面向少样本视频动作识别的动态多专家融合框架

胡 贝,欧阳君,王 晨

(武汉纺织大学 计算机与人工智能学院,湖北 武汉 430200)

摘 要: 多模态视频分析是理解复杂动态场景与人类行为的重要途径,但现有视觉语言模型常采用单一固化的时序建模范式,难以自适应地处理视频中多样的动态模式(静态背景、短程动作与长程周期性行为等),制约了视频分析深度与泛化能力。为此,提出一种动态多专家融合框架DMEF。首先,设计一个自适应专家调度器,根据输入视频的内容特征,动态选择并融合3种不同粒度的时序建模专家:全局静态专家用于维持静态语义泛化能力,短程动态专家专注捕捉局部帧间变化,长程上下文专家建模全局时序依赖关系;其次,建立一个可随视频内容自适应调节的智能建模架构,让视频内容分析的解释性得到提升。在少样本学习与跨域泛化任务上,DMEF方法在HMDB51、UCF101数据集上的性能最优;在SSv2数据集上,当K=8时性能最优。在跨域泛化任务中,DMEF方法在HMDB51数据集中的识别准确率优秀。

关键词: 视频动作识别;少样本学习;跨域泛化;视觉语言模型;时序建模;多模态视频分析

DOI: 10. 11907/rjdk. 251827 扫描二维码阅读全文:

中图分类号: TP391.4 **文献标识码:** A **文章编号:** 1672-7800(XXXX)0XX-0001-10

Dynamic Multi-Expert Fusion Framework for Few-Shot Video Action Recognition

HU Bei, OUYANG jun, WANG Chen

(*School of Computer Science and Artificial Intelligence*, *Wuhan Textile University*, *Wuhan* 430200, *China*)

Abstract: Multimodal video analysis serves as a crucial approach for understanding complex dynamic scenes and human behavior. However, existing vision-language models often rely on a single, rigid temporal modeling paradigm, which struggles to adaptively handle diverse dynamic patterns in videos (such as static backgrounds, short-range actions, and long-range periodic behaviors), thereby limiting the depth and generalization capabilities of video analysis. To address this, a Dynamic Multi-Expert Fusion (DMEF) framework is proposed. First, an adaptive expert scheduler is designed to dynamically select and integrate three different granularities of temporal modeling experts based on the content features of the input video: a global static expert maintains static semantic generalization, a short-range dynamic expert focuses on capturing local frame-to-frame changes, and a long-range contextual expert models global temporal dependencies. Second, an intelligent modeling architecture is established that can adaptively adjust according to video content, enhancing the interpretability of video content analysis. In few-shot learning and cross-domain generalization tasks, the DMEF method achieves optimal performance on the HMDB51 and UCF101 datasets; on the SSv2 dataset, it performs best when K=8. In cross-domain generalization tasks, the DMEF method demonstrates excellent recognition accuracy on the HMDB51 dataset.

Key Words: video action recognition; few-shot learning; cross-domain generalization; vision-language model; temporal modeling; multimodal video analysis

0 引言

随着短视频、直播与多模态内容生态的爆发,全球视频数据正以每秒PB级规模呈指数级增长,对人类分析和处理数据洪流的能力提出了严峻挑战。多模态视频分析被视为应对这一挑战的关键,能将海量视频转化为可交互的视觉摘要,提升决策效率。然而,此类模型的效能瓶颈

收稿日期: 2025-12-17

基金项目: 国家自然科学基金项目(62301373);湖北省自然科学基金项目(2023AFB294,2023AFB279)

作者简介: 胡贝(1999-),女,武汉纺织大学计算机与人工智能学院硕士研究生,研究方向为视频动作识别;欧阳君(1994-),女,博士,武汉纺织大学计算机与人工智能学院讲师、硕士生导师,研究方向为计算机视觉与模式识别;王晨(1993),女,博士,武汉纺织大学计算机与人工智能学院副教授、硕士生导师,研究方向为图像处理、机器视觉、模式识别。本文通讯作者:欧阳君、王晨。

在于:难以从时序上精准理解人类行为,且“黑箱”特性也存在可信度与可解释性问题,即模型不仅要准,更要“说得清”。

在此背景下,视频动作识别技术作为连接原始视频与高层语义分析的核心桥梁,其发展水平直接决定其在安防、交通、体育、医疗等关键领域的应用上限[1]。尽管,基于深度学习的方法已取得显著进展,但性能严重依赖大规模精细标注数据,但在危险事件监控、罕见动作识别等实际场景中,获取足量标注数据的成本高昂甚至不可行。

为此,小样本视频动作识别研究应运而生,仅凭少量标注样本便可识别未知类别动作,旨在从极少量样本中准确捕捉动作的时序演变规律。因此,开发一个能精准建模时序演变且推理清晰的小样本动作识别模型,对构建下一代可信、高效的智能系统具有重要意义。

1 相关工作

1.1 面向小样本学习的视觉语言模型

目前,视觉语言模型在超大规模图像—文本对上进行预训练,以得到一个共享的、语义丰富的多模态表示空间。例如,CLIP模型具有出强大的零样本泛化能力,无需下游任务标注,仅通过计算图像与文本特征的相似度即可完成分类[2,3]。该模型推动了一系列基于提示学习和适配器的高效迁移方法的发展,在图像分割、目标检测等任务上取得了显著成功[4-9]。

