第 13 卷 第 6 期
2024 年 11 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 13 No. 6
Nov. 2024

引文格式：
许清林, 乔宇, 王亚立. 基于领域上下文辅助的开放域行为识别 [J]. 集成技术, 2024, 13(6): 31-43.
Xu QL, Qiao Y, Wang YL. Open domain action recognition based on domain context assistance [J]. Journal of Integration Technology, 2024, 13(6): 31-43.

基于领域上下文辅助的开放域行为识别

许清林[1,2] 乔 宇[3] 王亚立[1,3*]

[1]（中国科学院深圳先进技术研究院 深圳 518055）
[2]（中国科学院大学 北京 100049）
[3]（上海人工智能实验室 上海 200232）

摘 要 如何将预训练模型中的知识迁移到视频理解下游任务是计算机视觉研究中的一个关键问题。在开放域场景中，由于不利的数据条件，知识迁移变得更具挑战性。受自然语言处理技术的启示，近期，许多多模态预训练模型通过设计文本提示进行迁移学习。作者利用大语言模型对开放域的理解能力，提出一种基于领域上下文辅助的开放域行为识别方法，提升模型在开放域场景下的理解能力。通过大语言模型对文本标签的上下文信息进行丰富，将视觉表示与人类行为的多层次描述进行对齐，实现鲁棒的分类。在开放域场景下进行了广泛的行为识别实验，在全监督设置中，该文方法在 ARID 数据集上得到了 71.86% 的 Top1 准确率，而在 Tiny-VARIT 数据集上得到了 80.93% 的平均精确率。此外，在无源视频领域自适应设置下，该研究得到了 48.63% 的 Top1 准确率，而在多源视频领域自适应设置中，该研究得到了 54.36% 的 Top1 准确率，实验结果表明了领域上下文辅助在各种开放域环境下的有效性。

关键词 开放域行为识别；迁移学习；大语言模型；多模态
中图分类号 TP183 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20231226001

Open Domain Action Recognition Based on Domain Context Assistance

XU Qinglin[1,2] QIAO Yu[3] WANG Yali[1,3*]

[1](*Shenzhen Institute of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[2](*University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)
[3](*Shanghai Artificial Intelligence Laboratory*, *Shanghai* 200232, *China*)
[*]*Corresponding Author*: *yl.wang@siat.ac.cn*

Abstract Effectively transferring knowledge from pre-trained models to downstream video understanding

收稿日期：2023-12-26 **修回日期**：2024-03-07
基金项目：国家重点研发计划(2022ZD0160505)，国家自然科学基金项目(62272450)
作者简介：许清林，硕士研究生，研究方向为计算机视觉、行为识别和多模态学习等；乔宇，博士，研究员，研究方向为计算机视觉、深度学习、行为识别、场景识别、人脸识别和目标检测等；王亚立(通讯作者)，博士，研究员，研究方向为计算机视觉、深度学习和行为识别等，E-mail：yl.wang@siat.ac.cn。

tasks is an important topic in computer vision research. Knowledge transfer becomes more challenging in open domain due to poor data conditions. Many recent multi-modal pre-training models are inspired by natural language processing and perform transfer learning by designing prompt learning. The paper leverages the comprehension ability of large language models over open domains and proposes a domain-context-assisted method for open-domain behavior recognition. This approach aligns visual representation with multi-level descriptions of human actions for robust classification, by enriching action labels with context knowledge in large language model. In the experiments of open-domain action recognition with fully supervised setting, it obtain a Top1 accuracy of 71.86% on the ARID dataset, and an mean average precision of 80.93% on the Tiny-VARIT dataset. More important, it can achieve Top1 accuracy of 48.63% in source-free video domain adaptation and 54.36% in multi-source video domain adaptation. The experimental results demonstrate the efficacy of domain context-assisted in a variety of open domain environments.

Keywords open-world action recognition; transfer learning; large language model; multi-modality

Funding This work is supported by National Key Research and Development Program of China (2022ZD0160505), National Natural Science Foundation of China (62272450)

1 引 言

近年来，众多多模态视觉语言模型[1-3]通过大规模图像-文本对进行对比学习预训练，在各类视觉任务上展现出卓越的迁移能力。利用这些强大的预训练模型进行迁移学习正逐渐成为行为识别领域的一个有效策略。已有众多研究成功将这些模型有效迁移到行为识别领域，并在多个学术数据集上取得显著成果。

在常规数据集中，现有通用方法[4-6]虽然展现出优异性能，但在面对开放域的复杂环境与苛刻数据条件时，往往表现不佳。在开放域场景中，视频数据通常受黑暗[7]、低分辨率[8]等多种复杂因素影响，同时，预训练数据与训练数据集的数据分布差异也加大了模型迁移的难度。此外，在复杂苛刻的场景下[7-8]，当缺少标签进行有监督学习时，现有的领域自适应方法往往因域差异过大而效果不佳。因此，如何无监督地将模型适应至目标领域成为另一关键挑战。针对上述问题，本研究提出一个研究目标：开发一种新的迁移学习策略，以提升多模态预训练模型对开放域视频中恶劣环境的适应性和识别准确性。

