无名稿件编号（由编辑插入）

CLIP引导的原型调制：小样本动作识别

Xiang Wang · 张世伟 · 岑俊 · 高昌新 · 张莹雅 ·赵德丽 · 桑农

arXiv:2303.02982v1 [cs.CV] 6 Mar 2023

收稿日期：日期 / 接受日期：日期

摘要 近年来，基于 CLIP 等大规模对比式语言-图像预训练模型进行学习，在广泛的下游任务中取得了显著成功，但在具有挑战性的小样本动作识别（FSAR）任务上的研究仍不充分。本文旨在将 CLIP 强大的多模态知识迁移过来，以缓解数据稀缺导致的原型估计不准确问题；这在低样本场景中是一个关键难题。为此，我们提出了一种名为 CLIP-FSAR 的 CLIP引导的原型调制框架，该框架由两个关键组件组成：视频-文本对比目标和原型调制。具体而言，前者通过对比视频及其对应的类别文本描述，弥合 CLIP 与小样本视频任务之间的任务差异；后者利用 CLIP 中可迁移的文本概念，通过时序 Transformer 自适应地优化视觉原型。借助这种方法，CLIP-FSAR 能够充分利用 CLIP 丰富的语义先验，获得可靠的原型并实现准确的少样本分类。在五个常用基准数据集上进行的大量实验表明，所提出的方法行之有效；在各种设置下，CLIP-FSAR 的表现均显著优于现有最先进方法。源代码和模型将公开发布于 https://github.com/alibaba-mmai-research/CLIP-FSAR。

关键词 小样本动作识别 · 多模态学习 · 多模态基础模型 · 大模型应用

Xiang Wang ·高昌新 · 桑农（通讯作者）教育部图像处理与智能控制重点实验室，人工智能与自动化学院，华中科技大学 电子邮箱：{wxiang, cgao, nsang}@hust.edu.cn 张世伟 ·张莹雅 · 赵德丽 阿里巴巴集团 电子邮箱：zhangjin.zsw@alibaba-inc.com, yingya.zyy@alibaba-inc.com 和 zhaodeli@gmail.com 岑俊 香港科技大学 电子邮箱：jcenaa@connect.ust.hk

1 引言

动作识别是视频理解领域的基础主题，近年来取得了显著进展 [33, 1, 4, 59, 60]。尽管如此，现代模型需要进行大规模数据标注，而数据收集可能耗时且繁琐。小样本动作识别是缓解数据标注问题的一个有前景的方向，旨在利用少量标注视频识别未见类别，并已受到广泛关注 [82, 3, 61]。

主流的小样本动作识别方法遵循基于度量的元学习范式 [55] 来优化模型：首先将输入视频映射到公共特征空间，然后通过预定义的度量规则计算查询样本与支持集原型之间的对齐距离，以进行分类。为了获得具有判别性的特征空间并促进学习，现有方法通常利用单模态监督预训练，例如例如，采用 ImageNet 初始化 [8]。尽管取得了显著进展，单模态初始化仍无法达到令人满意的性能，这是因为标注数据相对有限，且缺乏多模态对应信息。值得庆幸的是，对比式语言-图像预训练是一种强大的新兴范式，它通过利用互联网上大规模的图像-文本对进行对比学习，来学习高质量的可迁移表征 [43, 21]。以 CLIP [43] 为代表的典型工作成功展现了出色的迁移能力，并在各种下游任务 [81, 12, 75, 45,上取得了显著进展。

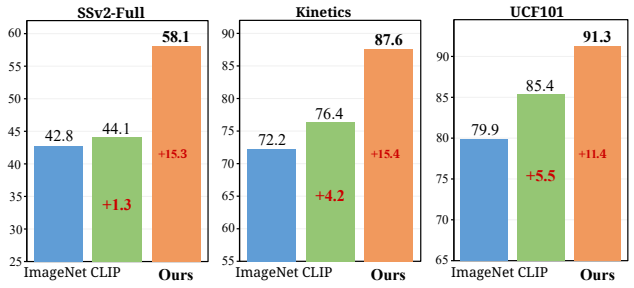


图 1：基于非典型小样本动作识别方法的性能比较，即，OTAM [3]。我们注意到，直接将原始 ImageNet 预训练骨干网络替换为 CLIP 的视觉编码器 [43]，并仅对 CLIP 模型进行微调，只能带来有限的性能提升。相比之下，我们的方法充分利用 CLIP 的多模态语义知识，从而显著提升了性能。

63]受这些成功实践的启发，我们尝试将 CLIP 的强大能力迁移到小样本动作识别任务中。一个直接的想法是用 CLIP 预训练的视觉编码器替代 ImageNet 初始化，以利用其强大的表征能力。然而，图1 中的实验结果表明，直接替换初始化方式并在下游数据集上微调，确实能够提升性能，但仍有局限。我们认为，这是因为尚未充分利用 CLIP 模型的多模态特性，因而无法在小样本场景中获得可靠的支持集原型。

受上述观察的启发，我们致力于充分利用 CLIP 模型强大的多模态知识来实现小样本动作识别，并据此提出了一种新颖的 CLIP 引导原型调制框架，称为 CLIP-FSAR。具体而言，为了使 CLIP 适应小样本视频任务，我们优化了视频—文本对比目标，以拉近视频特征与相应类别文本表征之间的距离。此外，为缓解小样本场景中视觉信息不足的困境，我们提出在 CLIP 中文本语义先验的引导下调制视觉原型。为此，我们采用时间Transformer，在支持集中自适应融合文本和视觉特征。基于这一方案，我们的 CLIP-FSAR 能够为小样本度量目标生成全面可靠的原型，从而获得稳健的分类结果。我们在五个标准基准上进行了大量实验，结果表明，我们的 CLIP-FSAR 显著优于基线方法。

总而言之，我们的主要贡献如下：(1) 我们提出了一种新颖的 CLIP-FSAR，用于小样本动作识别，充分利用了 CLIP 模型的多模态知识。据我们所知，这是首次将大规模对比式语言-图像预训练应用于小样本动作识别领域。(2) 我们设计了用于适配 CLIP 的视频-文本对比目标，以及一种用于生成可靠原型的原型调制方法。(3) 在五个具有挑战性的基准数据集上进行的大量实验，证明了我们方法的有效性，且 CLIP-FSAR 达到了最先进的性能。

2 相关工作

本研究与少样本图像分类、视觉-语言对比学习和小样本动作识别密切相关。下面简要介绍这些领域。

少样本图像分类。 少样本学习旨在利用少量样本识别新类别 [10, 16]。少样本研究大致可分为三类：数据增强、模型优化和基于度量的方法。第一类方法 [28, 17, 74, 69] 尝试利用数据生成策略提升少样本模型的泛化能力。第二类方法旨在为分类模型提供良好的初始化，使其只需少量梯度更新步数即可达到最优点。典型方法包括 MAML [11]及其其他变体 [20, 44, 47, 32, 46]。近年来， [81, 12]也尝试在测试时微调部分可学习权重，将 CLIP [43]稳健的先验迁移到少样本分类任务中。基于度量的方法通过学习映射空间并设置支持样本与查询样本的匹配规则，对查询样本进行分类；这些规则包括余弦相似度 [55, 68],、欧几里得距离 [50, 70],和可学习度量 [52, 26]。还有一些方法 [39, 40, 65, 66] 引入额外的属性信息来辅助少样本分类。在这些方法中，我们的方法属于基于度量的路线，无需在测试时进行微调即可对查询样本分类，并尝试利用 CLIP 可迁移的多模态知识，提升具有挑战性的小样本动作识别任务的性能。

视觉-语言对比学习。 探索视觉-语言预训练，是利用不同属性数据之间关联的一个极具热度和前景的方向 [31, 27, 30, 58, 57, 71, 78, 67]。近期，对比式语言-图像预训练 [43, 21, 36] 因其简洁高效而日益受到关注。以 CLIP [43] 为代表的工作通过两个独立的编码器，将图像和自然语言描述映射到公共特征空间进行对比学习，并通过在数亿个图像-文本对上进行预训练，实现了显著的“零样本”迁移能力。随后，这些预训练模型被扩展至各种下游任务，并展现出卓越的性能，包括图像分类 [81, 80],、目标检测 [48, 15], 、语义分割 [63, 45],和视频