后续,研究者为了在时序动态建模与小样本泛化能力之间取得平衡,尝试将CLIP模型适配于视频任务。现有方法大致可分为以下4类。

(1)逐帧编码后简单聚合。该方法主要解决从帧级特征到视频级表示的转换问题。例如,ActionCLIP模型通过设计多模态提示模板和轻量的时间注意力机制,将CLIP成功适配于视频动作分类任务,取得了接近全监督方法的性能,但时序建模能力相对较弱,对动作的精细时序顺序不够敏感[10]。

(2)引入额外时序模块。为了主动建模时序动态,在CLIP架构基础上引入额外的复杂模块。CLIP4Clip模型系统探索了多种视频帧编码器,包括序列类型和紧密类型。前者使用LSTM对帧序列进行编码;后者将文本与所有帧特征一同输入Transformer进行深度融合,用于视频—文本检索任务,显著提升了跨模态检索的精度[11-13]。然而,引入的复杂时序模块增加了模型参数量,在小样本场景下可会能引发过拟合问题,并一定程度上削弱了CLIP原有的零样本泛化优势。TimeSformer模型通过时空分离注意力机制建模视频时序关系,显著提升了模型对长程动作模式的理解,但参数量和计算成本较高,可能影响实际部署效率[14]。

(3)基于轻量级时序适配器或提示学习的方法。该方法是在模型复杂度和性能之间的折中方案,旨在通过调整参数高效地对CLIP进行时序适配。例如,ViFi-CLIP模型通过微调策略实现性能提升,验证了轻量适配的有效性,但时序建模能力有限[15]。文献[16]探索了视觉语言模型的提示学习机制在视频理解中的应用,通过设计可学习的时序提示向量,以极低的参数成本使模型适应视频任务,显著提升了训练效率与小样本性能。然而,这类方法对提示设计较为敏感,对复杂、长程时序动态的建模能力仍有待加强。

(4)高级聚合与知识集成方法。该方法致力于设计更精细的聚合策略或引入外部知识。文献[17]提出一种自适应分层图推理网络,通过构建一个能根据语义陈述自适应调整结构的3层图网络,并引入语义连贯性学习联合推理视频与文本,以提升视频—语言推理任务的性能。然而,此类方法通常依赖手工构建的图结构或外部知识库,性能与泛化能力受限于知识构建的质量与完备性。SkeletonCLIP++模型采用加权帧集成方式更细致地表征人体运动,可提升对运动细节的建模能力,但对骨架数据质量的依赖较大[18]。

尽管,研究者针对CLIP模型适配视频时序建模提出了多种方案,但普遍存在以下局限性:致力于寻找并固化一种“唯一最优”的时序建模范式,如图1所示。然而,视频中的动作本身具有高度多样性:有的依赖静态外观,例如“坐下的”;有的强调短程动态,例如“击掌”;有的呈现长程周期模式,例如“走路”。固化的模型难以自适应地应对这种变化,导致其在复杂的小样本场景下能力不足或发生过拟合。为此,本文构建一个能根据输入视频内容,动态调度与融合多种基础时序建模能力的自适应方法。

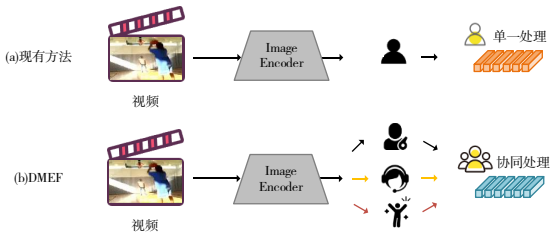


Fig. 1 Difference among DMEF and existing methods in time series modeling

图1 在时序建模上DMEF与现有方法的区别

1.2 多专家机制

混合专家(Mixture-of-Experts, MoE)模型通过动态路由机制稀疏地激活部分专家网络,已成为扩展模型参数规模、构建万亿参数大模型的关键[19,20]。尽管,这种以规模为导向的经典MoE范式,其目标与参数敏感的小样本视频动作识别需求存在根本差异,但通过专业化分工还是能提升模型能力。

近期,研究者不再追求参数规模的无限扩展,而为特定任务设计少量功能明确、互补的轻量级专家,实现稠密或软性融合机制的协同工作。这种“小而专”的新范式,既

保留了MoE的思想优势，又规避了经典结构在小样本场景下的冗余与冲突。例如，Immortal方法为3D目标检测设计了针对不同传感器的专家网络，有效融合了多源异构数据，显著提升了在极端环境下的鲁棒性，但模型结构较为复杂，对传感器同步与标定精度的依赖较高[21]。TMoE方法在视觉目标追踪中，利用多个专家分别建模不同类型的图像块间关系，通过动态融合提升了追踪的准确性，但未显式建模长期遮挡或目标ID切换，专家冗余度缺乏正则约束[22]。METransformer方法在编码器和解码器中，引入多个可学习的专家token来指导医学影像报告生成，增强了报告的专业性与多样性，但专家token数量会随报告长度呈线性增加，导致推理显存占用高于同规模基线[23]。MTE方法从单一预训练模型中蒸馏出多个辅助token以构建轻量专家，有效增强了自监督视觉表示学习的效果，但性能提升在一定程度上依赖原始大模型[24]。