在行为识别任务中应用多模态预训练模型时，通常将由类别标签生成的语义信息作为监督目标。然而，类别标签的形式是一个关键问题。例如，在暗光环境下喝水行为的视频可以有不同的描述：简单的“喝水”，描述性的“这是一个人在喝水的视频”，或具体的“人在黑暗场景下举起水杯喝水”。合适的手工描述对识别的结果至关重要[1]。与简单标签文本相比，手工设计专门的文本提示虽然能在一定程度上减轻领域差异，但手动编写耗时且难以适用于众多类别。近期出现的自动化提示学习[8]虽然在一定程度上解决了上述问题，但是在面对开放域复杂场景时，它的效果依然有限。

近期涌现的大语言模型(large language model，LLM)[9-10]展示了其对开放域的理解能力。本研究通过将领域先验知识和动作标签输入到 LLM

来丰富动作标签的上下文信息。具体来说，本文提出的领域上下文辅助方法利用 LLM 对每个标签进行深入分析，将标签文本与目标领域的描述作为输入，以获取包含目标领域上下文信息的多层次描述。通过上述方法，视频识别模型可更准确地将视觉信息与人类行为的多维描述对齐，从而提高分类的鲁棒性。由于 LLM 提供的辅助文本细致刻画了行为类别可能出现的场景，直接将视频与所有辅助文本进行匹配并不合适。因此本研究借鉴了多实例学习(multiple instance learning)策略，仅匹配最能精确描述视频核心特征的辅助文本。上述方法在多模态视频理解领域具备双重优势：一是更为精确的描述可更有效地与特定行为进行匹配；二是可为文本注入目标数据集的先验知识。例如，在黑暗环境中，被推动的物体通常难以辨认，行走和推动的动作难以区分。在这种情况下，通过使用更具体且包含领域信息的描述，如“使用手触摸不易辨认的物体”或“身体前倾推动物体”，模型将更容易准确识别这些行为。本文在全监督、无源视频领域自适应和多源视频领域自适应多个设定下对领域上下文辅助方法进行了评估。结果表明：在开放域场景下，与现有的多模态预训练模型迁移学习方法相比，本文方法达到较先进的性能。此外，本研究发现了多模态预训练模型在视频领域自适应场景中的应用潜力，与之前的方法相比，取得显著提升。

本文的贡献可总结为以下几点：(1)创新性地结合了 LLM 和多模态与训练模型的知识，探索了 LLM 在开放域场景方面的应用潜力；(2)通过生成具有领域上下文的多层次描述文本，可更好地进行多模态对齐，并可实现鲁棒分类；(3)在全监督和领域自适应的基准下进行大量的开放域行为识别实验，结果表明，本文的方法明显优于目前最先进的方法，尤其是在领域自适应场景，取得了显著的结果。

2　国内外研究现状

2.1　开放域视频理解

近年来，得益于各种模型的创新和大规模视频数据集的构建，深度学习技术在视频理解领域取得了显著的进展[6,11]。早期，视频理解研究主要分为两个方向：通过发展 3D 卷积神经网络[12-13]捕获视频中的时间信息；通过双流网络的设计[14-15]，以及并行处理 RGB 图像和光流信息，获取空间和时间维度信息。Transformer[16]架构出现之后，涌现了许多基于此技术的创新工作[17-19]。基础模型的涌现[1-3]使得许多研究聚焦于如何将其迁移到视频相关的下游任务，如 Frozen CLIP[20]和 UniFormerV2[21]展示了基于 CLIP[1]的模型在视频理解中的有效应用[22-23]。尽管如此，这些模型在实际应用中依然经常受限于复杂和恶劣的数据条件。

在复杂环境中，视频理解模型面临提取关键特征的挑战，如低光照条件可能导致模型忽略重要信息。DarkLight 研究[24]通过伽马校正技术增强视觉数据，减少低照度对模型性能的影响。此外，由于获取标注数据往往需要大量的人力成本，因此迫使模型必须能在数据匮乏的情况下有效学习。针对这一问题，视频领域自适应成了研究的重点，如 ATCoN[25]通过学习不同尺度上的视图一致性实现无监督迁移。然而，这些方法在面对具有明显域差异的开放域视频时表现不足。

2.2　提示学习

最初，提示学习(prompt learning)在自然语言处理领域被提出，并已在视觉多模态预训练模型的下游任务中证明其重要性。起初，研究者通常使用手动编写的提示语句(如“一个[cls]的视频”)作为视觉编码器的输入，但手动编写的提示语的质量显著影响识别效果，且其设计成本高昂。为应对这一挑战，CoOp 引入了一组参数化的文本上下文，针对特定数据集进行优化，以提

高提示的质量[26]。然而，该方法在处理未知类别时表现不佳。为此，CoCoOp 进一步在 CoOp 的基础上加入轻量级神经网络，为每个图像生成独特的向量，进一步优化提示效果[27]。与基于训练的提示方法不同，本研究利用大语言模型强大的理解能力生成具有特定数据集领域上下文的细致描述，以改善模型的识别能力。