理解 [34, 22, 35]。受这些成功的启发，我们在本文中提出首个简单而高效的框架，利用 CLIP 丰富的语义知识进行少样本动作识别。

小样本动作识别。大多数现有的小样本识别方法都采用基于度量的方式 [55]来优化模型，并设计稳健的对齐度量，以计算查询样本与支持样本之间的距离进行分类。一些方法 [84, 72,82, 83] 借鉴少样本图像分类领域的全局匹配思想 [50, 52] 进行少样本匹配，但由于度量过程中忽略了长期时间对齐信息，因此性能相对较差。为利用时间线索，以下方法 [3, 76, 42, 29, 64, 53, 61, 38, 19, 62, 77] 专注于查询视频与支持视频之间的局部帧级（或片段级）对齐。其中，OTAM [3]提出了一种动态时间规整技术的变体 [37] ，以显式利用支持视频与查询视频对中的时间顺序信息。ITANet [76]设计了一种混合空间与通道注意力机制，以学习具有代表性的特征。TRX [42] 穷举不同帧数的有序元组，以比较支持视频和查询视频。HyRSM [61] 引入自注意力来聚合视频之间的时序关系，并设计了双向平均豪斯多夫度量（Bi-MHM），以放宽匹配过程中的时间约束。Huang 等人 [19]提出了复合原型，采用多种度量函数进行少样本匹配。尽管这些方法取得了显著成功，但它们主要采用单模态预训练骨干网络，未利用多模态特征，这限制了性能的进一步提升。在我们的工作中，我们引入了强大的 CLIP 模型，并尝试利用其多模态语义知识来处理小样本动作识别任务。需要指出的是，所提出的 CLIP-FSAR 与原始线性探测CLIP之间有两个本质区别 [43]：1）原始线性探测CLIP [43]对动作识别进行线性探测评估，并需要在测试类别上微调；这与我们的基于度量的小样本设置截然不同，后者的训练类别与测试类别不相交，且不允许在测试类别上微调；2）我们的主要贡献是设计一种有效方法，利用 CLIP 的多模态知识生成可靠的原型，而原始线性探测CLIP仅对视觉骨干网络进行线性微调，并未利用多模态线索。此外，本文的关键见解是，将 CLIP 直接扩展到小样本动作识别任务显然并不充分（如图1所示）。因此，我们精心设计了 CLIP-FSAR，以更好地发挥 CLIP 强大多模态知识的优势，并希望为学术界提供一种新的可行方案，利用大型基础模型并针对下游任务进行微调。

3 方法论

我们首先简要回顾 CLIP 的背景知识 [43]，然后详细介绍我们的框架如何将 CLIP 应用于小样本动作识别任务。

3.1 CLIP预备知识

不同于使用人工标注标签进行监督训练的传统范式，[18, 8, 4], 对比语言-图像预训练（CLIP） [43]采用自然语言描述来监督表征学习，从而提升所学模型的可扩展性和迁移能力。具体而言，CLIP采用双编码器结构，由视觉编码器和文本编码器组成。视觉编码器旨在将输入图像映射到紧凑的嵌入空间，可通过 ResNet [18] 架构或ViT [9] 架构实现。文本编码器基于Transformer [54]，旨在从自然语言描述中提取高层语义特征。随后，对于匹配的图像特征和文本特征，将二者拉近；对于不匹配的，则将二者推远。总而言之，CLIP的学习目标是在公共特征空间中执行多模态对比学习。为进一步提升表征能力，CLIP在4亿个网络爬取的图文对上进行预训练。在迁移到下游任务（例如目标检测 [15]）时，为了与预训练期间的文本描述保持一致，输入文本通常采用提示模板“a photo of [CLS]”，其中 [CLS] 表示实际类别名称。

3.2 CLIP-FSAR

概述。 小样本动作识别旨在通过少量视频识别新的动作类别。在典型的小样本设置中，有两个类别互不重叠的数据集：用于训练的基础数据集 $\mathcal{D}_{train}$ 和用于测试的新类数据集 $\mathcal{D}_{test}$。为了模拟测试环境，通常会在训练过程中从 $\mathcal{D}_{train}$ 中采样大量少样本任务（也称为任务回合），以优化模型。对于标准的 N路 K样本任务，支持集 $S = \{s_1, s_2, ..., s_{N\times K}\}$ 由 N 个类别组成，每个类别包含K 个视频。该任务的目标是根据支持样本对查询视频 q 进行分类。为方便公式表述，我们以 N路 1样本（即，$K = 1$）任务来介绍我们的框架。遵循先前的工作 [3, 42, 76],，我们在输入视频上采用稀疏帧采样策略 [56] ，以降低计算负担。我们首先采用 CLIP 的视觉编码器 $\mathcal{V}$ 提取输入视频帧的特征，并使用文本编码器$\mathcal{T}$ 提取相应

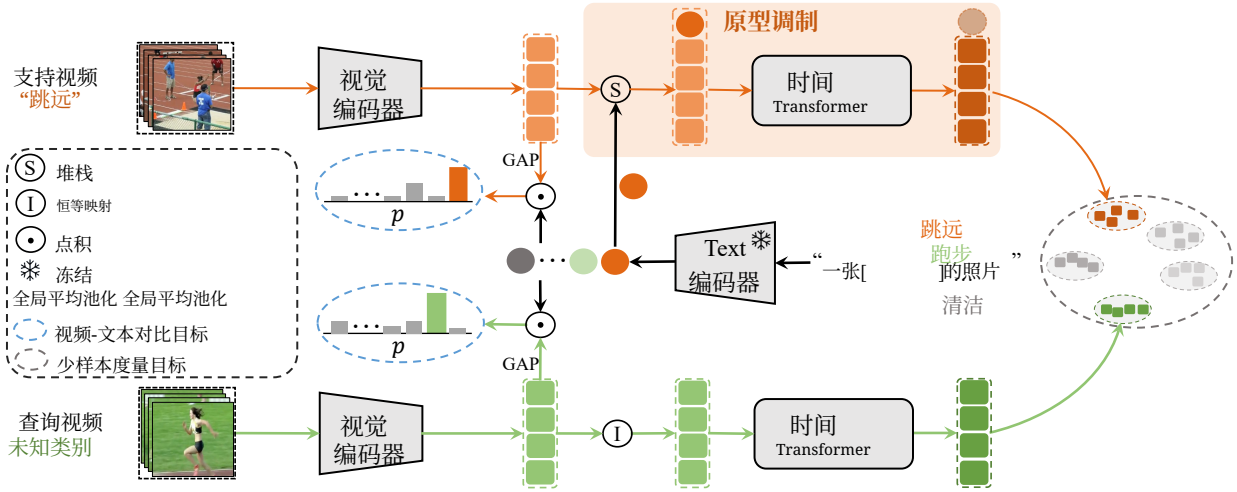


图 2： CLIP-FSAR 的架构。给定支持视频和查询视频，首先采用视觉编码器提取帧特征。随后，我们应用视频-文本对比目标，使得到的视频特征与相应的类别文本表征更加接近。之后，采用原型调制，利用文本语义特征优化视觉支持原型。最后，对得到的信息丰富的原型和查询视频施加少样本度量目标，以进行分类。

类别的自然语言描述。随后，我们对获得的帧特征和文本特征应用视频-文本对比目标，使 CLIP 适应小样本视频任务。此外，我们提出原型调制，以优化用于小样本度量目标的视觉原型。图 2展示了 CLIP-FSAR 的完整框架。

视频-文本对比目标。给定一个支持集 $S = \{s_1, s_2, ..., s_N\}$ 和一个查询视频 $q = \{q^1, q^2, ..., q^t\}$，其中 $s_i = \{s_i^1, s_i^2, ..., s_i^t\}$ 是由稀疏采样的 t 帧组成的支持视频。采用视觉编码器 $\mathcal{V}$ 对视频特征进行编码：

$$f_{s_i} = \mathcal{V}(s_i), \quad f_q = \mathcal{V}(q) \tag{1}$$

其中，$f_{s_i} \in \mathbb{R}^{t\times C}$、$f_q \in \mathbb{R}^{t\times C}$和 C 表示通道数。随后，我们使用文本编码器 $\mathcal{T}$，借助提示模板“a photo of [CLS]”提取基类的文本特征，并将得到的文本特征记为 $\{w_i\}_{i=1}^{B}$，其中 B 是基类集合 $\mathcal{D}_{train}$ 中的类别总数，$w_i \in \mathbb{R}^C$ 是一个特征向量。遵循以往的做法 [63, 48, 45], ，为保留文本编码器中原有的预训练迁移知识并减轻优化负担，我们在训练期间固定 $\mathcal{T}$，不进行更新。

为弥合 CLIP 与小样本视频任务之间的任务差异，我们模拟原始的 CLIP 训练目标：对于匹配的视频特征和文本特征，最大化它们的相似度；否则，则最小化相似度。为实现这一目标，我们首先计算视频-文本匹配概率，如下所示：

$$p_{(y=i|v)}^{video-text} = \frac{\exp(\text{sim}(\text{GAP}(f_v), w_i)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}((\text{GAP}(f_v), w_j)/\tau)} \tag{2}$$