在视频动作识别中，如何将“小规模稠密多专家”范式思想与时序建模这一核心挑战相结合，仍是一个有待探索的问题。现有工作尚未系统性地探索如何显式构建专家来捕捉视频中全局趋势、局部动态与关键语义等不同粒度的时序模式。基于这一现状，该方面的研究旨在实现一个范式转换：从“使用一个复杂模型处理所有时序”到“为不同类型的时序模式配备专属专家”。

为此，本文提出动态多专家融合框架（Dynamic Multi-Expert Fusion Framework，DMEF），一个全局静态专家（Global Static Expert，GSE）、一个短程动态专家（Short-term Dynamic Expert，SDE）与一个长程上下文专家（Long-range Context Expert，LCE）进行处理，核心创新并非专家本身，而在于自适应专家调度器。该机制能动态评估输入视频的特性，以此为3类基础专家分配权重，从而在面对多样化的动作类别时，具备更强大的泛化能力和鲁棒性。主要贡献为：①提出一种动态多专家融合框架，基于CLIP模型，根据输入视频内容自适应选择、融合最相关的时序建模方案，实现从“时序单一处理”到“分时序动态理解”的策略转变；②设计一个基于门控网络的自适应专家调度器，能动态协调并融合全局静态、短程动态与长程上下文等3类时序专家的贡献权重。

2　本文方法

2.1　方法介绍

DMEF的核心思想是为视频中不同复杂度的时序模式匹配专用的建模专家，并通过自适应专家调度器进行融合，如图2所示。首先，给定一个输入视频，利用预训练的CLIP图像编码器提取每一帧的视觉特征；其次，该帧级特征序列被并行送入3个专家模块进行协同工作：全局静态专家提供稳定、整体的场景视图，短程动态专家分析帧间细粒度的连续运动，长程上下文专家识别并聚焦整个序列中最关键的瞬间；再次，通过自适应专家调度器分析整个序列的全局内容，为每位专家生成一个融合权重；最后，视频的表征由各专家输出的加权和构成，并送入分类头进行预测。

2.2　全局静态专家

基于CLIP等视觉语言模型的方法，优势在于从大规模图像数据中学习强大的静态视觉概念先验，但复杂的时序建模可能会扭曲静态特征。为此，设计全局静态专家GSE，将其作为一个稳定的“锚点”，守护并提炼CLIP模型的静态语义理解能力，确保模型在面对时序变化时不会丢失最基本的场景识别基础。具体地，需要一个机制来评估每一帧作为独立静态图像的语义完整性与代表性。简单的平均池化会赋予所有帧同等权重，显然不够理想，原因为模糊、遮挡或信息量低的帧应该被抑制。因此，采用一个两层多层感知机（Multilayer Perceptron，MLP）$f(\cdot)$为每一帧特征X_t计算一个标量重要性分数S_t。

$$S_t = f(X_t) = \boldsymbol{W}_2^{\top}\sigma(\boldsymbol{W}_1 X_t + \boldsymbol{b}_1) + \boldsymbol{b}_2 \tag{1}$$

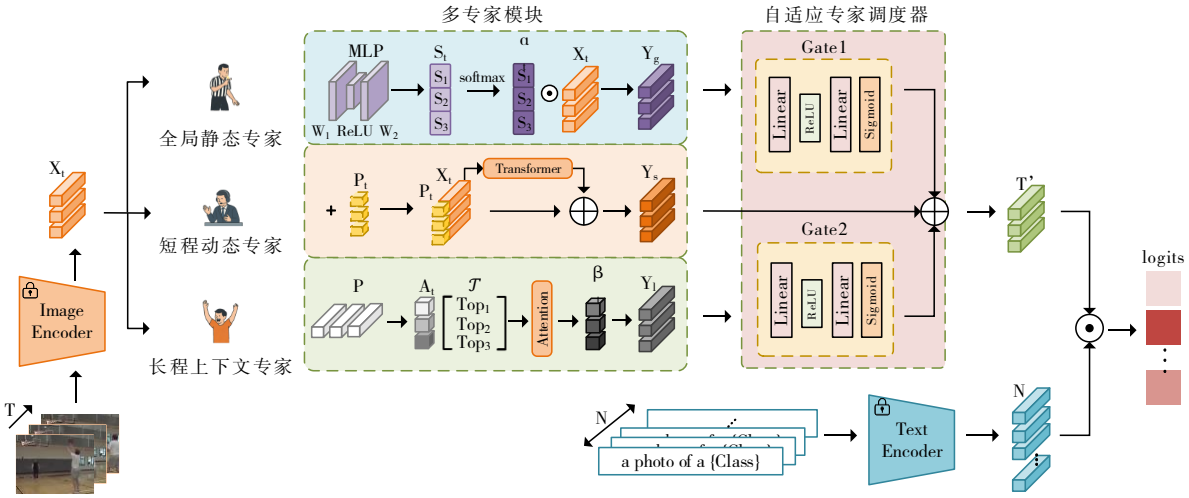


Fig. 2　Dynamic multi expert fusion framework

图2　动态多专家融合框架

式中：$\boldsymbol{W}_1 \in \boldsymbol{R}^{D/2 \times D}$、$\boldsymbol{b}_1 \in \boldsymbol{R}^{D/2}$为第一层全连接层的权重和偏置；$\sigma(\cdot)$表示激活函数ReLU；$\boldsymbol{W}_2 \in \boldsymbol{R}^{D/2}$、$\boldsymbol{b}_2 \in R$为第二层全连接层权重和偏置；$S_t$为该时间步的未归一化注意力分数。

