔稳定的条件下，训练一个不受模态间隔限制的视觉空间分类器，可以补偿模态间隔带来的一些负面影响，进一步提升模型能力。实验结果验证了我们方法的有效性。

局限性与未来工作 我们的研究重点在于模态间隔，且模型仅通过 LoRA 简单微调，未考虑其他微调方法。目前方法未引入专门设计的损失函数或参数约束来缓解遗忘问题。未来工作将探索将本方法与其他持续学习

方法结合，并研究合适的蒸馏策略。

致谢

本工作由国家自然科学基金（项目编号 62206135, 62225604）、中国科学技术协会青年拔尖人才支持计划（2023QNRC001）、“科技勇江 2035”关键技术突破计划项目（2024Z120）、深圳市科技计划（JCYJ20240813114237048）以及中央高校基本科研业务费（南开大学，070-63233085）资助。计算资源由南开大学超级计算中心支持。

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 7

[2] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Dual-teacher class-incremental learning with data-free generative replay. In *CVPR*, pages 3543–3552, 2021. 2

[3] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE TPAMI*, 44(7):3366–3385, 2021. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 5

[5] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *CVPR*, pages 9285–9295, 2022. 2

[6] Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Investigating approaches for improving cross-modal alignment in clip. *arXiv preprint arXiv:2406.17639*, 2024. 2

[7] Abrar Fahim, Alex Murphy, and Alona Fyshe. It's not a modality gap: Characterizing and addressing the contrastive gap. *arXiv preprint arXiv:2405.18570*, 2024. 3, 4

[8] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 3

[9] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 5

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6

[11] Xuefeng Hu, Ke Zhang, Lu Xia, Albert Chen, Jiajia Luo, Yuyin Sun, Ken Wang, Nan Qiao, Xiao Zeng, Min Sun, et al. Reclip: Refine contrastive language image pre-training with source free domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2994–3003, 2024. 2, 3

[12] Linlan Huang, Xusheng Cao, Haori Lu, and Xialei Liu. Class-incremental learning with clip: Adaptive representation adjustment and parameter fusion. In *European Conference on Computer Vision*, pages 214–231. Springer, 2024. 1, 2, 6

[13] Saurav Jha, Dong Gong, and Lina Yao. Clap4clip: Continual learning with probabilistic finetuning for vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2, 6

[14] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2

[15] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[16] Kuan-Ying Lee, Yuanyi Zhong, and Yu-Xiong Wang. Do pre-trained models benefit equally in continual learning? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6485–6493, 2023. 1

[17] Yunheng Li, Yuxuan Li, Quansheng Zeng, Wenhai Wang, Qibin Hou, and Ming-Ming Cheng. Unbiased region-language alignment for open-vocabulary dense prediction. *arXiv preprint arXiv:2412.06244*, 2024. 1

[18] Yunheng Li, Zhong-Yu Li, Quan-Sheng Zeng, Qibin Hou, and Ming-Ming Cheng. Cascade-CLIP: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 28243–28258. PMLR, 2024. 1

[19] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE TPAMI*, 40(12):2935–2947, 2017. 1, 2

[20] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 1, 3

[21] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017. 2

[22] Daniel Marczak, Bartłomiej Twardowski, Tomasz Trzciński, and Sebastian Cygert. Magmax: Leveraging model merging for seamless continual learning. In *European Conference on Computer Vision*, pages 379–395. Springer, 2024. 1, 2, 6, 7

[23] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE TPAMI*, 45(5):5513–5533, 2022. 1

[24] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 1

[25] Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. An empirical investigation of the role of pre-training in lifelong learning. *arXiv preprint arXiv:2112.09153*, 2021. 1

[26] Marco Mistretta, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Andrew D Bagdanov. Cross the gap: Exposing the intra-modal misalignment in clip via modality inversion. *arXiv preprint arXiv:2502.04263*, 2025. 2, 3

[27] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 7

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[29] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 2

[30] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1

[31] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *CVPR*, pages 11909–11919, 2023. 2, 6

[32] Zhicheng Sun, Yadong Mu, and Gang Hua. Regularizing second-order influences for continual learning. In *CVPR*, pages 20166–20175, 2023. 2

[33] Vishal Thengane, Salman Khan, Munawar Hayat, and Fahad Khan. Clip model is an efficient continual learner. *arXiv preprint arXiv:2210.03114*, 2022. 1, 2, 6