其中，$v \in \{s_1, s_2, ..., s_N, q\}$，sim 表示余弦相似度函数，GAP 是全局平均池化的缩写，τ表示可学习的温度因子。然后，我们对预测结果与实际类别标签之间施加交叉熵损失 $\mathcal{L}_{video-text}$，以优化该目标。

原型调制。现有的小样本动作识别方法，例如 OTAM [3]，通常通过比较查询视频 q 与支持视觉原型之间的时间对齐距离来对其进行分类。在小样本动作识别任务中，视频原型是一系列帧原型，更多详情请参阅 OTAM [3]。查询视频 q 与支持视频 s_i 之间的距离可表示为：

$$d_{q,s_i} = \mathcal{M}(f_q, f_{s_i}) \tag{3}$$

其中 $\mathcal{M}$ 表示一种时间对齐度量。在 OTAM 中，$\mathcal{M}$ 是动态时间规整 [37] 的一种变体，用于将帧对齐代价作为支持集—查询集距离进行度量。小样本性能很大程度上取决于原型估计 [50] 的准确性。一方面，在低样本场景中，由于数据稀缺，视觉信息往往不足，导致原型不准确。另一方面，视觉与文本模态 [49]，具有互补性，而 CLIP 的文本描述表示包含丰富的语义先验 [25]。基于此，为了提高支持集原型的可靠性，我们提出利用信息丰富的支持文本特征来优化原型。具体而言，基于支持视觉特征 f_{s_i}，我们沿时间维度将文本特征堆叠到相应视频 s_i 中，即即，$\mathbb{R}^{t\times C} \cup \mathbb{R}^C \rightarrow \mathbb{R}^{(t+1)\times C}$，并使用时间Transformer

以自适应地融合特征。我们将得到的增强视觉特征记为 f_{s_i} （不包括文本特征输出）。由于测试时我们不知道查询视频的真实类别信息，因此只将视觉查询特征输入时间Transformer，使输出查询特征 f_q 与支持特征能够在公共特征空间中匹配。随后，我们采用时间对齐度量计算查询集与支持集之间的距离：

$$d'_{q,s_i} = \mathcal{M}(\widetilde{f}_q, \widetilde{f}_{s_i}) \tag{4}$$

其中 $\mathcal{M}$ 在我们的 CLIP- FSAR 中默认采用 OTAM度量。需要注意的是，所提出的 CLIP-FSAR 是一个即插即用的框架，在后续实验部分，我们将把 CLIP-FSAR 插入其他现有的度量或方法中 [76, 42, 61] ，并通过实验证明其可插拔性。基于这些距离，查询视频 q 在支持类别上的概率分布可表示为：

$$p^{few-shot}_{(y=i|q)} = \frac{\exp(d'_{q,s_i})}{\sum_{j=1}^{N}\exp(d'_{q,s_j})} \tag{5}$$

遵循以往工作 [3, 61, 42], ，我们可以使用交叉熵损失 $\mathcal{L}_{few-shot}$ 来优化模型参数。我们的 CLIP-FSAR 最终训练目标为：

$$\mathcal{L} = \mathcal{L}_{video-text} + \alpha\mathcal{L}_{few-shot} \tag{6}$$

其中 α 是一个平衡因子。在少样本评估中，我们可以通过公式 5获得属于支持类别的匹配概率，与以往工作 [3, 42, 61] 类似。此外，由于我们所提出框架采用双目标设计，因此还可以结合视频-文本匹配结果（公式 2）和少样本分类结果（公式5）来获得合并后的预测：

$$p^{\dagger}_{(y=i|q)} = (p^{video-text}_{(y=i|q)})^{\beta} \cdot (p^{few-shot}_{(y=i|q)})^{1-\beta} \tag{7}$$

其中， $\beta \in [0,1]$ 是一个可调超参数，我们将上述集成方式记为 CLIP-FSAR†。请注意，上述零样本与小样本结果的结合仅是一种可选方案，本文主要关注小样本性能。

4 实验

本节在五个少样本动作识别基准数据集上进行了大量实验，并将 CLIP-FSAR 与当前最先进方法进行比较，以验证其有效性。此外，我们还进行了详细的消融研究，以分析 CLIP-FSAR 的特性。

4.1 实验设置

数据集。 我们在五个常用的少样本动作数据集上验证所提出的方法，包括 SSv2-Full [13], SSv2-Small [13],Kinetics [4], UCF101 [51],和 HMDB51 [24]。参照 [3, 82, 42, 76], 在 SSv2-Full、SSv2-Small 和 Kinetics 上的设置，为了公平比较，我们随机抽取 64 个类别作为基类，24 个类别作为新类。对于 UCF101 和 HMDB51，我们遵循 [72] 的数据集划分来评估我们的 CLIP-FSAR。

评估协议。遵循先前的工作 [3, 61, 42], ，我们报告 CLIP-FSAR 在 5-way$K$-shot 设置下的性能，其中 K 的范围为 1 到 5。

实现细节。我们的实验实现基于 PyTorch [41]库开发，并使用 Adam [23] 优化器进行训练。为简单起见，除非另有说明，我们采用 OTAM [3]作为基线对齐度量。遵循先前的方法 [3, 61], 8，我们从每个视频中均匀且稀疏地采样帧，以编码视频表示。在训练过程中，我们采用随机裁剪和颜色抖动等几种标准数据增强技术。我们使用两个 CLIP 模型验证 CLIP-FSAR，即 CLIP-RN50（ResNet-50） [18]和 CLIP-ViT-B（ViT-B/16） [9]。在多样本场景下（例如，5样本），我们采用简单而有效的平均原则 [50] ，在输入原型调制模块之前生成平均支持特征。在推理时，我们报告实验中随机采样的 10,000 个少样本任务的平均准确率。

基线方法。 现有的小样本动作识别方法，如 OTAM [3], TRX [42], ITANet [76], 和 HyRSM [61] ，均采用单模态 ImageNet预训练。为验证我们方法的有效性，我们主要考虑两种采用 CLIP 初始化的典型基线方法。第一种是 OTAM，我们将其原有的 ImageNet预训练替换为 CLIP 的视觉编码器，进行端到端训练。第二种是 CLIP-Freeze，它直接利用预训练视觉编码器输出的特征，并使用 OTAM度量进行少样本匹配，而无需在基础数据集上重新训练。

4.2 与最先进方法的比较

为验证所提出框架的有效性，我们在五个标准数据集上，将 CLIP-FSAR 的表现与当前最先进的小样本动作识别方法进行了比较。结果汇总于表1 和表2。根据实验结果，我们可以得出以下观察：(a) 与 OTAM 基线相比，我们的方法能够借助 CLIP 的多模态知识显著提升性能。例如，在 SSv2-Full 的 5-way 1样本设置下，

表1： 在 SSv2-Full 和 Kinetics 上与近期少样本动作识别方法的比较。实验在 5-way K-shot 设置下进行，其中 K 从 1 变化到 5。“-”表示已发表研究中未提供该结果。“$*$”表示我们实现得到的结果。“INet-RN50”指在 ImageNet 上预训练的 ResNet-50。“†”表示公式7中的集成方法，该方法结合了少样本结果和视频-文本匹配结果。