本文选择双层多层感知机的原因为：在理论表达与参数效率方面，双层结构更高效。虽然，通用近似定理指出仅含有一层隐藏层的前馈网络就足以有效表示任何连续函数，但此类“浅而宽”的网络需要指数级神经元，存在参数爆炸、易过拟合、难以泛化等风险[25]。因此，适当增加网络深度可在保持甚至提升表达能力的同时，显著压缩参数量。综上，本文采用双层MLP，在深度上延伸，在紧凑假设空间内捕获高阶特征交互，以获得更优的决策边界。

在认知合理性上，双层结构对应一种层次化评估机制，可理解为一种隐式的从“特征抽取”到“综合决策”的两阶段流程：第一层网络学习基础视觉概念，例如“主体完整性”“背景清晰度”；第二层对基础概念的得分进行非线性整合与加权决策，输出一个综合的重要性标量。该过程更贴近人类对图像信息进行分层、递进分析的认知过程，使模型决策更具可解释性。

其次，对所有时间步的分数进行softmax归一化，得到归一化的注意力权重。

$$\alpha_t = \frac{exp(S_t)}{\sum_{k=1}^{T} exp(S_k)} \tag{2}$$

式中：权重α_t表示帧t在整体时序中的重要程度。

最后，利用归一化权重对时序特征作加权和，得到全局时序聚合特征。

$$Y_g = \sum_{t=1}^{T} \alpha_t X_t \tag{3}$$

式中：$\boldsymbol{Y}_g$为该视频片段经过全局静态专家得到的时序聚合表示，形状为$[B, 1, D]$。

相较于直接使用CLS token或平均池化而言，这种内容感知的聚合方式能更鲁棒地产生视频的全局静态摘要，为核心的动作识别提供稳定可靠的基线信息。总而言之，该方法为整个框架提供了一个不受复杂时序动态干扰的、纯净的静态语义基线，在处理“坐在桌前”“躺在沙发上”等由场景和对象主导的动作时效果较好。

2.3 短程动态专家

全局静态专家与长程上下文专家分别专注于全局静态内容与关键判别帧，但均主动放弃了对帧间复杂依赖关系的显式建模。视频中存在大量动作，例如“穿衣”“握手”的本质是一系列连续的、具有因果关系的子动作构成的完整时序流程。为了捕捉这种中短程、结构化的动态演变，设计了短程动态专家SDE学习这些静态特征在时间维度上的动态演变残差。

综上，SDE的设计具有以下两大优势：①减轻优化负担：模型无需从零开始学习完整的时序动态，只需学习与基准静态特征之间的“变化量”，收敛更快、更稳定；②功能解耦，通过与GSE分工合作，可专注于纯动态信息，避免了时序模块“遗忘”静态语义的风险。

首先，为输入X_t加入可学习的位置编码P_t，用于提供帧间顺序信息。

$$\tilde{X}_t = X_t + \boldsymbol{P}_t \tag{4}$$

式中：P_t的形状为$[1 \times T \times D]$，在所有批次中共享。

其次，定义一个标准的Transformer编码器模块。本文使用一层编码器层，采用4个注意力头，将前馈网络的隐藏层维度设置为4D，以此平衡不同子空间的表征能力与模型复杂度。编码器借助自注意力机制捕捉帧间的全局依赖，输出上下文增强的特征H_t。

$$H_t = TransformerEncoder(\tilde{X})_t \tag{5}$$

最后，通过一个残差连接与层归一化，将原始的静态特征和学习到的动态变化相加，以获得富含时序信息的帧级特征。

$$\boldsymbol{Y}_s = LayerNorm\left(\tilde{X}_t + H_t\right) \tag{6}$$

式中：Y_s输出形状为$[B, T, D]$，是一个保留了所有时间步的、增强后的特征序列。这表明SDE提供了帧级别的、富含上下文的动态表征，给模型理解中等长度的动作演变提供了核心工具。

2.4 长程上下文专家

在视频动作识别的认知体系中，存在一类与时空背景紧密耦合的动作。其判别性既非来自单一的静态场景，也非依靠连续性的帧间动态，而是扎根于贯穿整个视频片段、稀疏分布的多个关键视觉概念之间的长程语义关联。例如，“调配鸡尾酒”的本质并非某个特定姿势，而是存在于“取瓶”“倾倒”“摇晃”等一系列离散但语义连贯的关键瞬间所构成的宏观叙事链。