2.3 基础模型在迁移学习中的应用

随着基础模型(foundation model)的不断发展，越来越多的研究集中在如何更有效地利用基础模型中的知识方面，以便有效地将其迁移到目标领域或应用到其他模型中。CaFo 专注于少样本学习领域[28]，首先使用 DALL-E[29]生成额外的训练数据，然后通过 CLIP[1]和 DINO[30]以 Tip-Adapter[31]的方式进行少样本学习，并使用生成的数据对分类头进行微调。Cap4Video[32]则充分利用视觉描述生成模型[33]和 GPT-3[9]生成的辅助文本提升 CLIP[1]在视频检索任务中的表现。LLaMA-Adapter[34]则是设计一种简单的适配器，只需微调少量参数，就可让 LLaMA[35]适用于不同场景。尽管如此，这些方法仍未充分挖掘视频领域自适应中基础模型的潜力。

3 领域上下文辅助的开放域行为识别方法

本节将详细介绍利用领域上下文辅助的开放域视频理解方法。针对开放域场景下现有多模态行为识别方法的局限性与提示学习在多模态学习中的重要性，本文提出一种新颖的方法：利用大语言模型扩展现有的标签文本，丰富视频内容的语义理解，并向视觉模型注入特定领域的先验知识，同时由于 LLM 生成的辅助文本覆盖了视频类别的广泛可能场景，直接将视频与全部辅助文本进行对应并不合适。因此，本方法采用多实例学习策略，旨在找到与视频内容最匹配的辅助文本。本文提出的方法旨在深入挖掘与特定环境(如黑暗场景)相关的隐含信息，以实现更准确和全面的视频理解。

3.1 领域扩展文本的生成

为丰富文本语义信息，实现更好的多模态匹配，之前的研究提出了一些方法。例如，ActionCLIP 为减少对手工文本提示的过度敏感，预设了 16 个模板，在训练过程中随机选择模板，推理结果则取所有模板结果的平均值[6]。CoOp 通过可学习向量自动化文本提示的学习，捕获数据集特有的领域信息[26]。X-CLIP 尝试将视频特征融入文本表示中，利用视频内容增强文本的表现力[36]。尽管这些方法在减少对手动文本提示的依赖上取得了进步，但对文本的语义丰富度仍有限。本研究利用 LLM 深度理解开放域的能力，通过生成细致的领域特定细粒度标签文本，有效补充了以往方法在信息量上的不足。

如图 1 左部分所示，本文向 ChatGPT[10]提供有关具体数据集的领域知识和标签文本集合等信息，指导 ChatGPT 根据每个标签文本生成一系列详细描述，这些描述通过展示目标领域数据集的特定上下文信息，可获得更全面和与上下文更紧密相关的理解。具体而言，每个标签文本 $T \in \mathcal{T}$，本文利用 ChatGPT 生成 K 个扩展文本，表示为 $\mathrm{LLM}(T)=\{e_1,e_2,\cdots,e_k\}$。其中，$e_k$ 为第 k 个扩展文本，$\mathcal{T}$ 为所有类别的文本标签。

3.2 领域扩展文本辅助策略

LLM 凭借对现实世界场景的深刻理解，能够为标签文本提供详细而丰富的扩展描述，形成 LLM 扩展文本。需要明确的是，并非所有视频都能与相应类别的 LLM 扩展文本准确对应。这是因为 LLM 扩展文本的目的是为每一指定类别绘制出更为细粒度的可能场景画面。因此，不应期望每个视频都与其对应类别的全部 LLM 扩展文本严格对齐。本研究借鉴多实例学习策略，视频内容应与最能准确描述其核心特征的 LLM 扩展文本进行匹配。例如：类别是“拿取

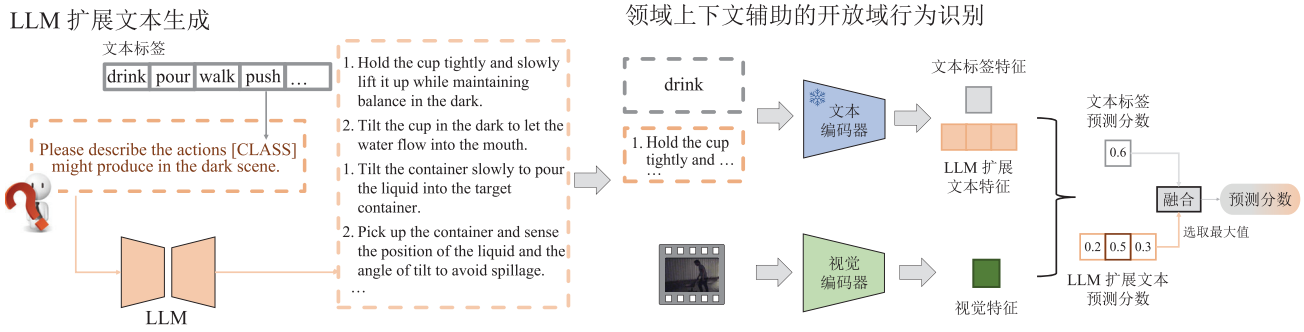

图 1　领域上下文辅助流程

Fig. 1　Domain context-assist process

行为”(pick)的扩展文本包括“有人举手触摸物体的视频”和“有人低头在桌子上找东西并拿取的视频”。当某个视频展示出一个人举起手在寻找高处的物体时，虽然“有人低头在桌子上找东西并拿取的视频”也属于“拿取行为”的扩展文本，但是本研究更倾向于将视频与“有人举手触摸物体的视频”的扩展文本对齐，因为它与视频的行为更为对应。