[34] Gido M Van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022. 1

[35] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. In *ECCV*, pages 398–414. Springer, 2022. 2

[36] Liyuan Wang, Xingxing Zhang, Kuo Yang, Longhui Yu, Chongxuan Li, HONG Lanqing, Shifeng Zhang, Zhenguo Li, Yi Zhong, and Jun Zhu. Memory replay

with data compression for continual learning. In *International Conference on Learning Representations*, 2021. 2

[37] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *ECCV*, pages 631–648. Springer, 2022. 2, 6

[38] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, pages 139–149, 2022. 2, 6

[39] Tz-Ying Wu, Gurumurthy Swaminathan, Zhizhong Li, Avinash Ravichandran, Nuno Vasconcelos, Rahul Bhotika, and Stefano Soatto. Class-incremental learning with strong pre-trained models. In *CVPR*, pages 9601–9610, 2022. 1

[40] Tz-Ying Wu, Gurumurthy Swaminathan, Zhizhong Li, Avinash Ravichandran, Nuno Vasconcelos, Rahul Bhotika, and Stefano Soatto. Class-incremental learning with strong pre-trained models. In *CVPR*, pages 9601–9610, 2022. 2

[41] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *CVPR*, pages 3014–3023, 2021. 2

[42] Chao Yi, Lu Ren, De-Chuan Zhan, and Han-Jia Ye. Leveraging cross-modal neighbor representation for improved clip classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27402–27411, 2024. 3

[43] Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23219–23230, 2024. 1, 2, 6, 7

[44] Jiang-Tian Zhai, Xialei Liu, Andrew D Bagdanov, Ke Li, and Ming-Ming Cheng. Masked autoencoders are efficient class incremental learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19104–19113, 2023. 2

[45] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 5

[46] Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. *arXiv preprint arXiv:2303.05118*, 2023. 2, 6

[47] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. Class-incremental learning via deep model consolidation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1131–1140, 2020. 2

[48] Wentao Zhang, Yujun Huang, Weizhuo Zhang, Tong Zhang, Qicheng Lao, Yue Yu, Wei-Shi Zheng, and Ruixuan Wang. Continual learning of image classes with language guidance from a vision-language model. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2, 3

[49] Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *ICCV*, pages 19125–19136, 2023. 2

[50] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Co-transport for class-incremental learning. In *ACM MM*, pages 1645–1654, 2021. 2

[51] Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision*, pages 1–21, 2024. 2, 5, 6

[52] Da-Wei Zhou, Yuanhan Zhang, Yan Wang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Learning without forgetting for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1, 2, 6

关注模态间隔：基于模态间隔保持和补偿的 CLIP 持续学习方法

Supplementary Material

1. 模态间隔的可视化

如图 6 所示，我们从 LAION-400M [30] 中随机采样了 512 对图像-文本对，使用 CLIP 提取特征，并应用 UMAP [24] 进行降维。结果显示，同一模态的特征明显聚集，而不同模态的特征之间保持一定距离。这一现象反映了模态间隔的存在。

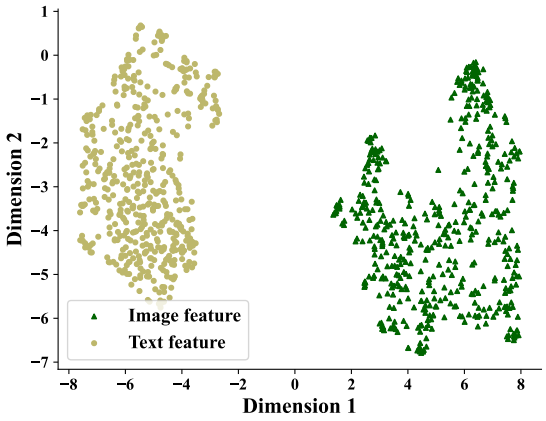


图 6. 使用 UMAP 可视化的图像及其对应文本的特征。

2. 图像空间分类器存在性及模态间隔约束证明

2.1. 图像空间内最优分类器的存在性

设 CLIP 的图像特征矩阵为 $\mathbf{X} \in \mathbb{R}^{d\times n}$，考虑一个在理想条件下达到最小交叉熵损失的分类器 $\mathbf{W}_* \in \mathbb{R}^{d\times C}$。我们证明总存在一个等价的分类器 $\mathbf{W}_{\text{opt}}$ 完全位于 $\mathbf{X}$ 的线性空间内，即 $\mathbf{W}_{\text{opt}} \in \text{span}(\mathbf{X})$。

