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：如何将预训练模型所获得的知识有效地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。受到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。本文提出了一种基于领域上下文辅助的开放域视频理解方法，通过大语言模型来深化模型对开放世界的理解。通过在大语言模型中融入领域的先验知识，在大语言模型中丰富动作标签的上下文知识，将视觉表示与人类行为的多层次描述进行对齐，实现了鲁棒的分类效果。我们在开放世界场景下进行了广泛的行为识别实验，在全监督设置中，ARID 数据集的预测准确度达到 71.86%，而 Tiny-VARIT 数据集在均值平均精度上取得了 80.93%。在无源领域自适应设置下，预测准确度实现了 48.63%；而多源领域自适应设置中，准确率为 54.36%，实验结果显示了领域上下文辅助在各种自适应环境下的有效性。

关键词:开放域行为识别；迁移学习；大语言模型；多模态

doi: 10.12146/j.issn.2095-3135.20231226001

Domain Context-Assisted for Open-World Action Recognition

Qinglin Xu[1,2],Yu Qiao[3], Yali Wang[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: Yali Wang E-mail：yl.wang@siat.ac.cn

Abstract: Effectively transferring knowledge from pre-trained models to downstream video understanding tasks is an important topic in computer vision research. Knowledge transfer becomes more challenging in open world due to poor data conditions. Many recent multimodal pre-training models are inspired by natural language processing and perform transfer learning by designing prompt learning. In this paper, we propose an LLM-powered domain context-assisted open-world action recognition method that leverages the open-world understanding capabilities of large language models. Our approach aligns visual representation with multi-level descriptions of human actions for robust classification, by enriching action labels with contextual knowledge in large language model. In the experiments of open-world action recognition with fully supervised setting, we obtain a Top-1 accuracy of 71.86% on the ARID dataset, and an mAP of 80.93% on the Tiny-VARIT dataset. More important, our method can achieve Top-1 accuracy of 48.63% in source-free video domain adaptation and 54.36% in multi-source video domain adaptation.

来稿日期：2023-12-26 修回日期：2024-03-07

基金项目：科技创新 2030——“新一代人工智能”重大项目（2022ZD0160505），国家自然科学基金资助项目（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，博士研究生导师，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，博士研究生导师，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。

Key words: Open-World Action Recognition; Transfer Learning; Large Language Model; Multi-Modality

Funding: This work was supported by the National Key R&D Program of China(NO.2022ZD0160505), the National Natural Science Foundation of China（Grant No. 62272450)

1 引 言

近年来，众多多模态视觉语言模型[1,2,3]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了卓越的迁移能力。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。已有众多研究成功探索如何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了显著成果。

尽管在常规数据集中这些方法[4,5,6]展现了优异性能，但在面对开放世界的复杂环境与苛刻数据条件时，它们往往表现不佳。在这种开放场景下，数据经常遭受诸如黑暗[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。同时在这个场景下，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，如何无监督地将模型适应至目标领域也成为另一关键挑战。针对以上问题，本研究提出了一个研究目标：开发一种新的迁移学习策略，旨在提升多模态预训练模型对开放世界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，常用的做法是利用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，在暗光环境下喝水类别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起到至关重要的结果[1]。尽管手工设计专门的文本提示相较于简单标签文本在一定程度上能够减轻领域差异，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这个问题，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]展示了其对开放世界的理解能力，本研究通过融入领域先验知识到 LLM，以丰富动作标签的上下文信息。具体来说，我们提出了领域上下文辅助方法，它利用大语言模型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。这样，模型能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，直接匹配视频与所有辅助文本并不适宜。因此我们借鉴了多实例学习（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。这种方法在多模态视频理解领域具备双重优势：首先，更为精确的描述能够更有效地与特定行为进行匹配。其次，这种方式可以为文本注入目标数据集的先验知识。例如，在黑暗环境中，由于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，通过采用更为具体具有领域信息的描述，如“使用手触摸不易辨认的物体”或“身体前倾推动物体”，模型将更容易准确地识别这些行为。我们在全监督、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。结果展示了在开放域的场景下，我们的方法相比较现有的多模态预训练模型迁移学习方法，达到先进的性能。尤其在视频领域自适应场景，我们发现了多模态预训练模型在视频领域自适应场景的潜力，相比较之前的方法，取得了显著的提升。