在模型训练的过程中，通过视觉和文本编码器，领域上下文辅助以视觉、标签文本和 LLM 扩展文本为输入，可以得到视觉特征$\left\{v_n \in \mathbb{R}^d \middle| n=1,2,\cdots,N\right\}$、文本特征$\left\{t_m \in \mathbb{R}^d \middle| m=1,2,\cdots,M\right\}$和 LLM 扩展文本特征$\left\{e_{mk} \in \mathbb{R}^d \middle| m=1,2,\cdots,M,\ k=1,2,\cdots,K\right\}$。其中，$d$ 为特征的维数；N 为取样视频的数量；M 为类别数；K 为每个类别的扩展文本数。值得指出的是，标签文本和 LLM 扩展文本均使用相同的文本编码器进行编码。通过计算视频与标签文本之间的余弦相似度，可评估它们的相关性，得到每个视频与所有类别的相似度：

$$S_{v_n t_m}=\mathrm{sim}\left(v_n,t_m\right) \tag{1}$$

其中，$S_{v_n t_m}$ 为视频 v_n 和文本标签 t_n 的相似度；sim 为余弦相似度函数。同样地，计算视频与 LLM 扩展文本之间的余弦相似度，并且为了视频能与最能够准确描述其核心特征的 LLM 扩展文本进行匹配，在每个类别的 LLM 扩展文本中，领域上下文辅助只保留其中的最大值来作为视频与该类 LLM 扩展文本的相似度：

$$S_{v_n e_m}=\arg\max_k \mathrm{sim}\left(v_n,e_{mk}\right) \tag{2}$$

其中，$S_{v_n e_m}$ 为视频 v_n 和 LLM 扩展文本 e_m 中所有描述的最大相似度；e_m 为 m 类别的所有扩展文本。在训练和推理过程中，领域上下文辅助采用式(3)将上述两个相似度分支结合起来：

$$S=\alpha S_{vt}+(1-\alpha)S_{ve} \tag{3}$$

其中，α 为预测结果融合的权值系数。最后，如果视频 n 和类别 m 是匹配的，那么就最大化 S_{nm} 的值，否则就最小化它，以这个为目的，最终的损失函数可表示为

$$\mathcal{L}=-\log\frac{\exp\left(S_{n\phi(n)}\right)}{\sum_{m=1}^{M}\exp\left(S_{nm}\right)} \tag{4}$$

其中，$\phi(n)$为第 n 个视频所对应的类别。

4　结　果

本节将详细介绍本文针对开放域场景下的实验设置及结果。由于有监督训练在这些场景中往往难以取得优异成绩，因此，本文选择全监督设定下的黑暗和低分辨率的数据集(即 ARID[7] 和 TinyVIRAT[8])进行验证实验。这些实验旨在证明领域上下文辅助方法面对显著域差异挑战时的有效性。在领域自适应方面，本研究选用当前主流的视频领域自适应数据集 Daily-DA[25]进行测

试，并获得了行业领先的成果。进一步地，本研究通过消融实验深入验证领域上下文辅助功能的重要性及文本在不同场景中的角色。最终，本文利用可视化展示 LLM 扩展文本如何有效纠正分类结果。

4.1 框架设定和数据集介绍

在本研究的实验评估中，由于 CLIP[1]在视频领域众多下游任务中展现出卓越的迁移能力，因此，本文选择其作为主干网络。具体而言，本研究采用 CLIP VIT-B/16 作为视觉编码器。该编码器基于 12 层的 Transformer[16]架构。在文本编码方面，本研究直接应用 CLIP 中的文本编码器生成标签和描述的特征表示，文本编码器是一个包含 12 层、维度为 512、拥有 8 个注意力头的 Transformer 模型。为了与 CLIP 的设计保持一致，本文选取“[CLS] token”对应的特征作为整个语句的特征表示[37]。公式(3)中的 α 取值为 0.2。

领域上下文辅助在 4 个基准数据集(ARID[7]、TinyVIRAT[8]、SFVDA Daily-DA[25]和 MSVDA Daily-DA[25])上进行了严格的评估。ARID 数据集专注于黑暗场景下的行为识别，包含 3 784 个视频片段，覆盖 11 个不同的动作类别。TinyVIRAT 聚焦于自然场景下的低分辨率活动，是一个多标签分类任务。该数据集从监控视频中提取，因此呈现出更加真实且挑战性的特点，涵盖 26 个动作类别和 20 258 个视频片段。SFVDA Daily-DA 是一个无源域适应的基准，包括 4 个数据集：正常光照条件下的 HMDB51[38](H51)、Moments in time[39](MIT)和 Kinetics 600[40](K600)，以及黑暗场景的 ARID (A11)。在该设定下，识别模型需要选择一个数据集作为源域，另一个数据集作为目标域。本文选择 3 个以 ARID 为目标域的跨领域任务，以突出领域上下文辅助在将正常光照条件下训练的模型无监督迁移到黑暗场景中的性能。MSVDA Daily-DA 数据集与 SFVDA 类似，但在多源视频域适应的设定中，本文将 3 个正常光照条件下的数据集作为源域，ARID 作为目标域。这样的选择突显了领域上下文辅助在将正常场景中训练得到的特征无监督迁移到黑暗场景中的显著能力。