因此，设计长程上下文专家LCE充当模型的“战略分析师”，摆脱局部与连续限制，从整个视频的宏观方面，识别并整合在语义上具有内在联系的决定性瞬间。它相较于GSE的“广度聚合”、SDE的“深度挖掘”具有明显差异，代表着一种“高度提炼”的认知维度。

为了实现对长程上下文的聚焦式建模，LCE引入一个可学习的动作原型向量$P \in \boldsymbol{R}^D$作为全局判别标准。首先，LCE通过计算每一帧特征X_t与动作原型P的余弦相似度，以此衡量该帧对动作识别的潜在贡献度。

$$A_t = \frac{\boldsymbol{P}^{\top} X_t}{\|\boldsymbol{P}\| \, \|X\|} \tag{7}$$

式中：$A_t \in [-1, 1]$，值越高表示该帧特征方向和学习到的动作本质概念匹配程度越高；原型向量$\boldsymbol{P}$在训练过程中与网络其他参数一同优化，目的是收敛至一个能描绘“何谓判别性动作特征”的跨类别统一方向。

为了防止信息稀释并强化模型的决策聚焦性，对所有帧的分数$\{A_t\}_{t=1}^{T}$执行一次硬选择操作，仅保留分数最高的K个帧，从而建立了关键帧集合$\mathcal{T}$。

$$T = TopK\left(\{A_t\}_{t=1}^{T}, K\right) \tag{8}$$

式中：K为一个超参数，通常被设置成一个较小的整数（例如$K = 3$）。该步骤主动摒弃了大部分帧，迫使模型仅关注与动作原型匹配度最高、最具判别性的稀疏瞬间，不但大幅降低了后续的计算量，还模拟了人类在理解长视频时忽略冗余信息、聚焦关键片段的认知过程。

其次，LCE仅在关键帧集合T内部进行注意力加权聚合，用以融合多个关键视角的信息，同时保持其相对的判别性强度。

$$\beta_t = \frac{exp\left(A_t/\tau\right)}{\sum_{k\in T} exp\left(A_k/\tau\right)}, \forall t \in T \tag{9}$$

式中：$\tau > 0$为可调节的温度系数。当$\tau < 1$时，Softmax函数输出的权重分布会更加“尖锐”，使绝大部分权重集中在分数最高的1–2个帧上；当$\tau > 1$时，分布会更加平滑。通过调节τ，可控制LCE输出的“聚焦程度”。

最后，利用归一化的注意力权重β_t对原始关键帧特征进行加权求和，得到LCE的判别性时序表征。该表征有两个核心特性：①天然具有稀疏性，因为它仅来源于少数几个关键帧；②具有极强的可解释性，权重β_t直接表明哪些帧对当前决策起到了最大的贡献作用。

$$\boldsymbol{Y}_l = \sum \beta_t X_t \tag{10}$$

式中：$Y_l \in \boldsymbol{R}^{[B\times 1\times D]}$为LCE模块的输出。

尽管GSE与LCE在顶层都采用了加权聚合形式，但在目标、机制与输出特性上存在根本差异。具体地，GSE运用非线性MLP评估单帧的静态完整性，并对所有帧进行软融合，目的是获得一个稳定的场景基线；LCE使用原型匹配评估帧的动态判别性，先进行硬选择再聚合，旨在提炼一个尖锐的动作特征。如此分工使多专家架构能协同覆盖从场景理解到瞬间捕捉的完整认知谱系。

2.5　自适应专家调度器

传统多分支网络通常采用静态加权（如平均池化）或基于注意力的无差别融合方式，隐含地假设不同专家对所有输入的贡献是恒定或同质的。然而，视频数据有着内在多样性：有的视频主要是静态场景，而有的视频包含复杂的演变过程。因此，一个出色的融合机制应赋予模型“元认知”能力，使其能先诊断输入视频的时序特性，再动态、有区分度地调度最适合的专家。

为此，设计了自适应专家调度器。首先，统一表征空间与全局上下文感知，确保所有专家的输出位于一个能比较的语义空间；其次，给定3位专家的输出：Y_g、Y_l的形状为$[B, D]$、Y_s的形状为$[B, T, D]$；最后，对序列输出Y_s进行时序维度的平均池化，将其转换为视频级表征$\bar{Y}_s$，从而将所有专家的输出统一到相同的语义空间。

$$\bar{\boldsymbol{Y}}_s = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{Y}_s[:, t, :] \tag{11}$$

本文对SDE的输出采用mean池化，是处于以下考量：①SDE作为主干专家，借助了Transformer层充分建模时序依赖，mean池化可提取稳健的全局时序表征；②与辅助专家（GSE、LCE）的输出维度保持一致，便于门控加权融合；③在小样本场景下，避免过度复杂的时序建模有助于防止过拟合。

其次，为GSE、LCE专家分别配备独立的门控网络，网络以专家自身的输出作为输入，旨在让专家进行“自我评估”，根据自身输出蕴含的信息特性来生成其对于最终决策的置信度权重。GSE的门控权重g_g为：

$$g_g = \sigma\left(W_{g_2}\cdot \mathrm{ReLU}\left(W_{g_1}Y_g\right) + b_{g_1}\right) + b_{g_2} \tag{12}$$