本文的主要贡献有：（1）本文创新性地结合了大语言模型和多模态与训练模型的知识，探索了其在开放域场景方面的应用潜力。（2）通过生成了具有领域上下文的多层次描

述文本，以更好地进行多模态对齐，实现鲁棒分类。（3）在全监督和领域自适应的基准下进行了大量的开放域行为识别实验，本文的方法明显优于目前最先进的方法，尤其在领域自适应场景，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，研究主要分为两个方向：发展 3D CNN[12,13]以捕获视频中的时间信息；以及双流网络的设计[14,15]，通过并行处理 RGB 图像和光流信息，获取空间和时间维度信息。Transformer[16]架构的出现之后涌现了许多基于此技术的创新工作[17,18,19]。基础模型的涌现，许多研究聚焦于如何将其迁移到视频相关的下游任务，如 FrozenCLIP[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往往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。Dark-light[24]研究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，获取标注数据往往需要大量的人力成本，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，视频领域自适应成为了研究的重点，如 ATCoN[25]通过学习不同尺度上的视图一致性来实现无监督迁移。然而，这些方法在面对具有明显域差异的开放世界视频时表现不足。

2.2 提示学习(Prompt Learning)

提示学习最初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[26]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。然而，该方法在处理未知类别时表现不佳。为此，CoCoOp[27]进一步在 CoOp 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。与基于训练的提示方法不同，我们的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（Foundation Model）在迁移学习中的应用

随着基础模型的不断发展，越来越多的研究集中在探索如何更有效地利用基础模型中的知识，以便有效地将其迁移到目标领域或应用到其他模型中。CaFo[28]专注于少样本学习领域，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]则充分利用了视觉描述生成模型[33]和 GPT-3[9]生成的辅助文本，以提升 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，就可以让 LLaMA[35]适用于不同场景。尽管如此，这些方法尚未充分挖掘视频领域自适应应用中基础模型的潜力。

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

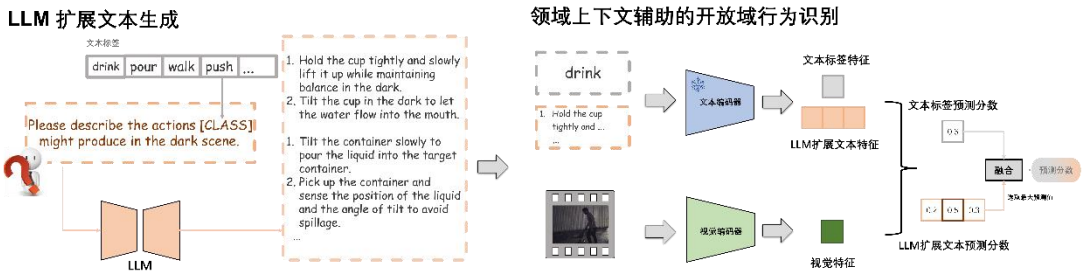

图 1 领域上下文辅助流程

Fig. 1 Domain context-assist process

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，我们提出了一种新颖的方法：利用大语言模型扩展现有的标签文本，从而丰富视频内容的语义理解，并向其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（例如黑暗场景）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

为了丰富文本语义信息以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[6]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。CoOp[26]通过可学习向量自动化文本提示的学习，旨在捕获数据集特有的领域信息。X-CLIP[36]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。尽管这些方法在减少对手动文本提示依赖上取得了进步，但对文本的语义丰富度仍有限。本研究利用大语言模型深度理解开放世界的能力，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，我们向 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。接下来，我们指导 ChatGPT[10]根据每个标签文本生成一系列详细描述，这些描述展示了目标领域数据集的特定上下文信息，以获得更全面和与上下文更紧密相关的理解。具体而言，对于每个标签文本 $T \in \mathcal{T}$ ，我们利用 ChatGPT 生成 K 个扩展文本，表示为 $LLM(T)=\{e_1,e_2,...,e_k\}$ ，其中 e_k 表示第 k 个扩展文本， $\mathcal{T}$ 表示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，形成了所谓的 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类别的全部 LLM 扩展文本严格对齐。因此我们借鉴多实例学习的策略，视频内容应与能够最准确地描述其核心特征的 LLM 扩展文本进行匹配。