任意分类器 $\mathbf{W}_*$ 可分解为：

$$\mathbf{W}_* = \mathbf{W}_{\parallel} + \mathbf{W}_{\perp}, \tag{7}$$

其中 $\mathbf{W}_{\parallel} \in \text{span}(\mathbf{X})$，$\mathbf{W}_{\perp} \perp \text{span}(\mathbf{X})$。输入特征 $\mathbf{x}_i \in \mathbf{X}$ 及其标签 y 对交叉熵损失的贡献为：

$$\mathcal{L}_{ce} = -\sum_{i}^{n} \log \frac{\exp(\mathbf{w}_y^\top \mathbf{x}_i)}{\sum_j^C \exp(\mathbf{w}_j^\top \mathbf{x}_i)} \tag{8}$$

显然，$\mathbf{W}_{\perp}$ 不参与分类决策，因此不影响损失函数。由此，存在一个等价分类器满足：

$$\mathbf{W}_{\text{opt}}^\top \mathbf{X} = \mathbf{W}_{\parallel}^\top \mathbf{X} = \mathbf{W}_*^\top \mathbf{X}. \tag{9}$$

这证明了存在一个最优分类器完全包含于图像特征空间。

2.2. 模态间隔对文本分类器的限制

对 $\mathbf{W}_{\text{opt}}$ 进行奇异值分解（SVD）：

$$\mathbf{W}_{\text{opt}} = \mathbf{U}\,\mathbf{V}^\top, \tag{10}$$

其中 $\mathbf{U} \in \mathbb{R}^{d\times r'}$ 是 $\text{span}(\mathbf{X})$ 的正交基。文本特征矩阵 $\mathbf{T}$ 可分解为：

$$\mathbf{T} = \mathbf{T}_{\parallel} + \mathbf{T}_{\perp}, \tag{11}$$

其中 $\mathbf{T}_{\parallel} = \mathbf{U}_r\mathbf{A}$（秩为 $r \le r'$）位于图像特征子空间内，$\mathbf{T}\perp$ 是其正交补空间。

由于 $\mathbf{T}_{\perp}$ 不影响分类，最优拟合问题可转化为：

$$\min_{\mathbf{A}} \|\mathbf{U}_r\mathbf{A} - \mathbf{W}_{\text{opt}}\|_F^2. \tag{12}$$

其最优解为：

$$\mathbf{A}^* = \mathbf{U}_r^\top \mathbf{W}_{\text{opt}}. \tag{13}$$

代回原式，可得失配误差的下界为：

$$\|\mathbf{T}_{\parallel} - \mathbf{W}_{\text{opt}}\|_F^2 \ge \|\mathbf{U}_r\mathbf{U}_r^\top\mathbf{W}_{\text{opt}} - \mathbf{W}_{\text{opt}}\|_F^2 = \sum\nolimits_{i=r+1}^{r'} s_i^2, \tag{14}$$

其中 s_i 是 $\mathbf{W}_{\text{opt}}$ 的奇异值。

该结论表明，只有当文本特征子空间的秩足够大以完整表示 $\mathbf{W}_{\text{opt}}$（即 $r = r'$）时，才能实现完美对齐。然而，由于模态间隔的存在，文本分类器的有效秩通常较低（$r < r'$），这导致分类性能存在内在限制。

3. 图像空间与分类器空间分析的实现细节

为分析这些关系，我们首先对图像特征矩阵进行奇异值分解（SVD），并提取对应的基向量，记为 $\mathbf{B}_i \in \mathbb{R}^{d\times r}$。由于图像特征数量庞大，我们仅保留能够覆盖总能量 95% 的基向量，以减少噪声同时保留关键信息。

	CIFAR100		ImageNet100	
	Avg	Last	Avg	Last
ours w/o replay	86.79	80.40	87.31	78.38
ours w/ replay	88.48	82.58	88.50	80.74

表 6. 我们方法在使用重放数据情况下的实验结果

对于文本特征分类器和视觉空间分类器，我们采用 QR 分解获得它们各自的基向量，分别记为 $\mathbf{B}_t$ 和 $\mathbf{B}_{vc}$。此外，我们还计算了由两者联合生成的空间的基向量，记为 $\mathbf{B}_{t+vc}$。