式中：W_{g_1}、W_{g_2}、b_{g_1}、b_{g2}为GSE的门控网络权重和偏置；$\sigma(\cdot)$为Sigmoid函数，保证输出权重在$[0, 1]$区间。

同理，LCE的门控权重g_l为：

$$g_l = \sigma\left(W_{l_2}\cdot ReLU\left(W_{l_1}Y_l\right) + b_{l_2}\right) + b_{l_2} \tag{13}$$

这种“自我评估”机制的核心优势在于高效性与针对性，强迫每位专家不仅产生特征，还要对特征在当前输入下的“有用性”进行打分。例如，当一个视频内容为剧烈运动时，GSE的门控网络会倾向于输出一个较低的g_g，因为静态信息可能是次要的；相反，对于一个“谈话”视频，LCE的门控网络可能会输出较低的g_l，原因为内容缺乏决定性的关键帧。

最后，融合结果Y为三部分加权相加的和。

$$Y = \bar{Y}_s + g_g \odot Y_g + g_l \odot Y_l \tag{14}$$

式中：$\odot$表示元素按批次广播的标量乘法；g_g、g_l维度从$[B, 1]$自动扩展到$[B, D]$；$\bar{Y}_s$为SDE的均值池化输出，不带门控权重，起主干作用；g_g、g_l为动态自适应权重，决定两个辅助专家对融合结果的贡献度。

式（14）体现了自适应专家调度器的核心思想：提供一种“稳健基准+动态调制”的策略。首先，在信息架构上实现了各专家明确的分工。SDE的核心职责是：通过Transformer显式地建模复杂的帧间依赖与时序动态。对SDE的输出Y_s进行平均池化的目的并非丢弃时序信息，而是将经过充分交互建模的时序特征序列提炼为一个全局的、稳健的时序上下文基准。GSE、LCE则分别提供全局静态特征与关键帧上下文这两种特定视角的增强信息，将$\bar{Y}_s$作为主干，无论输入视频内容如何变化，都能确保融合结果立足于一个可靠的时序表示。同时，由两个带门控的辅助专家Y_g、Y_l作为场景自适应的调制器，他们的门控权重g_g、g_l根据当前视频内容动态生成，负责对主干基准进行针对性的增强与修正。例如，当动作高度依赖场景时，增大g_g以强化静态信息；当动作有关键帧时，增大g_l以聚焦关键帧上下文。

其次，在模型优化上提供至关重要的训练稳定性。将$\bar{Y}_s$设置为固定权重，为梯度回传提供了一条恒定、无衰减的“绿色通道”。训练初期，当门控网络输出不稳定时，如果所有专家均受门控所影响，将会使梯度变得嘈杂甚至消失，导致训练困难。为此，本文确保了模型在训练的全周期内，始终能获得来自SDE清晰、稳定的强梯度信号，以显

著提升模型收敛速度与训练稳定性。

3 实验结果与分析

3.1 数据集

HMDB51数据集是一个经典的人体动作识别数据集，包含51个动作类别，约7 000个视频片段，动作类别涵盖“笑”“喝”等面部动作，“鼓掌”“跳跃”等肢体动作、“抽烟”“拔剑”等物体交互动作，因其多样的视频来源与动作模式，常被用于评估模型泛化能力[26]。HMDB51的独特价值在于其视频来源的多样性，大部分来源于老电影，在光照、拍摄角度、视频质量和背景等方面存在巨大差异。这种固有的真实世界复杂性与相对有限的样本数量，使其成为评估模型鲁棒性与泛化能力的试金石。综上，所提模型在HMDB51上的表现，能有效反映其在有限且嘈杂的数据中学习本质动作特征，而非过拟合于背景关联的能力。

UCF101数据集是动作识别领域常用的大规模基准数据集，包含101个动作类别和超过13 000个视频。这些视频源自真实的网络视频，动作被划分为人与物体交互、人体动作、人与人交互、乐器演奏与体育运动等五大类[27]。高质量的帧序列与清晰的动作定义，使其成为动作识别领域核心的评估基准。相较于HMDB51而言，UCF101的视频通常具有更高的分辨率、更清晰的动作定义，但仍存在相机抖动、遮挡、类内差异及背景杂乱等问题。UCF101因在数据规模与质量上的平衡，长期被视为衡量模型通用识别性能的核心基准。在该数据集上的优异表现，通常可证明模型具备了强大的空间及表观特征理解能力。

Something-Something V2（SSv2）数据集是一个大规模视频数据集，专注于人类基础性、交互式动作，包含约220 000个视频片段，覆盖174个细粒度的动作类别[28]。这些动作大多表现为简单的手部与物体的互动，例如“放下某物”“拿起某物”“用手指将某物向右推”等。该数据集的根本特性在于：动作的识别强烈依赖对时序动态与因果关系的精确理解，而非静态的场景或物体外观。例如，区分“拿起某物”“放下某物”的关键在于物体相对于手移动的方向；识别“假装打开某物”与“真正打开某物”时需要理解动作的意图与结果。因此，SSv2对模型的时序建模能力提出了极其严苛的要求，使其成为评估视频理解模型是否真正“看懂”了动作的权威基准。在该数据集上的性能，直接反映了模型捕捉帧间微妙变化和理解动作演变逻辑的能力。