例如如果类别是“拿取行为”（Pick）的扩展文本就包括“有人举手触摸物体的视频”、“有人低头在桌子上找东西并拿取的视频”。当具体某个视频展示了一个人举起手在寻找高处的物体，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，但是我们更倾向于将视频与“有人举手触摸物体的视频”的扩展文本对齐，因为他与视频的行为更为对应。

训练的过程中，通过视觉和文本编码器，我们以视觉、标签文本和 LLM 扩展文本作为输入，可以得到视觉特征： $\{v_n \in \mathbb{R}^d \mid n=1,2,...,N\}$ ，文本特征： $\{t_m \in \mathbb{R}^d \mid m=1,2,...,M\}$ 和

LLM 扩展文本特征：$\{e_{mk} \in \mathbb{R}^d \mid m = 1,2,...,M，\mathrm{k} = 1,2,...,K\}$。$N$ 是取样视频的数量、M 是类别数，K 是每个类别的扩展文本数。值得指出的是，标签文本和 LLM 扩展文本都是使用相同的文本编码器进行编码的。首先我们通过计算视频与标签文本之间的余弦相似度来评估它们的相关性，得到每个视频与所有类别的相似度：

$$S_{\mathrm{v}_n\mathrm{t}_m} = sim\left(\mathrm{v}_n, t_m\right), \tag{1}$$

其中 $S_{\mathrm{v}_n\mathrm{t}_m}$ 表示视频 v_n 和文本标签 t_n 的相似度，sim 为余弦相似度函数。同样的，我们计算视频与 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地描述其核心特征的 LLM 扩展文本进行匹配，在每个类别的 LLM 扩展文本中，我们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：

$$S_{\mathrm{v}_n\mathrm{e}_m} = \arg\max_k sim\left(\mathrm{v}_n, e_{mk}\right), \tag{2}$$

其中 $S_{\mathrm{v}_n\mathrm{e}_m}$ 表示视频 v_n 和 LLM 扩展文本 e_m 中所有描述的最大相似度，e_m 代表 m 类别的所有扩展文本。在训练和推理过程中，我们采用以下融合方程将这两个相似度分支结合起来：

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

其中 α 为预测结果融合的权值系数。最后,我们的目标是如果视频 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:

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

其中 $\phi(n)$ 表示第 n 个视频所对应的类别。

4 结 果

在本章节将详细介绍了我们针对开放域场景下的实验设置及其结果。由于有监督训练在这些场景中往往难以取得优异成绩，我们选择了全监督设定下的黑暗和低分辨率的数据集，即 ARID[7]和 Tiny-VIRIT[8]，进行了验证实验。这些实验旨在证明我们的方法有效应对显著的域差异挑战。在领域自适应方面，我们选用了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，我们通过消融实验深入验证了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，我们利用可视化展示了 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的卓越迁移能力，我们选择其作为主干网络。具体而言，我们采用了 CLIP[1] 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，我们直接应用了 CLIP[1]中的文本编码器来生成标签和描述的特征表示，这是一个包含 12 层、维度为 512、拥有 8 个注意力头的 Transformer[16]模型。与 CLIP[1]的设计保持一致，选取了“[CLS] token”对应的特征作为整个语句的特征表示[40]。公式（3）中的 α 我们采用其值为 0.2。

我们的方法在四个基准数据集上进行了严格的评估：ARID[7]、Tiny-VIRIT[8]、SFVDA Daily-DA[25]和 MSVDA Daily-DA[25]。ARID[7]数据集专注于黑暗场景下的行为识别，包含

3784 个视频片段，覆盖 11 个不同的动作类别。TinyVIRAT[8]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