方法	参考文献	预训练	SSv2-Full					Kinetics				
			1样本	2样本	3样本	4样本	5样本	1样本	2样本	3样本	4样本	5样本
MatchingNet [55]	NeurIPS'16	INet-RN50	-	-	-	-	-	53.3	64.3	69.2	71.8	74.6
MAML [11]	ICML'17	INet-RN50	-	-	-	-	-	54.2	65.5	70.0	72.1	75.3
Plain CMN [82]	ECCV'18	INet-RN50	-	-	-	-	-	57.3	67.5	72.5	74.7	76.0
CMN++ [82]	ECCV'18	INet-RN50	34.4	-	-	-	43.8	-	-	-	-	-
TRN++ [79]	ECCV'18	INet-RN50	38.6	-	-	-	48.9	-	-	-	-	-
TARN [2]	BMVC 2019	INet-RN50	-	-	-	-	-	64.8	-	-	-	78.5
CMN-J [83]	TPAMI'20	INet-RN50	-	-	-	-	-	60.5	70.0	75.6	77.3	78.9
ARN [72]	ECCV 2020	-	-	-	-	-	-	63.7	-	-	-	82.4
OTAM [3]	CVPR 2020	INet-RN50	42.8	49.1	51.5	52.0	52.3	72.2*	75.9	78.7	81.9	84.2*
ITANet [76]	IJCAI 2021	INet-RN50	49.2	55.5	59.1	61.0	62.3	73.6	-	-	-	84.3
TRX ($\{1\}$)	CVPR'21	INet-RN50	38.8	49.7	54.4	58.0	60.6	63.6	75.4	80.1	82.4	85.2
TRX ($\{2,3\}$) [42]	CVPR'21	INet-RN50	42.0	53.1	57.6	61.1	64.6	63.6	76.2	81.8	83.4	85.9
TA^{2}N [29]	AAAI'22	INet-RN50	47.6	-	-	-	61.0	72.8	-	-	-	85.8
MTFAN [64]	CVPR'22	INet-RN50	45.7	-	-	-	60.4	74.6	-	-	-	87.4
STRM [53]	CVPR'22	INet-RN50	43.1	53.3	59.1	61.7	68.1	62.9	76.4	81.1	83.8	86.7
HyRSM [61]	CVPR'22	INet-RN50	54.3	62.2	65.1	67.9	69.0	73.7	80.0	83.5	84.6	86.1
Nguyen 等人 [38]	ECCV 2022	INet-RN50	43.8	-	-	-	61.1	74.3	-	-	-	87.4
Huang 等人 [19]	ECCV 2022	INet-RN50	49.3	-	-	-	66.7	73.3	-	-	-	86.4
HCL [77]	ECCV 2022	INet-RN50	47.3	54.5	59.0	62.4	64.9	73.7	79.1	82.4	84.0	85.8
OTAM（基线） [3]	CVPR 2020	CLIP-RN50	44.1	52.4	55.6	58.1	59.8	76.4	84.0	86.1	87.7	88.9
CLIP-Freeze [43]	ICML'21	CLIP-RN50	28.5	32.5	34.5	37.1	37.8	68.2	78.3	82.1	84.0	85.3
CLIP-FSAR	-	CLIP-RN50	58.1	59.0	60.7	61.4	62.8	87.6	89.3	90.7	91.7	91.9
CLIP-FSAR†	-	CLIP-RN50	58.7	59.2	60.7	61.4	62.8	90.1	90.4	90.8	91.9	92.0
OTAM（基线） [3]	ICML'21	CLIP-ViT-B	50.2	57.4	63.8	65.5	68.6	88.2	92.4	93.6	94.2	94.8
CLIP-Freeze [43]	ICML'21	CLIP-ViT-B	30.0	34.4	38.2	40.6	42.4	78.9	87.0	90.2	91.4	91.9
CLIP-FSAR	-	CLIP-ViT-B	**61.9**	**64.9**	**68.1**	**70.9**	**72.1**	**89.7**	**92.9**	**94.2**	**94.8**	**95.0**
CLIP-FSAR†	-	CLIP-ViT-B	62.1	65.0	68.3	70.9	72.1	94.8	95.0	95.0	95.1	95.4

我们的方法采用 CLIP-RN50 和 CLIP-ViT-B 时，性能分别提升了 14.0% 和 11.7%。值得注意的是，基于 CLIP-ViT-B 的 CLIP-FSAR 始终优于其他最先进的方法，证明了我们方法的有效性。(b) 对比采用相同 CLIP 视觉编码器的 OTAM 和 CLIP-Freeze 的结果，可以发现，在小样本动作识别任务上重新训练 CLIP，能够使其适应下游任务并进一步提升性能。(c) 基于 CLIP-ViT-B 的 CLIP-FSAR 通常优于基于 CLIP-RN50 的方法，这表明更强的预训练模型能够带来更好的小样本泛化能力。此外，多模态预训练的 CLIP 表现优于采用 ImageNet 预训练的对应模型。(d) 与基线方法相比，CLIP-FSAR 的性能优势在样本数较少时更为显著，并随着样本数增加而逐渐缩小。我们认为，这是因为在视觉信息不足时，引入文本语义线索会更加有效。同样，与基线方法相比，基于 CLIP-RN50 的性能提升幅度大于基于 CLIP-ViT-B 的方法。在 Kinetics 数据集上，采用 CLIP-RN50 时，1样本的性能差距为 11.2%（87.6% 对 76.4%），而采用 CLIP-RN50 上为 1.5%（89.7% 对比 88.2%），而 CLIP-VIT-B 上为 1.5%。(e) 将视频-文本匹配的结果融入少样本分类中，即 CLIP-FSAR†，性能也得到了一定提升。

为进一步验证我们框架的可插拔性，我们将所提出的方法作为即插即用组件应用于现有的度量方法或方法，例如 Bi-MHM [61],ITANet [76], 和 TRX [42]。从表3中可以发现，将我们的 CLIP-FSAR 扩展到这些技术时，同样能够取得显著的性能提升，这表明我们的方法是一种通用架构。

4.3 消融研究

为研究 CLIP-FSAR 中各组件的贡献，我们开展了一系列详细的消融研究。除非另有说明，我们采用 CLIP-RN50 模型作为对比实验的默认设置。

组件分析。在表4中，我们研究了视频-文本对比目标和原型调制在所提出的 CLIP-FSAR 中的作用。通过执行视频-

表 2： 在 UCF101、SSv2-Small 和 HMDB51 数据集上与近期少样本动作识别方法的比较。实验采用 5-way K-shot 设置，其中 K 从 1 变至 5。“-”表示已发表研究中未提供该结果。“$*$”表示我们实现所得的结果。“INet-RN50”表示在 ImageNet 上预训练的 ResNet-50。“†”表示公式 7中的集成方法，该方法结合了小样本结果与视频-文本匹配结果。

方法	参考文献	预训练	UCF101			SSv2-Small			HMDB51		
			1样本	3样本	5样本	1样本	3样本	5样本	1样本	3样本	5样本
MatchingNet [55]	NeurIPS'16	INet-RN50	-	-	-	31.3	39.8	45.5	-	-	-
MAML [11]	ICML'17	INet-RN50	-	-	-	30.9	38.6	41.9	-	-	-
Plain CMN [82]	ECCV'18	INet-RN50	-	-	-	33.4	42.5	46.5	-	-	-
CMN-J [83]	TPAMI'20	INet-RN50	-	-	-	36.2	44.6	48.8	-	-	-
ARN [72]	ECCV 2020	-	66.3	-	83.1	-	-	-	45.5	-	60.6
OTAM [3]	CVPR 2020	INet-RN50	79.9	87.0	88.9	36.4	45.9	48.0	54.5	65.7	68.0
ITANet [76]	IJCAI 2021	INet-RN50	-	-	-	39.8	49.4	53.7	-	-	-
TRX [42]	CVPR'21	INet-RN50	78.2	92.4	96.1	36.0	51.9	56.7*	53.1	66.8	75.6
TA^{2}N [29]	AAAI'22	INet-RN50	81.9	-	95.1	-	-	-	59.7	-	73.9
MTFAN [64]	CVPR'22	INet-RN50	84.8	-	95.1	-	-	-	59.0	-	74.6
STRM [53]	CVPR'22	INet-RN50	80.5	92.7	96.9	37.1	49.2	55.3	52.3	67.4	77.3
HyRSM [61]	CVPR'22	INet-RN50	83.9	93.0	94.7	40.6	52.3	56.1	60.3	71.7	76.0
Nguyen等人 [38]	ECCV 2022	INet-RN50	84.9	-	95.9	-	-	-	59.6	-	76.9
Huang等人 [19]	ECCV 2022	INet-RN50	71.4	-	91.0	38.9	-	61.6	60.1	-	77.0
HCL [77]	ECCV 2022	INet-RN50	82.5	91.0	93.9	38.7	49.1	55.4	59.1	71.2	76.3
OTAM（基线） [3]	CVPR 2020	CLIP-RN50	85.4	92.5	94.5	39.9	48.2	51.9	63.0	71.2	76.6
CLIP-Freeze [43]	ICML'21	CLIP-RN50	84.6	92.9	94.5	26.8	33.7	36.3	51.4	65.2	71.0
CLIP-FSAR	-	CLIP-RN50	91.3	95.1	97.0	52.0	54.0	55.8	69.2	77.6	80.3
CLIP-FSAR†	-	CLIP-RN50	92.4	95.4	97.0	52.1	54.0	55.8	69.4	78.3	80.7
OTAM（基线） [3]	CVPR 2020	CLIP-ViT-B	95.8	98.2	98.8	43.3	53.5	57.5	72.5	81.6	83.9
CLIP-Freeze [43]	ICML'21	CLIP-ViT-B	89.7	94.3	95.7	29.5	38.0	42.5	58.2	72.7	77.0
CLIP-FSAR	-	CLIP-ViT-B	**96.6**	**98.4**	**99.0**	**54.5**	**58.6**	**61.8**	**75.8**	**84.1**	**87.7**
CLIP-FSAR†	-	CLIP-ViT-B	97.0	98.5	99.1	54.6	59.4	61.8	77.1	84.1	87.7