3.2 实验设置

本文在PyTorch环境下进行实验，配置了一块NVIDIA RTX 3090 GPU，所有输入视频帧被预处理为224×224大小，并随机采样32帧作为模型输入，Few-shot和base-to-novel泛化实验均参照ViFi-CLIP进行设定[15]。

在Few-shot实验中，采用在Kinetics-400上预训练的ViT-B/16作为视觉主干网络，文本token长度固定为77。同时，启用视觉与文本双模态的Prompt Tuning，在每个模态的12层网络中分别插入10个可学习的prompt token。模型训练优化器为AdamW，基础学习率设为8×10-3（SSv2数据集单独设置为2×10-4）。模型的权重衰减系数均为0.001，共训练50个epoch，批处理大小为4，采用每4步累积一次梯度的策略以适应单卡显存限制，训练时每10个epoch保存一次检查点。

在base-to-novel泛化实验中，Prompt Tuning的配置调整为每个模态插入16个prompt token，深度为9层。优化过程采用每16步累积一次梯度，训练轮数为11，学习率设置为2×10-2（SSv2数据集的学习率单独设置为1×10-2）。

3.3 实验结果

为评估DMEF在数据稀缺场景下的泛化能力，在HMDB-51、UCF-101和SSv2数据集上进行少样本学习实验（K=2、4、8、16），并与一系列基于CLIP的先进方法进行比较，如表1所示。其中，最佳结果以粗体显示，次优结果以下划线显示。由此可知，DMEF方法在HMDB-51、UCF-101数据集上优势明显，领先所有比较方法。具体地，在极具挑战的4-shot设置下，DMEF在HMDB-51、UCF-101上的准确率分别为66.3%、93.5%，相较于T2L提升5.2%，证明DMEF能更高效地利用极其有限的标注样本，学习到更具判别性的特征表示[33]。

Table 1 Comparison with the latest methods in few sample learning

表1 与最新方法在少样本学习上的比较

方法	发表出处	HMDB-51				UCF-101				SSv2			
		K=2	*K*=4	*K*=8	*K*=16	*K*=2	*K*=4	*K*=8	*K*=16	*K*=2	*K*=4	*K*=8	*K*=16
X-CLIP[29]	ECCV 2022	53.0	57.3	62.8	62.4	71.4	79.9	83.7	91.4	3.9	4.5	6.8	10.0
A5[16]	ECCV 2022	46.7	50.4	61.3	65.8	76.3	84.4	90.7	93.0	4.5	6.7	7.2	9.5
ViFi-CLIP[15]	CVPR 2023	57.2	62.7	64.5	66.8	80.7	85.1	90.0	92.7	6.2	7.4	8.5	12.4
MAXI[30]	ICCV 2023	58.0	60.1	65.0	66.5	86.8	89.3	92.4	93.5	7.1	8.4	9.3	12.4
ViLT-CLIP[31]	AAAI 2024	60.6	61.9	66.9	69.6	85.3	90.0	91.3	93.8	**7.8**	**9.4**	10.3	13.2
TC-CLIP[32]	ECCV 2024	57.3	62.3	67.3	68.6	85.9	89.9	92.5	94.6	7.3	8.6	9.3	**14.0**
T2L[33]	TMLR 2025	57.3	61.1	65.4	67.7	84.7	88.3	91.3	92.8	6.8	8.7	9.6	13.2
DMEF	–	**61.1**	**66.3**	**68.6**	**70.9**	**89.8**	**93.5**	**94.7**	**96.3**	6.8	9.0	**10.4**	12.9

值得深入分析的是，在更具时序推理挑战性的SSv2数据集上，DMEF表现并没有达到最优，反映了不同数据集对动作建模的要求不同。文献[34]发现，在UCF101、HMDB51等数据集上，模型对时序扰动普遍表现出鲁棒性，但在SSv2上则极度敏感，性能显著下降，原因为SSv2数据集中存在许多动作，例如“放入”“取出”本身就由精确的时序顺序定义，任何时序扰动都会从根本上改变动作的语义，从而证明了时序信息对SSv2的动作识别具有决定性作用，而对外观和短程动态主导的数据集影响不大。

本文研究及多数比较方法均基于CLIP模型，核心优势在于从海量图像—文本对中学习强大的通用静态视觉先验。因此，当前基于CLIP的视频识别方法，本质上是利用并适配这一强大的静态特征，以理解和补充视频中的时序信息。当前SDE专家更侧重于建模相邻帧间的局部运动模式，但对于跨越多帧的、复杂的长程时序关系推理SSv2数据集，其建模深度尚未构成决定性的优势。换而言之，DMEF提供了一个高效的集成框架，设计重点在于灵活融合视频在时间维度上的多尺度特征，但对于SSv2所需的深度长程时序推理这一特定核心能力，当前框架的建模能力仍然不够，导致DMEF在SSv2上的性能无法达到最优。