在无源域适应和多源域适应的设定下，本研究先运用预训练模型生成目标域数据的高质量伪标签，再利用这些伪标签进行跨领域的无监督迁移学习。

4.2 全监督行为识别实验

本研究分别在黑暗场景和低分辨率场景下进行全监督行为识别实验。表 1 展示了 ARID 和 Tiny-VIRIT 数据集上的实验结果，并将领域上下文辅助与当前最先进的几个基于 CLIP 的行为识别方法进行了比较。由表 1 可知，领域上下文辅助在两个场景下均实现了最优性能。具体来说，在黑暗场景的 ARID 数据集上，本研究方法的 Top1 准确率为 71.86%，与之前最佳的 XCLIP[36]方法相比，提高了 4.54%。在低分辨率场景的 TinyVIRAT 数据集上，本研究方法的平均精确率为 80.93%，与之前最优的 Frozen CLIP[20]方法相比，提高了 1.60%。上述实验结果表明：领域上下文辅助在处理多模态预训练模型面临的域差异性大挑战时，可有效提升识别性能。

表 1 开放域行为识别的全监督结果

Table 1 Result for fully supervised on open domain action recognition

模型方法	ARID	TinyVIRAT
	Top1 准确率 (%)	平均精确率 (%)
ActionCLIP[6]	67.15	77.82
XCLIP[36]	67.32	78.80
Frozen CLIP[20]	66.22	79.33
领域上下文辅助	71.86	80.93

4.3 无源视频领域自适应行为识别实验

本文将领域上下文辅助方法与当前主流的一些无源视频领域自适应方法(如 SFDA[41]、

SHOT[42]、MA[43]和 ATCoN[25]等）进行了比较。此外，本文方法还与专门设计用于无监督域适应的方法进行了对比，如 DANN[44]、MK-MMD[45]和 TA^{3}N[46]等。为确保评估的公平性，本文还报告了使用相同的多模态预训练模型作为主干网络的 ActionCLIP 的结果。与领域上下文辅助类似，ActionCLIP 首先通过在源数据上的训练得到正常光照情况下的识别模型，随后使用目标域的无监督数据生成的伪标签进行迁移学习。如表 2 所示，领域上下文辅助在 3 个跨域基准上均取得了最佳性能，且明显优于之前的无源域适应方法，在具体的平均性能指标上，领域上下文辅助远优于之前最佳的方法。由此表明多模态预训练模型在领域自适应场景下的强大能力。此外，通过引入 LLM 扩展文本作为辅助，本研究的模型比使用相同编码器的 ActionCLIP 的性能好，进一步证明 LLM 扩展文本在帮助模型适应域差异较大的场景下进行迁移学习方面的有效性。

4.4　多源视频领域自适应行为识别实验

本研究对比了基于领域上下文辅助的开放域视频理解方法与先前采用的方法，包括基于领域对抗[44,50-54]（adversarial-based）和基于分布差异[43,55-61]（discrepancy-based）的方法。为保证评估的公平性，本文还展示了与本研究方法采用相同主干网络的 ActionCLIP 方法进行比较的结果。与本研究方法一致，ActionCLIP 模型先在源数据上进行训练，获得正常光照条件下的识别模型，再利用目标域的无监督数据生成的伪标签进行迁移学习。参照 Peng 等[57]的做法，本文报告了在相同网络配置下进行 5 次实验的 Top1 平均值和标准偏差。如表 3 所示，本文方法在所有对比方法中表现最佳，并以显著优势领先于先前的方法。值得注意的是，即便在使用相同的预训练模型的情况下，通过 LLM 扩展文本的辅助，本文方法实现了额外的 2.25% 的性能提升。上述结果不仅凸显了多模态预训练模型在面对领域自适应场景时的强大泛化能力，还突出了 LLM 扩展文本在此过程中的关键辅助作用。

4.5　消融实验

4.5.1　领域上下文辅助与主流 prompt 方法对比

基于领域上下文辅助的开放域视频理解方法的细节及其与当前主流提示学习方法的对比如表 4 所示。具体地，比较了以下几种策略：直接将文本标签作为文本编码器的输入；手动设计的

表 2　Daily-DA 无源视频领域适应的结果

Table 2　Result for source-free video domain adaptation on Daily-DA

方法	是否无源	Daily-DA 数据集实验准确率 (%)			
		K600→A11	MIT→A11	H51→A11	平均值
DANN[44]	×	21.18	22.81	14.20	19.40
MK-MMD[45]	×	21.66	21.02	20.35	21.01
TA^{3}N[46]	×	19.87	21.57	14.38	18.60
SFDA[41]	√	12.57	15.96	13.08	13.87
SHOT[42]	√	12.03	15.28	13.50	13.60
SHOT++[47]	√	12.57	14.90	15.98	14.48
MA[43]	√	12.76	17.75	12.90	14.47
BAIT[48]	√	12.69	16.93	13.65	14.42
CPGA[49]	√	13.06	18.08	13.14	14.76
ATCoN[25]	√	17.21	27.23	17.92	20.79
ActionCLIP[6]	√	47.89	48.59	45.20	47.22
领域上下文辅助	√	48.91	50.00	46.95	48.63