表 3： 泛化实验。我们将所提出的方法应用于多种现有的少样本动作识别方法中。

方法	预训练	SSv2-Small		Kinetics	
		1样本	5样本	1样本	5样本
Bi-MHM [61]	INet-RN50	38.0	48.9	72.3	84.5
Bi-MHM [61]	CLIP-RN50	40.6	52.2	77.1	88.9
我们的方法（Bi-MHM）	CLIP-RN50	**52.3**	**55.6**	**87.7**	**92.1**
ITANet [76]	INet-RN50	39.8	53.7	73.6	84.3
ITANet [76]	CLIP-RN50	40.2	56.5	78.5	89.0
我们的方法（ITANet）	CLIP-RN50	**49.6**	**58.2**	**79.3**	**89.6**
TRX [42]	INet-RN50	36.0	56.7	63.6	85.9
TRX [42]	CLIP-RN50	39.4	56.8	71.8	89.2
我们的方法（TRX）	CLIP-RN50	**51.5**	**57.1**	**86.7**	**92.2**

表 4： 我们 CLIP-少样本动作识别中各组件的消融研究。

视频-文本原型调制		SSv2-Small		Kinetics	
		1样本	5样本	1样本	5样本
		39.9	51.9	76.4	88.9
✓		41.4	55.5	80.4	91.6
	✓	50.8	55.3	86.5	91.3
✓	✓	**52.0**	**55.8**	**87.6**	**91.9**

文本对比学习，我们在 Kinetics 上分别取得了 1样本和 5样本设置下 4.0% 和 2.7% 的提升。这一持续的性能提升表明，将 CLIP 适配到小样本视频任务中至关重要。同样，在 SSv2-Small 数据集上，采用原型调制后，性能较基线分别提升了 10.9% 和 3.4%。我们观察到，性能提升在 1样本场景下尤为显著，这表明它会更有效以在视觉信息非常有限时补充文本语义信息。此外，同时训练这两个组件能够取得最佳性能，这从实证上验证了视频-文本对比目标与原型调制具有互补性。

改变时间 Transformer 层的数量。为探究所应用的时间 Transformer 层数量的影响，我们在 SSv2-Small 和 Kinetics 上开展了消融研究。结果如表 5所示。我们可以注意到，在 SSv2-Small 上，随着 Transformer 层数量的增加，性能缓慢提升，并且当数量超过某个特定值后开始出现过拟合。不同的是，单个 Transformer 层就足以获得最佳的

表 5： 在 SSv2-Small 和 Kinetics 数据集上，不同数量的时间 Transformer 层的比较结果。

设置	SSv2-Small		Kinetics	
	1样本	5样本	1样本	5样本
时间Transformer ×1	52.0	55.8	**87.6**	**91.9**
时间Transformer ×2	52.1	56.4	87.4	91.4
时间Transformer ×3	**52.4**	**56.6**	86.7	91.2
时间Transformer ×4	51.6	55.8	85.8	90.9
时间Transformer ×5	50.8	55.1	85.2	90.4

表 6： 不同原型调制方案的比较。“仅使用时间Transformer”表示使用时间Transformer对时间关系进行建模，而不引入文本特征进行原型调制。

设置	SSv2-Small		Kinetics	
	1样本	5样本	1样本	5样本
仅时间Transformer	40.5	52.3	77.9	90.0
双向长短期记忆网络	45.2	55.3	80.0	89.1
Bi-GRU	47.1	55.5	80.5	89.5
时间Transformer（我们的方法）	**52.0**	**55.8**	**87.6**	**91.9**

表 7： 共享权重与非共享权重的比较。

设置	SSv2-Small		Kinetics	
	1样本	5样本	1样本	5样本
非共享	51.1	55.0	79.5	84.7
共享（我们的方法）	**52.0**	**55.8**	**87.6**	**91.9**

Kinetics数据集性能，我们认为这是因为该数据集相对更依赖外观信息，而时序建模对于识别的重要性较低 [5]。为了平衡精度与效率，我们采用单个时间 Transformer 层作为默认设置。

原型调制方案的影响。在表6中，我们进行了消融实验，将几种常用的时间聚合算子，如 Bi-LSTM [14]，Bi-GRU [6]，和时间 Transformer [54]进行比较，以进一步分析不同原型调制方案的效果。我们发现，时间 Transformer 始终优于其他竞争方法，这得益于其灵活的信息交互能力以及 Transformer 强大的泛化能力 [7, 61]。此外，在原型调制机制中，Transformer 的输出不受输入语义特征和视觉特征相对位置的影响，这一特性使其更适用于我们的框架。为进一步分析应用于支持视频和查询视频的时间 Transformer 是否应共享权重，我们在表7中进行了实验比较。实验结果表明，共享权重可以使查询样本和支持样本处于相同的特征空间，从而取得更好的小样本动作识别结果。

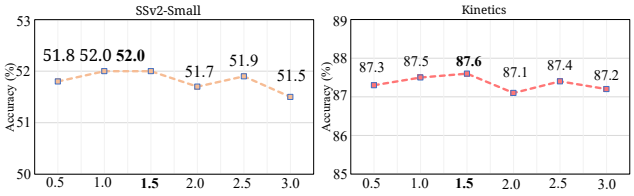


图 3： 在 SSv2-Small 和 Kinet- ics 上对 α 的敏感性分析。

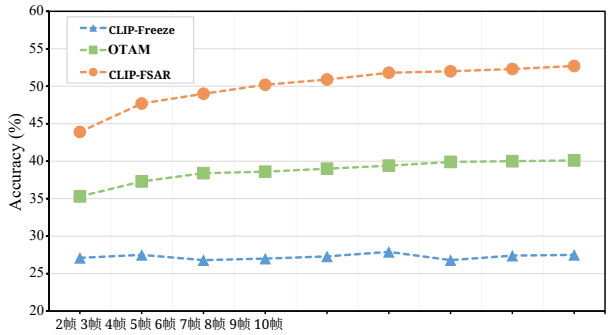


图 4： 在 SSv2-Small 上进行 5-way 1样本设置时，不同数量输入视频帧的性能对比。

超参数敏感性分析。在公式6中，我们引入了一个超参数 α 来平衡两个损失项。如图 3所示，我们报告了消融实验结果。可以观察到，该参数对性能的影响相对较小，且 CLIP-FSAR 始终保持领先性能。

改变输入视频帧的数量。为与先前方法进行公平比较，[3, 42, 76]，我们在实验中从每个视频采样 8 帧，以编码视频表征。我们进一步在图 4中分析了不同输入视频帧数量的影响。结果表明，随着输入视频帧数量增加，CLIP-FSAR 的性能缓慢提升并最终趋于饱和。值得注意的是，我们的方法始终优于基线方法，体现了 CLIP-FSAR 的可扩展性。一个有趣的现象是 CLIP-Freeze 的结果出现波动，我们认为这是因为 CLIP 最初是在图像领域进行预训练的，若不经过适配，便无法感知视频中的运动模式。

N-路分类。我们还进行了消融实验，探究不同N对小样本性能的影响。 N-路 1样本对比结果如图 5所示，其中 N 从 5 变化到 10。我们可以观察到， N 越大，分类难度越高，性能也越低。例如，CLIP-FSAR 在 SSv2-Small 上的 10 路结果相比 5 路结果下降了 14.3%。不过，

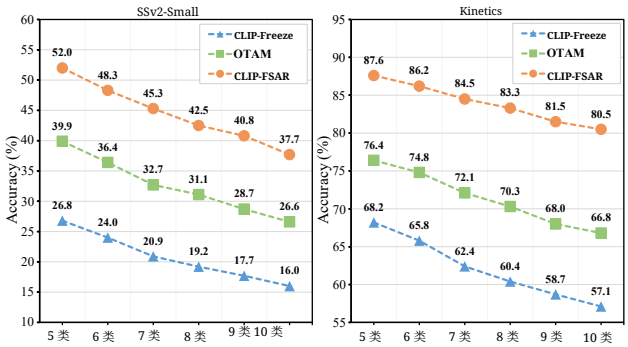


图 5： 我们的 CLIP=FSAction 与其他基线方法在 N 从 5 变化到 10 时的 N路1样本结果。

表 8： 对使用若干个在 ImageNet 上预训练的骨干网络替换我们 CLIP-FSAR 的视觉编码器进行的实证分析。

方法	预训练	SSv2-Small		Kinetics	
		1样本	5样本	1样本	5样本
OTAM	INet-RN18	35.8	47.1	68.6	78.3
CLIP-FSAR	INet-RN18	**46.0**	**52.4**	**72.9**	**84.1**
OTAM	INet-RN34	36.2	47.9	71.0	80.4
CLIP-FSAR	INet-RN34	**46.5**	**53.0**	**73.9**	**85.3**
OTAM	INet-RN50	36.4	48.0	72.3	84.2
CLIP-FSAR	INet-RN50	**46.7**	**53.3**	**75.5**	**85.9**