模型方法	准确率(%)
	Top1
ActionCLIP[6]	67.15
XCLIP[36]	67.32
FrozenCLIP[20]	66.22
Ours	**71.86**

模型方法	平均精确率
	mAP
ActionCLIP[6]	77.82
XCLIP[36]	78.80
FrozenCLIP[20]	79.33
Ours	**80.93**

表 3　无源领域自适应场景下 Daily-DA 数据集实验结果　　果

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

方法	是否无源	Daily-DA			
		K600→A11	**MIT→A11**	**H51→A11**	**Avg**
DANN[42]	×	21.18	22.81	14.20	19.40
MK-MMD[43]	×	21.66	21.02	20.35	21.01
TA^{3}N[44]	×	19.87	21.57	14.38	18.60
SFDA[45]	√	12.57	15.96	13.08	13.87
SHOT[46]	√	12.03	15.28	13.50	13.60
SHOT++[47]	√	12.57	14.90	15.98	14.48
MA[48]	√	12.76	17.75	12.90	14.47
BAIT[49]	√	12.69	16.93	13.65	14.42
CPGA[50]	√	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
Ours	√	**48.91**	**50.00**	**46.95**	**48.63**

方法	方法类别	Daily-DA
		Daily→A11
s-DANN[42]	Adversarial-	22.03±0.35

性的特点，涵盖 26 个动作类别和 20258 个视频片段。SFVDA Daily-DA[25]是一个无源域适应的基准，包括四个数据集：正常光照条件下的 HMDB51[37]（H51）、Moments-in-Time[38]（MIT）和 Kinetics[39]（K600），以及黑暗场景的 ARID[7]（A11）。在这个设定下，模型需要选择一个数据集作为源域，另一个作为目标域。我们选择了三个以 ARID[7]为目标域的跨领域任务，以突出我们方法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[25]数据集与 SFVDA 类似，但在多源视频域适应的设定中，我们将三个正常光照条件下的数据集作为源域，ARID[7]作为目标域。这样的选择突显了我们的模型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，本研究首先运用预训练模型生成目标域数据的高质量伪标签。随后，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

在本研究中，我们在黑暗场景和低分辨率场景下进行了全监督行为识别实验。在表 1

s-ADDA[51]	based	22.30±0.21
s-TA^{3}N[44]		21.76±0.16
s-ACAN[52]		23.44±0.16
c-DANN[42]		22.15±0.33
c-ADDA[51]		22.65±0.25
c- TA^{3}N[44]		22.24±0.20
c- ACAN[52]		23.95±0.28
MDAN[53]		23.75±0.38
DCTN[54]		24.94±0.36
MDDA[55]		22.73±0.26
s-MMD[56]	Discrepancy-based	21.62±0.22
s-MCD[57]		23.80±0.28
s-CORAL[58]		21.51±0.15
c-MMD[56]		24.28±0.36
c-MCD[57]		25.68±0.28
c-CORAL[58]		23.96±0.16
MSDA[59]		24.98±0.12
MCC[60]		22.65±0.35
MOST[61]		26.28±0.46
M3SDA[62]		24.83±0.23
TAMAN[63]		29.95±0.35
ActionCLIP[6]	-	52.11±0.99
Ours		**54.36**±0.83

表 4　多源领域自适应下 Daily-DA 实验结果

Table 4　Result for MSVDA on Daily-DA

和表 2 中，我们分别展示了在 ARID[7]和 Tiny-VIRIT[8]数据集上的实验结果，并对我

方法	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
Ours	**71.86**	**48.63**	**54.36**

们的方法与当前最先进的几个基于 CLIP[1]的行为识别方法进行了比较。从结果中可以观察到，我们的模型在两个场景下均实

方法	ARID	SFVDA	MSVDA
不用文本	67.56	43.75	47.57
使用文本	**71.86**	**48.63**	**54.36**

现了最优性能。具体来说，在黑暗场景的 ARID[7]数据集上，我们的方法取得了 Top-1 71.86%的识别准确率，相比之前最佳的

表 6 文本信息作用的对比实验

Table6　Ablation study on the effect of the text ；

XCLIP[36]方法提高了 4.54%。在低分辨率场景的 Tiny-VIRIT[8]数据集上，我们的方法在平均精确率（mAP）指标上达到了 80.93%，比最优的 FrozenCLIP[20]方法提高了 1.60 个百分点。这些实验结果充分证明了我们的方法在处理多模态预训练模型面临的域差异性大挑战时，能够有效提升识别性能。