4. 更多实验

4.1. 与重放方法的兼容性

我们的方法不依赖重放样本，但仍兼容重放机制。我们测试了结合简单随机采样重放数据（总数保持为 2000）的情况。实验结果如表 6所示，表明我们的方法能从重放数据中获益，且与重放方法兼容。

4.2. CLIP ViT-L/14 主干网络的实验

我们在另一种 CLIP 模型上评估了我们方法的有效性。为此，我们将所有方法的主干网络替换为 OpenAI 更强大的 ViT-L/14 模型。实验结果如表 7 所示，表明我们的方法仍然优于其他方法。

4.3. 超参数选择

我们对所有数据集使用相同的超参数。以下是在单个数据集上进行的超参数选择实验。

LoRA 的秩 如表 8 所示，我们的方法对 LoRA 的秩不敏感。我们选择秩为 8 作为所有数据集的实验设置。

输出集成权重 $\boldsymbol{\beta}$ 如表 9 所示，随着分配给视觉空间分类器输出的权重 $\boldsymbol{\beta}$ 增加，整体性能先提升后下降。这表明，视觉空间分类器的高置信度预测能有效弥补文本分类器的不足，而低置信度预测对整体结果影响较小。然而，当 $\boldsymbol{\beta}$ 过大时，即使是视觉分类器的低置信度预测也会显著影响文本分类器的输出，导致低分的错误预测占主导。因此，适当选择权重能更好地实现两个分类器间的互补。

Method	CIFAR100		ImageNet-R	
	Avg	Last	Avg	Last
PROOF	89.87	83.59	91.25	87.33
CLAP	87.94	84.86	92.12	88.63
SLCA	90.12	84.62	89.99	86.83
RAPF	90.25	85.29	91.96	88.32
L2P++	85.68	77.86	90.49	86.73
DualPrompt	86.63	79.12	90.66	87.14
CODA	85.82	78.67	89.11	84.56
Continual-CLIP	80.48	73.46	86.99	83.05
Aper-Adapter	80.21	71.95	89.17	85.4
MOE4CL	90.98	85.83	93.27	90.42
CLAP*	74.41	71.57	91.10	87.55
MagMax	90.16	86.06	93.22	89.55
ours	91.78	87.03	93.66	91.08

表 7. 使用 ViT-L/14 主干模型的实验结果

Rank	4	8	16	32
Last	78.02	78.38	78.32	78.26

表 8. 在 ImageNet-100 上不同 LoRA 秩的实验

$\boldsymbol{\beta}$	1	2	4	6	8
Last	77.58	77.92	78.38	77.98	77.32

表 9. 在 ImageNet-100 上，不同 $\boldsymbol{\beta}$ 的实验结果

$\boldsymbol{\alpha}$	5%	10%	20%	30%
Last	77.88	78.38	77.34	76.9

表 10. 在 ImageNet-100 上，不同 $\boldsymbol{\alpha}$ 值的实验结果

参数 $\boldsymbol{\alpha}$ 的影响 如表 10 所示，当 $\boldsymbol{\alpha}$ 过小时，会对模型训练施加过多限制，导致模型无法充分学习新任务，从而限制性能表现。相反，当 $\boldsymbol{\alpha}$ 过大时，会破坏预训练知识，导致性能逐渐下降。我们选择 10% 作为 $\boldsymbol{\alpha}$ 的默认值。

	Defocus	Contrast	Frost	Gaussian
CLIP	41.95	55.07	38.23	43.20
MagMax	43.51	52.10	36.75	41.43
MOE4CL	44.24	54.82	34.75	36.09
Ours	44.65	56.02	37.71	42.47

表 11. 在 CIFAR-100 类别增量学习后，模型在 ImageNet-C 上的零样本性能表现。

	100-shot	50-shot	25-shot	5-shot
MagMax	75.82	75.02	72.53	67.66
MOE4CL	75.52	75.40	74.98	68.54
Ours	78.30	78.04	77.04	75.10