表 3 Daily-DA 多源视频领域自适应的结果

Table 3 Result for multi-source video domain adaptation on Daily-DA

方法	方法类别	Daily-DA 数据集实验平均准确率 (%)
		Daily→A11
s-DANN[44]	Adversarial-based	22.03±0.35
s-ADDA[50]	Adversarial-based	22.30±0.21
s-TA^{3}N[46]	Adversarial-based	21.76±0.16
s-ACAN[51]	Adversarial-based	23.44±0.16
c-DANN[44]	Adversarial-based	22.15±0.33
c-ADDA[50]	Adversarial-based	22.65±0.25
c- TA^{3}N[46]	Adversarial-based	22.24±0.20
c- ACAN[51]	Adversarial-based	23.95±0.28
MDAN[52]	Adversarial-based	23.75±0.38
DCTN[53]	Adversarial-based	24.94±0.36
MDDA[54]	Adversarial-based	22.73±0.26
s-MMD[43]	Discrepancy-based	21.62±0.22
s-MCD[55]	Discrepancy-based	23.80±0.28
s-CORAL[56]	Discrepancy-based	21.51±0.15
c-MMD[43]	Discrepancy-based	24.28±0.36
c-MCD[55]	Discrepancy-based	25.68±0.28
c-CORAL[56]	Discrepancy-based	23.96±0.16
MSDA[57]	Discrepancy-based	24.98±0.12
MCC[58]	Discrepancy-based	22.65±0.35
MOST[59]	Discrepancy-based	26.28±0.46
M3SDA[60]	Discrepancy-based	24.83±0.23
TAMAN[61]	Discrepancy-based	29.95±0.35
ActionCLIP[6]	—	52.11±0.99
领域上下文辅助	—	54.36±0.83

表 4 与其他提示方法的比较

Table 4 Comparison with other prompting methods

方法	Top1 准确率 (%)		
	ARID	SFVDA	MSVDA
w/o prompt	67.34	46.92	51.96
手工提示	67.15	47.22	52.11
CoOp[26]	68.21	47.56	52.42
领域上下文辅助	71.86	48.63	54.36

文本提示(在本文中是“A video of person doing {label}”)；自动化文本提示学习方法 CoOp。结果显示，此方法在多个方面超越了现有技术。在全监督的环境下，本方法在 ARID 数据集上取得了 3.65% 的准确率提升，在 SFVDA 设定下达到了 1.07% 的提升，并在 MSVDA 设定下实现了 1.94% 的提升。这些结果强调了利用 LLM 生成的细粒度领域特定文本提示对提高模型在开放域行为识别中的迁移学习能力和适应性的重要性。

4.5.2 文本在模型识别中的作用

本文对文本信息对模型性能的影响进行了评估。具体而言，比较了本文方法和一个替代方案：利用随机初始化的全连接层代替通过文本编码器生成的标签文本的语义特征。如表 5 所示，本文方法在无监督视频领域自适应(SFVDA)和多源视频领域自适应(MSVDA)的实验设置中分别带来了 4.88% 和 6.79% 的显著性能提升。这一发现突出显示了文本中的语义信息在无监督数据条件下对视频识别任务的重要贡献。

表 5 文本信息作用的消融实验

Table 5 Ablation experiment on the effect of the text information

方法	Top1 准确率 (%)		
	ARID	SFVDA	MSVDA
不用文本	67.56	43.75	47.57
使用文本	71.86	48.63	54.36

4.6 结果可视化

LLM 扩展文本在行为识别任务中的应用效果如图 2 所示。通过引入 LLM 扩展文本，模型可匹配更精细的文本信息，显著提高了行为识别的准确性。图 2 的预测结果明确表明，LLM 扩展文本可有效提升视频理解模型在细粒度层面的识别能力，特别是在面对复杂和模糊的行为场景时。例如，在图 2 中第一列第二行的视频样本中，模型原先将动作“一个人从一个容器向另一个容器转移物质”误识别为“拿(pick)”。整合 LLM 扩展文本后，模型在扩展文本“A video of a person transferring a substance from one container to another.”的辅助下成功地将该动作准确识别为

图 2　ARID 结果可视化

Fig. 2　ARID results visualization

“倒(pour)”。该文本描述是由大语言模型根据“倒(pour)”类别视频可能的行为生成的。这也验证了详细、具体的行为描述对提升模型识别特定视频内容的精确度具有重要作用。

5　讨论与分析

通过迁移多模态预训练模型理解开放域场景中的视频内容是现实世界中的一项重大挑战。现有研究主要针对学术数据集，而在处理开放域的复杂环境和质量较差的数据时，常常难以达到理想的性能。这一局限性的原因是显著的域差异和有监督数据的不足使得模型难以有效匹配视觉表征和简单的文本标签。为克服这些难题，本文提出一种新颖的方法，该方法通过大语言模型生成的富含领域上下文知识的动作标签增强视觉表征与人类行为的多层次描述之间的联系，从而促进鲁棒性分类。在全监督设置中，该方法在 ARID(黑暗场景)和 TinyVIRAT(低分辨率场景)数据集上分别获得了 71.86% 的 Top1 准确率和 80.93% 的平均精确率，比当前最先进的方法分别提高了 4.54% 和 1.60%。在无源视频领域自适应和多源视频领域自适应设置中，本文方法在 Daily-DA 数据集上的 Top1 准确率和平均精确率分别比使用相同主干网络的 ActionCLIP 方法提升了 1.41% 和 2.25%。与之前的最优方法相比，在上述两个场景下，都实现了平均精确率超过 20% 的显著进步。这一成果强调了利用大语言模型生成的细粒度、领域特定辅助文本在提高模型的迁移学习能力、适应新领域知识及减少域偏见方面的重要性。同时，这也展示了多模态预训练模型在领域自适应应用中的巨大潜力。本研究初步探索了利用大语言模型生成的辅助文本辅助多模态预训练模型迁移学习的潜力。尽管如此，本研究尚未完全挖掘文本与视觉信息间相互作用的全部潜力，目前仅限于使用生成文本辅助视觉识别，未来工作可进一步探索文本与视觉信息间的双向互补性。