4.3 无源视频领域自适应行为识别实验

我们将我们的方法与当前最先进的一些无源视频领域自适应方法进行了比较，这些方法包括 SFDA[45]、SHOT[46]、MA[48]和 ATCoN[25]等。此外，我们还将其与专门设计用于无监督域适应的方法进行了对比，如 DANN[43]、MK-MMD[43]和 TA^{3}N[44]等。为了确保评估的公平性，我们还报告了使用与我们相同的多模态预训练模型作为主干网络的 ActionCLIP[6]的结果。与我们的方法类似，ActionCLIP[6]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。在表 3 中展示了 Top-1 准确率的结果。结果显示，我们的方法在三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，我们的方法远远优于之前最佳的方法。这一成果充分展示了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 扩展文本的辅助，我们的模型相比于使用相同编码器的 ActionCLIP[6]实现了性能提升，这进一步证明了 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

在本研究中，我们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

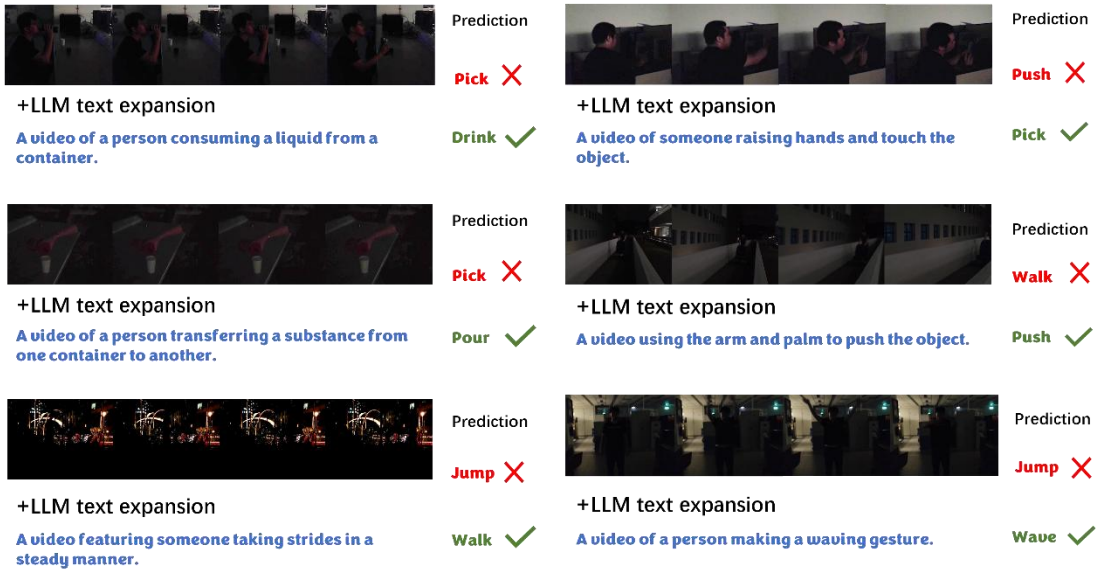

图 2 ARID 结果可视化

Fig. 2 ARID results visualization

法，包括基于领域对抗[42,51-55]（adversarial-based）和基于分布差异[56-63]（discrepancy-based）的方法。为了保证评估的公平性，我们还展示了使用与我们方法相同的多模态预训练模型 ActionCLIP[6]作为主干网络的结果。与我们的方法一致，ActionCLIP 模型首先在源数据上进行训练，以获得正常光照条件下的识别模型，随后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 等人（2019）[59]的做法，我们报告了在相同网络配置下进行 5 次实验的 Top-1 平均值和标准偏差。表 4 中的结果显示，我们的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，我们的方法实现了额外的+2.25%的性能提升。这些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