尽管如此，CLIP-FSAR 在各种设置下仍显著优于比较方法，这表明我们的方法有效。

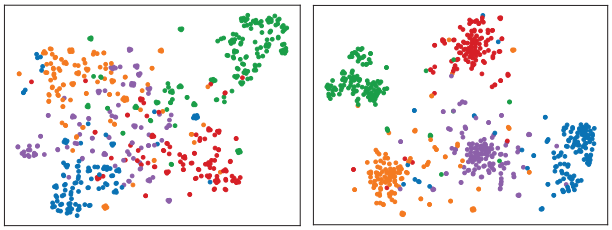
泛化能力分析。 我们的 CLIP-FSAR 基于 CLIP 预训练的视觉编码器和文本编码器进行对比实验。为了进一步探究我们的框架能否扩展到标准的单模态预训练初始化设置，我们将视觉编码器替换为在 ImageNet 上预训练的模型，并保持文本编码器不变。如表 8所示，我们分别在三种深度不同、以 ImageNet 初始化的模型 ResNet-18、ResNet-34 和 ResNet-50 上进行了实验。结果充分表明，我们的方法能够有效提升这些基线方法的性能，例如，使用 ResNet-18 时，CLIP-FSAR 将 SSv2-Small 上的 1样本性能从 35.8% 提升至 46.0%，充分证明了我们框架的泛化能力。在表 9中，我们还将 CLIP-FSAR 框架应用于自监督 DINO [73]方法。具体而言，我们将 CLIP-FSAR 中的视觉编码器替换为一个可公开获取的 DINO ResNet-50 模型，该模型在 ImageNet 上经过自监督预训练，并保持文本编码器不变。从结果可以看出，与朴素基线相比，基于 OTAM 和 TRX 时，我们的框架仍然取得了显著的性能提升，这表明我们的方法具有很强的通用性。

表 9： 将我们的 CLIP-FSAR 框架迁移至自监督 DINO 视觉编码器的泛化能力实验。

方法	预训练	SSv2-Small		Kinetics	
		1样本	5样本	1样本	5样本
OTAM	DINO-RN50	39.8	52.5	66.3	77.5
我们的方法（OTAM）	DINO-RN50	**47.3**	**53.4**	**73.3**	**84.6**
TRX	DINO-RN50	38.2	57.5	58.7	78.4
我们的方法（TRX）	DINO-RN50	**44.8**	**58.9**	**70.2**	**85.1**

表 10： 使用 CLIP 的 5-way零样本性能比较。

方法	编码器	SSv2-Small	Kinetics	HMDB51
CLIP-Freeze	RN-50	30.6	89.5	55.1
CLIP-FSAR	RN-50	**43.3**	**89.7**	**62.6**
CLIP-Freeze	ViT-B	36.0	94.2	64.4
CLIP-FSAR	ViT-B	**46.8**	**94.3**	**75.6**

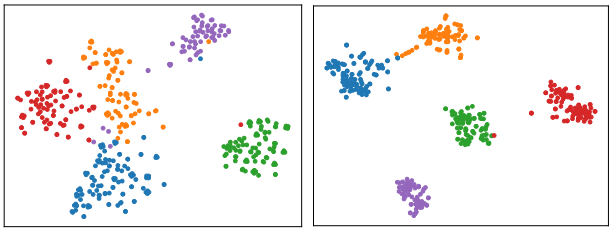


(a) 未使用原型调制 **(b)** 使用原型调制**图 6：** SSv2-Small测试集上五个动作类别的T-SNE分布可视化。建议在彩色PDF中查看。

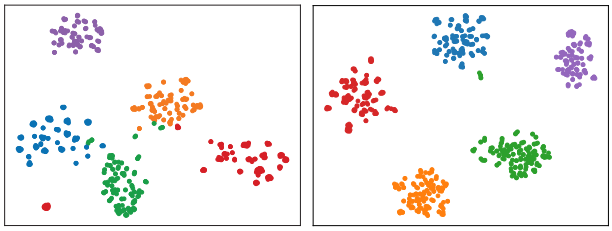
零样本性能。 尽管我们的CLIP-FSAR最初是为小样本动作识别任务设计的，但我们也可以利用视频-文本对比目标进行零样本分类。为方便起见，我们采用与少样本评估相同的数据划分，在训练集上训练，并在测试集上进行零样本分类。5-way零样本实验对比结果如表10所示。可以发现，与CLIP-Freeze相比，我们的方法性能更好。例如，在采用CLIP-ViT-B的HMDB数据集上，我们的方法比CLIP-Freeze提升了11.2%。这可以归因于我们的CLIP-FSAR对视频任务的适配。

4.4 可视化分析

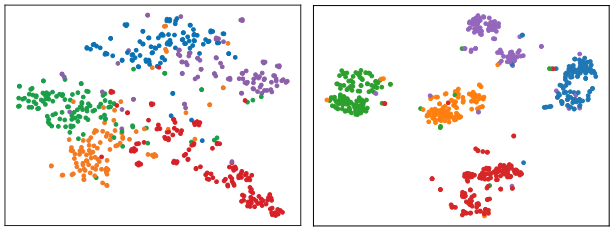
为了进一步定性分析我们的 CLIP-FSAR，我们在图 6 中可视化了测试阶段进行原型调制前后特征分布的变化。通过比较结果，我们可以观察到明显的提升：经过原型调制后，类内分布更加紧凑，类间特征也更具区分性。



(a) 未进行原型调制 (b) 进行原型调制**图 7：** Kinetics 测试集上五个动作类别的 T-SNE 分布可视化。这五个类别分别是“玩大富翁”、“呼啦圈”、“切西瓜”、“剪羊毛”和“打鼓”。



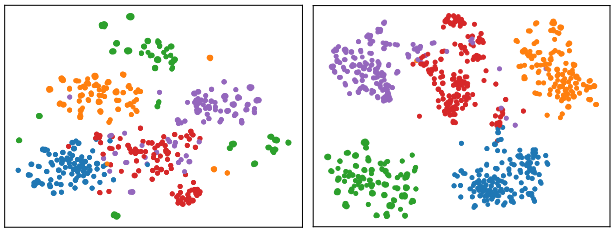
(a) 未进行原型调制 (b) 进行原型调制**图 8：** UCF101 测试集上五个动作类别的 T-SNE 分布可视化。这五个类别分别是“萨尔萨旋转”、“跳水”、“出拳”、“网球挥拍”和“在厨房切东西”。



(a) 未进行原型调制 (b) 进行原型调制 图 9：SSv2-Full 测试集上五个动作类别的 T-SNE 分布可视化。这五个类别分别是“用我们的摄像机靠近 [某物] ”、“从 [某物]中取出 [某物] ”、“将 [某物]从右向左推”、“从 [某物]中倒出 [某物] ”和“在 [某物]旁展示 [某物]。”

原型调制。上述发现验证了我们提出的 CLIP-FSAR 的合理性和优越性。

我们进一步可视化了另外四个数据集上的类别分布结果，包括 Kinetics（图7）、UCF101（图8）、SSv2-Full（图9）和 HMDB51（图10）。我们从这些数据集的测试集中随机抽取五个类别进行可视化。结果表明，引入原型调制后，各类别彼此变得更加可区分，且类内分布更加紧密。例如，如图7所示，在未引入原型调制时，



(a) 无原型调制 (b) 有原型调制**图10：** HMDB51测试集上五个动作类别的T-SNE分布可视化。这五个类别分别是“跳水”、“攀爬”、“爬楼梯”、“喝水”和“运球”。

表11： 不同提示模板的对比实验。

提示模板	SSv2-Small		Kinetics	
	1样本	5样本	1样本	5样本
“[CLS]”	**53.2**	**55.9**	87.3	91.6
“ [CLS]的一张照片”	52.0	55.8	87.6	91.9
“ [CLS]， 一种动作”	51.6	54.6	**88.4**	**92.1**

“玩大富翁”、“呼啦圈”和“剪羊毛”这三个类别相互混杂。然而，引入原型调制后，这些类别之间变得清晰可区分。此外，如图9所示，两个细粒度动作类别“取出 [某物] ， 从 [某物]’ 和“倒出 [某物]， 从 [某物]’的可区分性也因引入原型调制而显著提升，这得益于利用 CLIP 强大的文本语义先验来补充视觉识别。

4.5 局限性

在我们的CLIP-FSAR中，我们采用广泛使用的提示模板，即“ [CLS]的一张照片”，作为默认设置。在表11中，我们探究了不同文本提示的影响，发现不同提示模板在不同数据集上的表现并不一致。例如，“[CLS]”在SSv2-Small上表现最佳，而在Kinetics数据集上表现最差。研究通用且有效的文本提示形式的设计将是一项很有价值的工作。此外，我们主要关注从视觉方面调制原型，并未考虑文本端的一些潜在改进，例如利用视觉特征来优化文本特征。我们将上述讨论留待未来工作。