表 12. 在少样本设置下，CIFAR-100（10 任务）的最终准确率，展示了我们方法在有限数据条件下的持续提升。

4.4. 零样本能力

如表 11所示，我们在从 CIFAR-100 持续学习后，对四种 ImageNet-C [8]（严重程度为 3）的腐蚀类型进行了零样本测试。**散焦模糊（Defocus blur）**：由于 CIFAR-100 的分辨率较低，所有方法相比原始 CLIP 都有小幅提升，其中我们的方法表现最佳。**其他腐蚀类型**：其他方法的鲁棒性有所下降，而我们的方法则保持了 CLIP 的性能，几乎没有下降，并在对比度（Contrast）上略有提升。这表明我们的方法在保持 CLIP 的泛化能力方面优于其他方法。

4.5. 少样本能力

如表 12所示，我们报告了在有限数据条件下，CIFAR-100 10 任务的持续学习最终准确率。我们的方法在有限数据条件下表现出更大的优势。

4.6. 额外任务设置

我们进一步在不同任务设置下评估了我们的方法。如表 13所示，虽然我们在 CIFAR-100（5 任务）上的表现低于最佳基线，但在更具挑战性的 20 任务设置下，我们的方法取得了最高准确率。

	CIFAR-100		ImageNet-R	
	5task	20task	5task	20task
MagMax	82.07	76.84	82.75	80.18
MOE4CL	78.96	76.20	81.37	79.58
LGVLM[48]	83.84	77.26	82.46	79.32
Ours	81.47	79.31	83.13	82.12

表 13. 在 CIFAR-100 和 ImageNet-R 不同任务设置下的最终准确率。

	Ours	+ EMA	+ Epoch Est.
Avg	87.58	87.73	87.77
Last	82.67	82.92	82.82

表 14. 在 ImageNet-R（10 任务）上的辅助策略研究。

4.7. 辅助策略研究

为探索潜在的性能提升，我们测试了两种辅助策略：（1）**指数移动平均（EMA）**和（2）**每任务轮次估计**。如表 14所示，两者均带来小幅提升，但会增加计算开销。尤其是每任务轮次估计将前向计算次数翻倍。为了保持效率，我们采用不包含这些附加策略的方法。

5. 算法伪代码

训练的整体流程如伪代码所示：

Algorithm 1 训练流程

Input: $\mathbf{D} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_t\}$ ▷ 所有任务的训练数据
Input: $f^0_{\text{clip}}(\cdot)$ ▷ 原始 CLIP
require: $f^T_{clip}(\cdot)$ ▷ T 个任务微调后的 CLIP
require: $\mathbf{W}_v$ ▷ 视觉空间的余弦分类器
Initialize: $e = 0$ ▷ 微调轮次
for $t = 1$ to T **do**
    $f^{t,0}_{\text{clip}}(\cdot) = f^{t-1}_{\text{clip}}(\cdot)$ ▷ 初始化当前任务模型
    if $t = 1$ **then** ▷ 计算微调轮次
        $neg^0 = Eq.2(f^{1,0}_{\text{clip}}(\cdot), \mathbf{X}_1)$
        $f^{1,1}_{\text{clip}}(\cdot) = FINETUNE(f^{1,0}_{\text{clip}}(\cdot), \mathbf{X}_1)$
        $neg^1 = Eq.2(f^{1,1}_{\text{clip}}(\cdot), \mathbf{X_1})$
        while $Eq.3(neg^0, neg^{e+1}) < \alpha$ **do**
            $e+ = 1$
            $f^{1,e+1}_{\text{clip}}(\cdot) = FINETUNE(f^{1,e}_{\text{clip}}(\cdot), \mathbf{X}_1)$
            $neg^{e+1} = Eq.2(f^{1,e+1}_{\text{clip}}, \mathbf{X_1})$
        $e = max(1, e)$
    else
        for $i = 1$ to e **do**
            $f^{t,i}_{\text{clip}}(\cdot) = FINETUNE(f^{t,i-1}_{\text{clip}}(\cdot), \mathbf{X}_t)$
    $f^t_{\text{clip}}(\cdot) = f^{t,e}_{\text{clip}}(\cdot)$
    Initialize: $\mathbf{W}^t_v$ ▷ 通过类别原型初始化 $\mathbf{W}^t_v$
    $W^t_v = TRAIN(\mathbf{W}^t_v, f^t_{\text{clip}}(\mathbf{X_t}))$
return $f^T_{\text{clip}}(\cdot), \mathbf{W}^T_v$