表 5 展示了我们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。具体地，我们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；手动设计的文本提示（在本文中是“A video of person doing {label}”）；以及自动化文本提示学习方法 CoOp。结果显示，我们的方法在多个方面超越了现有技术。在全监督的环境下，我们的方法在 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大语言模型(LLM)生成的细粒度领域特定文本提示对于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。

4.5.3 文本在模型识别中的作用

我们对文本信息对模型性能的影响进行了评估。具体而言，我们比较了我们的方法和一个替代方案：使用随机初始化的全连接层来代替通过文本编码器生成的标签文本的语义特征。如表 6 所展示的，我们的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。

4.6 结果可视化

图 2 展示了 LLM 扩展文本在行为识别任务中的应用效果。通过引入 LLM 扩展文本，模型能够匹配更精细的文本信息，这显著提高了行为识别的准确性。这些结果明确表明，LLM 扩展文本可以有效地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，模型成功地将该动作在扩展文本“A video of a person transferring a substance from one container to another.”的辅助下准确识别为“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、具体的行为描述对于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

迁移多模态预训练模型以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，而在处理开放世界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。这一局限性部分源于显著的域差异和有监督数据的不足，使得模型难以有效地匹配视觉表征和简单的文本标签。为了克服这些难题，我们提出了一种新颖的方法，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，来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。采用我们的方法后，在全监督设置中，我们在 ARID[7]（黑暗场景）和 Tiny-VARIT[8]（低分辨率场景）数据集上分别获得了 71.86%和 80.93%的准确率，相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，我们的方法在 Daily-DA 数据集上分别比使用相同主干网络的 ActionCLIP[6]方法提升了 1.41%和 2.25%。与之前的最优方法相比，我们在这两个场景下都实现了超过 20 个百分点的显著进步。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辅助文本来辅助多模态预训练模型迁移学习的潜力，尽管如此，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来辅助视觉识别，未来工作可以进一步探索这两种模态间的双向互补性。

6 结 论

本研究提出了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，我们采用了类别文本和领域特定的先验知识作为输入，生成了针对每个标签的细粒度描述，并在此过程中融合了领域上下文信息。在执行多模态图文匹配时，模型被设计为将视频内容与这些细化的描述符进行比较，而不是简单地匹配类别名称。结合了 ChatGPT[10]和 CLIP[1]技术的使用，我们的方法在多个开放域场景设定中取得了优异的表现。这种方法极大地提升了模型适应新领域的能力，减少了潜在的偏差，并且在视频识别任务中提高了性能。

参考文献

[1] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR 2021: 8748-8763.

[2] Yu J, Wang Z, Vasudevan V, et al. Coca: Contrastive captioners are image-text foundation models[J]. arXiv preprint arXiv:2205.01917, 2022.

[3] Li J, Li D, Xiong C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International Conference on Machine Learning. PMLR, 2022: 12888-12900.

[4] Luo H, 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 M, Wang Z, Liu Z, et al. Clip4caption: Clip for video caption[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 4858-4862.

[6] Wang M, Xing J, Liu Y. Actionclip: A new paradigm for video action recognition[J]. arXiv preprint arXiv:2109.08472, 2021.

[7] Xu Y, Yang J, Cao H, et al. Arid: A new dataset for recognizing action in the dark[C]//Deep Learning for Human Activity Recognition: Second International Workshop, DL-HAR 2020, Held in Conjunction with IJCAI-PRICAI 2020, Kyoto, Japan, January 8, 2021, Proceedings 2. Springer Singapore, 2021: 70-84.

[8] Demir U, Rawat Y S, Shah M. Tinyvirat: Low-resolution video action recognition[C]//2020 25th international conference on pattern recognition (ICPR). IEEE, 2021: 7387-7394.

[9] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[10] Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023. , 2: 3.

[11] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks[C]//Proceedings of the 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.