5 结论

本文提出一种 CLIP-FSAR 方法，利用 CLIP 模型解决少样本动作识别问题，并充分挖掘 CLIP 的多模态知识。我们采用视频-文本对比目标来

使 CLIP 模型适应少样本视频任务。此外，我们提出利用文本特征，通过实现时间Transformer，自适应地调节视觉支持原型。在五个常用基准数据集上进行的大量实验表明，我们的 CLIP-FSAR 显著优于当前最先进方法。

数据可用性声明。在本研究期间生成和/或分析的数据集可在我们的开源代码库中获取。

致谢 本工作得到国家自然科学基金委员会项目U22B2053的资助，并由阿里巴巴集团通过阿里巴巴研究实习项目提供支持。

参考文献

1. Arnab A、Dehghani M、Heigold G、Sun C、Lu˘ci´c M、Schmid C（2021）Vivit：视频视觉Transformer。载于：ICCV，第6836–6846页 1 2. Bishay M、Zoumpourlis G、Patras I（2019）TARN：用于小样本和零样本动作识别的时间注意力关系网络。载于：BMVC，BMVA Press，第154页，网址：https://bmvc2019.org/wp-content/uploads/papers/0650 6 3 -paper.pdf。Cao K、Ji J、Cao Z、Chang CY、Niebles JC（2020）通过时间对齐实现小样本视频分类。载于：CVPR，第10618–10627页 1、2、3、4、5、6、7、8 4. Carreira J、Zisserman A（2017）动作识别何去何从？一种新模型与Kinetics数据集。载于：CVPR，第6299–6308页 1、3、5 5. Chen CFR、Panda R、Ramakrishnan K、Feris R、Cohn J、Oliva A、Fan Q（2021）深入分析基于CNN的用于动作识别的时空表征。载于：CVPR，第6165–6175页 8 6 2014 – 。Chung J、Gulcehre C、Cho K、Bengio Y（ ）门控循环神经网络在序列建模上的实证评估。arXiv预印本，arXiv:14123555 8 7. Dai Z、Yang Z、Yang Y、Carbonell J、Le QV、Salakhutdinov R（2019）Transformer-XL：超越固定长度上下文的注意力语言模型。arXiv预印本，arXiv:190102860 8 8. Deng J、Dong W、Socher R、Li LJ、Li K、Fei-Fei L（2009）ImageNet：大规模层次化图像数据库。载于：CVPR，第248–255页 1 、 3 9. Dosovitskiy A 、 Beyer L 、 Kolesnikov A 、 Weissenborn D、Zhai X、Unterthiner T、Dehghani M、Minderer M、Heigold G、Gelly S等（2021）一张图像胜过16×16个词：用于大规模图像识别的Transformer。载于：国际学习表征会议 3、5 10. Fei-Fei L、Fergus R、Perona P（2006）物体类别的单样本学习。IEEE模式分析与机器智能汇刊28 611 2 11 4 594（）：–。Finn C、Abbeel P、Levine S（2017）用于深度网络快速适应的模型无关元学习。载于：ICML，PMLR，第1126–1135页 2、6、7 12. Gao P、Geng S、Zhang R、Ma T、Fang R、Zhang Y、Li H、Qiao Y（2021）CLIP-Adapter：借助特征适配器改进视觉-语言模型。arXiv预印本，arXiv:211004544 1、2 13. Goyal R、Michalski V、Materzy J、Westphal S、Kim H、Haenel V、Yianilos P、Mueller-freitag M、Hoppe F、Thurau C、Bax I、Memisevic R（2017）用于学习和评估视觉常识的“Something Something”视频数据库。载于：ICCV，网址：https://www.

14. Graves A, Mohamed Ar, Hinton G（2013）基于深度循环神经网络的语音识别。载于：ICASSP，第 6645–6649 页 8 15 2022 – . Gu X, Lin TY, Kuo W, Cui Y（ ）通过视觉与语言知识蒸馏实现开放词汇目标检测。载于：ICLR 2, 3 16. Guo Y, Codella NC, Karlinsky L, Codella JV, Smith JR, Saenko K, Rosing T, Feris R（2020）对跨域小样本学习的更广泛研究。载于：ECCV，Springer，第 124–141 页 2 17 2017 – . Hariharan B, Girshick R（ ）通过收缩和特征幻觉生成实现低样本视觉识别。载于：ICCV，第 3018–3027 页 2 18 2016 – . He K, Zhang X, Ren S, Sun J（ ）用于图像识别的深度残差学习。载于：CVPR，第 770–778 页 3, 5 19 2022 – . Huang Y, Yang L, Sato Y（ ）用于小样本动作识别的复合原型匹配。载于：ECCV 3, 6, 7 20 2019 . Jamal MA, Qi GJ（ ）面向小样本学习的任务无关元学习。载于：CVPR，第 11719–11727 页 2 21 – . Jia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, Le Q, Sung YH, Li Z, Duerig T（2021）通过噪声文本监督扩展视觉与视觉-语言表征学习。载于：ICML，PMLR，第 4904–4916 页 1, 2 22 – . Ju C, Han T, Zheng K, Zhang Y, Xie W（2022）通过提示视觉-语言模型实现高效视频理解。载于：ECCV 3 23 2014 . King ma DP, Ba J（ ）Adam：一种随机优化方法。arXiv 预印本 arXiv:14126980 5 24. Kuehne H, Serre T, Jhuang H, Garrote E, Poggio T, Serre T（2011）HMDB：一个用于人体动作识别的大型视频数据库。载于：ICCV，DOI 10 6126543 5 25 1109 2011 . / ICCV. . . Li B, Weinberger KQ, Belongie S, Koltun V, Ranftl R（2022）语言驱动的语义分割。载于：ICLR 4 26 2019 . Li H, Eigen D, Dodge S, Zeiler M, Wang X（ ）通过类别遍历为小样本学习寻找任务相关特征。载于：CVPR，第 1–10 页 2 27 – . Li J, Selvaraju R, Gotmare A, Joty S, Xiong C, Hoi SCH（2021）先对齐再融合：通过动量蒸馏进行视觉与语言表征学习。NeurIPS 34:9694–9705 2 28. Li K, Zhang Y, Li K, Fu Y（2020）用于小样本学习的对抗性特征幻觉生成网络。载于：CVPR，第 13470–13479 页 2 29. Li S, Liu H, Qian R, Li Y, See J, Fei M, Yu X, Lin W（2022）Ta2n：用于小样本动作识别的两阶段动作对齐网络。载于：AAAI，第 1404–1411 页 3, 6, 7 30 – . Li W, Gao C, Niu G, Xiao X, Liu H, Liu J, Wu H, Wang H（2020）Unimo：通过跨模态对比学习实现统一模态理解与生成。arXiv 预印本 arXiv:201215409 2 31. Li X, Yin X, Li C, Zhang P, Hu X, Zhang L, Wang L, Hu H, Dong L, Wei F, 等（2020）Oscar：面向视觉-语言任务的对象-语义对齐预训练。载于：ECCV，Springer，第 121–137 页 2 32. Li Z, Zhou F, Chen F, Li H（2017）Meta-sgd：快速学习小样本学习的元学习方法。arXiv 预印本 arXiv:170709835 2 33 2019 . Lin J, Gan C, Han S（ ）TSM：用于高效视频理解的时间移位模块。载于：ICCV，第 7083–7093 页 1 34 – . Lin Z, Geng S, Zhang R, Gao P, de Melo G, Wang X, Dai J, Qiao Y, Li H（2022）冻结的 CLIP 模型是高效的视频学习器。载于：ECCV 3 35. Liu Y, Xiong P, Xu L, Cao S, Jin Q（2022）TS2-Net：用于文本-视频检索的令牌移位与选择变换器。载于：ECCV 3 36. Luo J, Li Y, Pan Y, Yao T, Chao H, Mei T（2021）CoCo-BERT：通过对比式跨模态匹配和去噪改进视频-语言预训练。载于：ACMMM，第

5600 5608 2 – .