6　结　论

本研究提出一种基于领域上下文辅助的开放域视频理解策略，目标是将多模态预训练模型有效地迁移到开放域视频内容的理解中。此方法利用了大语言模型内嵌的丰富语言知识，尤其是与视觉类别相关的知识。具体来说，本文采用类

别文本和领域特定的先验知识作为输入，生成针对每个标签的细粒度描述，并在此过程中融合领域上下文信息。在执行多模态图文匹配时，模型将视频内容与细粒度描述进行比较，而不是简单地匹配类别名称。结合 ChatGPT 和 CLIP 技术的使用，本文方法在多个开放域场景设定中取得了优异的表现，极大地提升了模型适应新领域的能力，减少了潜在偏差，并在视频识别任务中提高了性能。

参 考 文 献

[1] Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision [C] // Proceedings of the 38th International Conference on Machine Learning, 2021: 8748-8763.

[2] Yu JH, Wang ZR, Vasudevan V, et al. CoCa: contrastive captioners are image-text foundation models [Z/OL]. arXiv Preprint, arXiv: 2205.01917, 2022.

[3] Li JN, Li DX, Xiong CM, et al. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation [C] // Proceedings of the 39th International Conference on Machine Learning, 2022: 12888-12900.

[4] Luo HS, Ji L, Zhong M, et al. CLIP4Clip: an empirical study of CLIP for end to end video clip retrieval and captioning [J]. Neurocomputing, 2022, 508: 293-304.

[5] Tang MK, Wang ZY, Liu ZH, et al. CLIP4Caption: clip for video caption [C] // Proceedings of the 29th ACM International Conference on Multimedia, 2021: 4858-4862.

[6] Wang MM, Xing JZ, Liu Y. ActionCLIP: a new paradigm for video action recognition [Z/OL]. arXiv Preprint, arXiv: 2109.08472, 2021.

[7] Xu YC, Yang JF, Cao HZ, et al. ARID: a new dataset for recognizing action in the dark [C] // Proceedings of the International Workshop on Deep Learning for Human Activity Recognition, 2021: 70-84.

[8] Demir U, Rawat YS, Shah M. TinyVIRAT: low-resolution video action recognition [C] // Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), 2021: 7387-7394.

[9] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners [C] // Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020: 1877-1901.

[10] Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report [Z/OL]. arXiv Preprint, arXiv: 2303.08774, 2023.

[11] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks [C] // Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1725-1732.

[12] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks [C] // Proceedings of the IEEE International Conference on Computer Vision, 2015: 4489-4497.

[13] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6299-6308.

[14] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [J]. Advances in Neural Information Processing Systems, 2014, 27: 568-576.

[15] Wang LM, Xiong YJ, Wang Z, et al. Temporal segment networks for action recognition in videos [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(11): 2740-2755.

[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] // Proceedings of the 31st Conference on Neural Information Processing

Systems (NIPS 2017), 2017: 1-11.

[17] Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? [C] // Proceedings of the 38th International Conference on Machine Learning, 2021: 813-824.

[18] Liu Z, Ning J, Cao Y, et al. Video swin transformer [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 3202-3211.

[19] Fan HQ, Xiong B, Mangalam K, et al. Multiscale vision transformers [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 6824-6835.

[20] Lin ZY, Geng SJ, Zhang RR, et al. Frozen CLIP models are efficient video learners [C] // Proceedings of the European Conference on Computer Vision, 2022: 388-404.

[21] Li KC, Wang YL, He YN, et al. UniFormerV2: spatiotemporal learning by arming image vits with video uniformer [Z/OL]. arXiv Preprint, arXiv: 2211.09552, 2022.

[22] Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset [Z/OL]. arXiv Preprint, arXiv: 1705.06950, 2017.

[23] Goyal R, Ebrahimi Kahou S, Michalski V, et al. The “something something” video database for learning and evaluating visual common sense [C] // Proceedings of the IEEE International Conference on Computer Vision, 2017: 5842-5850.

[24] Chen R, Chen JJ, Liang ZX, et al. DarkLight networks for action recognition in the dark [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2021: 846-852.

[25] Xu YC, Yang JF, Cao HZ, et al. Source-free video domain adaptation by learning temporal consistency for action recognition [C] // Proceedings of the European Conference on Computer Vision, 2022: 147-164.

[26] Zhou KY, Yang JK, Loy CC, et al. Learning to prompt for vision-language models [J]. International Journal of Computer Vision, 2022, 130(9): 2337-2348.

[27] Zhou KY, Yang JK, Loy CC, et al. Conditional prompt learning for vision-language models [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 16816-16825.