[15] Wang L, Xiong Y, 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] Transformer Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[17] Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding?[C]// Proceedings of 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 H, 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 Z, Geng S, Zhang R, et al. Frozen clip models are efficient video learners[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 388-404.

[21] Li K, Wang Y, He Y, et al. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer[J]. arXiv preprint arXiv:2211.09552, 2022.

[22] Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset[J]. 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 J, Liang Z, et al. Darklight networks for action recognition in the dark[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 846-852.

[25] Xu Y, Yang J, Cao H, et al. Source-free video domain adaptation by learning temporal consistency for action recognition[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 147-164.

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

[27] Zhou K, Yang J, Loy C C, 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 R, Hu X, Li B, 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 M D M, Basha M S M, Hari M M C, 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 R, Fang R, Zhang W, et al. Tip-adapter: Training-free clip-adapter for better vision-language modeling[J]. arXiv preprint arXiv:2111.03930, 2021.

[32] Wu W, Luo H, 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 J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//International conference on machine learning. PMLR, 2023: 19730-19742..

[34] Zhang R, Han J, Zhou A, et al. Llama-adapter: Efficient fine-tuning of language models with zero-init attention[J]. arXiv preprint arXiv:2303.16199, 2023.

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

[36] Kuehne H, Jhuang H, Garrote E, et al. HMDB: a large video database for human motion recognition[C]//2011 International conference on computer vision. IEEE, 2011: 2556-2563.

[37] Monfort M, Andonian A, Zhou B, 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.

[38] Carreira J, Noland E, Banki-Horvath A, et al. A short note about kinetics-600[J]. arv preprint arv:1808.01340, 2018.

[39] 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.

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

[41] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation[C]//International conference on machine learning. PMLR, 2015: 1180-1189.

[42] Long M, Cao Y, Wang J, et al. Learning transferable features with deep adaptation networks[C]//International conference on machine learning. PMLR, 2015: 97-105.

[43] Chen M H, 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.

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

[45] Liang J, Hu D, Feng J. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation[C]//International conference on machine learning. PMLR, 2020: 6028-6039.

[46] Liang J, Hu D, Wang Y, 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.

[47] Yang S, Wang Y, Van De Weijer J, et al. Unsupervised domain adaptation without source data by casting a bait[J]. arv preprint arv:2010.12427, 2020, 1(2): 5.

[48] Agarwal P, Paudel D P, Zaech J N, et al. Unsupervised robust domain adaptation without source data[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022: 2009-2018.

[49] Qiu Z, Zhang Y, Lin H, et al. Source-free domain adaptation via avatar prototype generation and adaptation[C]//International Joint Conferences on Artificial Intelligence Organization,.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 Y, Cao H, Mao K, et al. Aligning correlation information for domain adaptation in action recognition[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022.

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

[53] Xu R, Chen Z, Zuo W, 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 S, Wang G, Zhang S, et al. Multi-source distilling domain adaptation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 12975-12983.

[55] Long M, Cao Y, Wang J, et al. Learning transferable features with deep adaptation networks[C]//International conference on machine learning. PMLR, 2015: 97-105.

[56] 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.

[57] Sun B, Feng J, Saenko K. Return of frustratingly easy domain adaptation[C]//Proceedings of the AAAI conference on artificial intelligence. 2016, 30(1).

[58] Peng X, Bai Q, Xia X, et al. Moment matching for multi-source domain adaptation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1406-1415.

[59] Wang H, Xu M, Ni B, et al. Learning to Combine: Knowledge Aggregation for Multi-source Domain Adaptation[C]//European Conference on Computer Vision. 2020: 727-744.

[60] Nguyen T, Le T, Zhao H, et al. Most: Multi-source domain adaptation via optimal transport for student-teacher learning [J]. Proceedings of Machine Learning Research, 2021, 161: 225-235.

[61] Jin Y, Wang X, Long M, et al. Minimum Class Confusion for Versatile Domain Adaptation[C]//European Conference on Computer Vision. 2020: 464-480.

[62] Xu Y, Yang J, Cao H, et al. Multi-source video domain adaptation with temporal attentive moment alignment network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023.