37 ¨ 2007

Muller M（ ）动态时间规整。《音乐与动作信息检索》第 69–84 页 3, 4 38 – . Nguyen KD, Tran QH, Nguyen K, Hua BS, Nguyen R（2022）通过外观和时间对齐进行归纳式与传导式小样本视频分类。载于：ECCV 3, 6, 7

39. Pahde F, Ostapenko O, Hnichen PJ, Klein T, Nabi M（2019）用于多模态小样本学习的自步对抗训练。载于：WACV，IEEE，第 218–226 页 2 40 – . Pahde F, Puscas M, Klein T, Nabi M（2021）用于小样本学习的多模态原型网络。载于：WACV，第 2644–2653 页 2 41. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, 等（2019）PyTorch：一种命令式风格、高性能的深度学习库。NeurIPS 32 5 42. Perrett T, Masullo A, Burghardt T, Mirmehdi M, Damen D（2021）用于小样本动作识别的时序关系交叉变换器。载于：CVPR，第 475–484 页 3, 5, 6, 7, 8 43 – . Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, 等（2021）通过自然语言监督学习可迁移视觉模型。载于：ICML，PMLR，第 8748–8763 页 1, 2, 3, 6, 7 44 – . Rajeswaran A, Finn C, Kakade SM, Levine S（2019）基于隐式梯度的元学习。载于：NeurIPS，第 32 卷 2 45. Rao Y, Zhao W, Chen G, Tang Y, Zhu Z, Huang G, Zhou J, Lu J（2022）DenseCLIP：通过上下文感知提示进行语言引导的密集预测。载于：CVPR，第 18082–18091 页 1, 2, 4 46 2017 – . Ravi S, Larochelle H（ ）将优化视为小样本学习的模型。载于：ICLR 2 47. Rusu AA, Rao D, Sygnowski J, Vinyals O, Pascanu R, Osindero S, Hadsell R（2019）基于潜在嵌入优化的元学习。载于：ICLR 2 48. Shi H, Hayat M, Wu Y, Cai J（2022）ProposalCLIP：利用 CLIP线索进行无监督开放类别目标提议生成。载于：CVPR，第 9611–9620 页 2, 4 49 – . Shi Z, Liang J, Li Q, Zheng H, Gu Z, Dong J, Zheng B（2021）多模态多动作视频识别。载于：ICCV，第 13678–13687 页 4 50. Snell J, Swersky K, Zemel R（2017）用于小样本学习的原型网络。NeurIPS 30 4087 2, 3, 4, 5 51 4077 2012 101 : – . Soomro K, Zamir AR, Shah M（ ）UCF101：一个包含来自自然场景视频的 101 类人体动作的数据集。arXiv URL http://arxiv.org/abs/1212 0402 5 51 0402, 1212. . . Sung F, Yang Y, Zhang L, Xiang T, Torr PH, Hospedales TM（2018）学会比较：用于小样本学习的关系网络。载于：CVPR，第 1199–1208 页 2, 3 53 – . Thatipelli A, Narayan S, Khan S, Anwer RM, Khan FS, Ghanem B（2022）面向小样本动作识别的时空关系建模。载于：CVPR，第 19958–19967 页 3, 6, 7 54 – . Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I（2017）注意力就是你所需要的一切。载于：NeurIPS，第 5998–6008 页 3, 8 55 – . Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K, Wierstra D（2016）单样本学习的匹配网络。载于：NeurIPS，arXiv:1606 2 1, 2, 3, 6, 7 56 04080. v . Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L（2016）时序片段网络：迈向深度动作识别的良好实践。载于：ECCV，Springer，第 20–36 页 3 57 2021 – . Wang M, Xing J, Liu Y（ ）ActionCLIP：视频动作识别的新范式。arXiv 预印本 arXiv:210908472 2 58. Wang T, Jiang W, Lu Z, Zheng F, Cheng R, Yin C, Luo P（2022）VLMixer：通过跨模态CutMix进行非配对视觉语言预训练。载于：ICML，PMLR，第 22680–22690 页 2 59 – . Wang X, Zhang S, Qing Z, Shao Y, Gao C, Sang N （ 2021 ）用于半监督时序动作提议的自监督学习。载于：CVPR，第 1905–1914 页 1 60 – . Wang X, Zhang S, Qing Z, Shao Y, Zuo Z, Gao C, Sang N（2021）OADTR：基于Transformer的在线动作检测。载于：ICCV，第 7565–7575 页 1 61. Wang X, Zhang S, Qing Z, Tang M, Zuo Z, Gao C, Jin R, Sang N（2022）用于小样本动作识别的混合关系引导集合匹配

识别。载于：CVPR，第 19948–19957 页 1, 3, 5, 6, 7, 8 62 – 。Wang X, Zhang S, Qing Z, Zuo Z, Gao C, Jin R, Sang N (2023) HyRSM++：混合关系引导的时序集合匹配，用于小样本动作识别。arXiv 预印本 arXiv:230103330 3 63 2022 。Wang Z, Lu Y, Li Q, Tao X, Guo Y, Gong M, Liu T ( ) Cris：由CLIP驱动的指代表达图像分割。载于：CVPR，第 11686–11695 页 2, 4 64。Wu J, Zhang T, Zhang Z, Wu F, Zhang Y (2022) 用于小样本动作识别的运动调制时序片段对齐网络。载于：CVPR，第 9151–9160 页 3, 6, 7 65 – 。Xing C, Rostamzadeh N, Oreshkin B, O Pinheiro PO (2019) 自适应跨模态小样本学习。NeurIPS 32 2 66 2022 。Xu W, Xian Y, Wang J, Schiele B, Akata Z ( ) 用于任意样本学习的属性原型网络。国际计算机视觉期刊 130 1753 2 67 7 1735 2022( ): – 。Yang J, Li C, Zhang P, Xiao B, Liu C, Yuan L, Gao J ( ) 图像-文本-标签空间中的统一对比学习。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第 19163–19173 页 2 68 2020– 。Ye HJ, Hu H, Zhan DC, Sha F ( ) 通过集合到集合函数的嵌入适配进行小样本学习。载于：CVPR，第 8808–8817 页 2 69。Ye HJ, Hu H, Zhan DC (2021) 为广义小样本学习合成自适应分类器。国际计算机视觉期刊 129 1953 2 70 1930 2019： – 。Yoon SW, Seo J, Moon J ( ) Tapnet：通过任务自适应投影增强的神经网络，用于小样本学习。载于：ICML，PMLR，第 7115–7123 页 2 71 – 。Zhai X, Wang X, Mustafa B, Steiner A, Keysers D, Kolesnikov A, Beyer L (2022) Lit：通过锁定图像文本调优实现零样本迁移。载于：CVPR，第 18123–18133 页 2 72 – 。Zhang H, Zhang L, Qi X, Li H, Torr PH, Koniusz P (2020) 采用置换不变注意力的小样本动作识别。载于：ECCV，Springer，第 525–542 页 3, 5, 6, 7 73 – 。Zhang H, Li F, Liu S, Zhang L, Su H, Zhu J, Ni L, Shum H (2022) DINO：采用改进去噪锚框的DETR，用于端到端目标检测。载于：ICLR 9 74。Zhang R, Che T, Ghahramani Z, Bengio Y, Song Y (2018) Meta-gan：一种用于小样本学习的对抗方法。NeurIPS 31 2 75。Zhang R, Fang R, Gao P, Zhang W, Li K, Dai J, Qiao Y, Li H (2022) Tip-adapter：用于改进视觉-语言建模的免训练CLIP-Adapter。载于：ECCV 1 76 2021 。Zhang S, Zhou J, He X ( ) 学习小样本视频分类的隐式时序对齐。载于：国际人工智能联合会议 3, 5, 6, 7, 8 77 2022 。Zheng S, Chen S, Jin Q ( ) 采用层次匹配和对比学习的小样本动作识别。载于：ECCV，Springer 3, 6, 7 78。Zhong Y, Yang J, Zhang P, Li C, Codella N, Li LH, Zhou L, Dai X, Yuan L, Li Y, 等 (2022) Regionclip：基于区域的语言-图像预训练。载于：CVPR，第 16793–16803 页 2 79 2018– 。Zhou B, Andonian A, Oliva A, Torralba A ( ) 视频中的时序关系推理。载于：ECCV，第 803–818 页 6 80 2022 – 。Zhou K, Yang J, Loy CC, Liu Z ( ) 面向视觉-语言模型的条件提示学习。载于：CVPR，第 16816–16825 页 2 81 2022 – 。Zhou K, Yang J, Loy CC, Liu Z ( ) 面向视觉-语言模型的学习提示。国际计算机视觉期刊 130 2348 1, 2 82 9 2337 2018 ( ): – 。Zhu L, Yang Y ( ) 用于小样本视频分类的复合记忆网络。载于：ECCV，第 751–766 页 1, 3, 5, 6, 7 83 2020 – 。Zhu L, Yang Y ( ) 用于半监督小样本视频分类的标签无关记忆。IEEE模式分析与机器智能汇刊 3, 6, 7 84。Zhu X, Toisoul A, Perez-Rua JM, Zhang L, Martinez B, Xiang T (2021) 采用原型中心注意力学习的小样本动作识别。arXiv 预印本 arXiv:210108085 3