[28] Zhang RR, Hu XF, Li BH, et al. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 15211-15222.

[29] Reddy DM, Basha SM, Hari MC, et al. DALL-E: creating images from text [J]. UGC Care Group I Journal, 2021, 8(14): 71-75.

[30] Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 9650-9660.

[31] Zhang RR, Fang RY, Zhang W, et al. Tip-Adapter: training-free clip-adapter for better vision-language modeling [Z/OL]. arXiv Preprint, arXiv: 2111.03930, 2021.

[32] Wu WH, Luo HP, Fang B, et al. Cap4Video: what can auxiliary captions do for text-video retrieval? [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 10704-10713.

[33] Li JN, Li DX, Savarese S, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models [C] // Proceedings of the 40th International Conference on Machine Learning, 2023: 19730-19742.

[34] Zhang RR, Han JM, Liu C, et al. LLaMA-Adapter: efficient fine-tuning of language models with zero-init attention [Z/OL]. arXiv Preprint, arXiv: 2303.16199, 2023.

[35] Touvron H, Lavril T, Izacard G, et al. LLaMA:

open and efficient foundation language models [Z/OL]. arXiv Preprint, arXiv: 2302.13971, 2023.

[36] Ni B, Peng H, Chen M, et al. Expanding language-image pretrained models for general video recognition [C] // European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 1-18.

[37] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners [J]. OpenAI Blog, 2019, 1(8): 9.

[38] Kuehne H, Jhuang H, Garrote E, et al. HMDB: a large video database for human motion recognition [C] // Proceedings of the 2011 International Conference on Computer Vision, 2011: 2556-2563.

[39] Monfort M, Andonian A, Zhou BL, et al. Moments in Time dataset: one million videos for event understanding [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(2): 502-508.

[40] Carreira J, Noland E, Banki-Horvath A, et al. A short note about kinetics-600 [Z/OL]. arXiv Preprint, arXiv: 1808.01340, 2018.

[41] Kim Y, Cho D, Han K, et al. Domain adaptation without source data [J]. IEEE Transactions on Artificial Intelligence, 2021, 2(6): 508-518.

[42] Liang J, Hu DP, Feng JS. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation [C] // Proceedings of the 37th International Conference on Machine Learning, 2020: 6028-6039.

[43] Li R, Jiao QF, Cao WM, et al. Model adaptation: unsupervised domain adaptation without source data [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 9641-9650.

[44] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation [C] // Proceedings of the 32nd International Conference on Machine Learning, 2015: 1180-1189.

[45] Long MS, Cao Y, Wang JM, et al. Learning transferable features with deep adaptation networks [C] // Proceedings of the 32nd International Conference on Machine Learning, 2015: 97-105.

[46] Chen MH, Kira Z, AlRegib G, et al. Temporal attentive alignment for large-scale video domain adaptation [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 6321-6330.

[47] Liang J, Hu DP, Wang YB, et al. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11): 8602-8617.

[48] Yang SQ, Wang YX, Van De Weijer J, et al. Unsupervised domain adaptation without source data by casting a BAIT [Z/OL]. arXiv Preprint, arXiv: 2010.12427, 2020.

[49] Qiu Z, Zhang YF, Lin HB, et al. Source-free domain adaptation via avatar prototype generation and adaptation [C] // Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021: 2921-2927.

[50] Tzeng E, Hoffman J, Saenko K, et al. Adversarial discriminative domain adaptation [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7167-7176.

[51] Xu YC, Cao HZ, Mao KZ, et al. Aligning correlation information for domain adaptation in action recognition [J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 35(5): 6767-6778.

[52] Scalbert M, Vakalopoulou M, Couzinié-Devy F. Multi-source domain adaptation via supervised contrastive learning and confident consistency regularization [Z/OL]. arXiv Preprint, arXiv: 2106.16093, 2021.

[53] Xu RJ, Chen ZL, Zuo WM, et al. Deep cocktail network: multi-source unsupervised domain adaptation with category shift [C] // Proceedings of the IEEE Conference on Computer Vision and

Pattern Recognition, 2018: 3964-3973.

[54] Zhao SC, Wang GZ, Zhang SH, et al. Multi-source distilling domain adaptation [C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12975-12983.

[55] Saito K, Watanabe K, Ushiku Y, et al. Maximum classifier discrepancy for unsupervised domain adaptation [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 3723-3732.

[56] Sun BC, Feng JS, Saenko K. Return of frustratingly easy domain adaptation [C] // Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016: 2058-2065.

[57] Peng XC, Bai QX, Xia XD, et al. Moment matching for multi-source domain adaptation [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1406-1415.

[58] Wang H, Xu MH, Ni BB, et al. Learning to combine: knowledge aggregation for multi-source domain adaptation [C] // Proceedings of the European Conference on Computer Vision, 2020: 727-744.

[59] Nguyen T, Le T, Zhao H, et al. Most: multi-source domain adaptation via optimal transport for student-teacher learning [C] // Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, PMLR, 2021: 225-235.

[60] Jin Y, Wang XM, Long MS, et al. Minimum class confusion for versatile domain adaptation [C] // Proceedings of the European Conference on Computer Vision, 2020: 464-480.

[61] Xu YC, Yang JF, Cao HZ, et al. Multi-source video domain adaptation with temporal attentive moment alignment network [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 3860-3871.