跨域泛化任务是评估模型在学习已知类别（基类）后，将知识迁移到未见类别（新类）上的能力，实验结果如表2所示，其中，“Base”“Novel”“HM”分别代表基类准确率、新类准确率和调和平均值；加粗为最优结果，下划线为次优结果。在HMDB-51、UCF-101数据集上，DMEF的HM分别为65.3、84.7，显著领先于其他方法，验证了模型从有限基类中学习到的表征能有效迁移至未见新类，未出现灾难性遗忘现象。在更具时序挑战的SSv2数据集上，DMEF的HM为14.6，进一步彰显了其时序建模优势。该数据集中基类与新类在时序逻辑上存在显著差异，DMEF通过其自适应专家调度器，能从具体动作类别中抽离、提炼出通用的基本时序动态模式。这些模式作为与类别无关的“时序常识”，可被灵活重组，用于理解、识别未曾学习过的、但包含相似基本动态的新动作，从而实现优异的跨类别泛化性能。

为了评估模型效率，在单张NVIDIA RTX 3090 GPU环境下，比较DMEF与现有方法的计算成本，包括参数量（Params）、浮点运算量（GFLOPs）、吞吐量（Throughput），实验结果如表3所示。由此可知，DMEF方法的总参数量为131.19 M，仅0.15 M的参数需要训练（占0.11%），具有极高的参数效率。同时，极低的可训练参数量能大幅减少梯度计算和存储开销，特别适用于数据稀缺的少样本学习场景。

Table 2　Comparison with the latest methods in cross domain generalization ability

表2　与最新方法在跨域泛化能力上的比较

方法	发表出处	HMDB-51			UCF-101			SSv2		
		Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
Vanilla CLIP[1]	ICML 2021	53.3	46.8	49.8	78.5	63.6	70.3	4.9	5.3	5.1
A5[16]	ECCV 2022	70.4	51.7	59.6	95.8	71.0	81.6	12.9	5.7	7.9
X-CLIP[29]	ECCV 2022	75.8	52.0	61.7	95.4	74.0	83.4	14.2	11.0	12.4
ViFi-CLIP[15]	CVPR 2023	73.8	53.3	61.9	92.9	67.7	78.3	16.2	12.1	13.9
SC-CLIP[35]	PR 2025	72.2	57.2	63.8	93.5	**77.0**	84.5	15.8	12.6	14.0
DMEF	–	**76.4**	57.1	**65.3**	**96.6**	75.1	**84.7**	**16.9**	**12.9**	**14.6**

Table 3　Comparison of model complexity and efficiency of different methods

表2　不同方法的模型复杂度与效率比较

方法	发表出处	HMDB-51			UCF-101			SSv2		
		Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
Vanilla CLIP[1]	ICML 2021	53.3	46.8	49.8	78.5	63.6	70.3	4.9	5.3	5.1
A5[16]	ECCV 2022	70.4	51.7	59.6	95.8	71.0	81.6	12.9	5.7	7.9
X-CLIP[29]	ECCV 2022	75.8	52.0	61.7	95.4	74.0	83.4	14.2	11.0	12.4
ViFi-CLIP[15]	CVPR 2023	73.8	53.3	61.9	92.9	67.7	78.3	16.2	12.1	13.9
SC-CLIP[35]	PR 2025	72.2	57.2	63.8	93.5	**77.0**	84.5	15.8	12.6	14.0
DMEF	–	**76.4**	57.1	**65.3**	**96.6**	75.1	**84.7**	**16.9**	**12.9**	**14.6**

在计算方面，DMEF的单视频推理计算量为456 GFLOPs，实际吞吐量达到9.6 samples/sec。相较于ViFi-CLIP，DMEF以60%的计算量增加换取了12.8%的吞吐量，在可训练参数量上得到了优化（0.15 M vs 0.18 M）。DMEF在参数效率、计算效率和实际性能之间取得了良好平衡，特别适合资源受限的应用场景。

综上，DMEF框架通过构建多专家协同机制和自适应专家调度策略，有效增强了视频时序多尺度建模能力。在HMDB-51、UCF-101和SSv2数据集中，所提框架均展现出了优越的分类性能和泛化能力。

3.4　消融实验

为系统性验证DMEF中各组件的贡献，在HMDB51数据集的4-shot设定下进行消融研究。所有实验均采用相同的ViFi-CLIP视觉编码器与超参数设置，确保结论的公平性与可靠性[15]。同时，通过顺序去除每个组件来训练模型，以逐一评估各时序专家的独立效能与协同效应，实验结果如表4所示。

Table 4　Results of ablation experiment

表4　消融实验结果

专家模块	GSE	SDE	LCE	准确率/%
DMEF w/o GSL				65.10
DMEF w/o SL	√			65.37
DMEF w/o GL		√		65.81
DMEF w/o GS			√	65.58
DMEF(Full)	√	√	√	66.34

（1）DMEF w/o GSL：在不引入任何专家模块的情况下，使用原始ViFi-CLIP作为基准模型。此时仅对提取的帧级特征进行全局平均池化，模型准确率为65.10%。

（2）DMEF w/o SL：除去SDE、LCE专家的模型。模型

的准确率提升至65.37%(+0.27%),这一明确且稳健的增益证实了DMEF的核心设计动机:通过内容感知注意力机制对帧级特征进行重要性筛选与加权聚合,可有效守护、增强CLIP模型强大的静态视觉先验。具体为,GSE通过提供一种稳定且鲁棒的全局视频摘要,为模型识别静态表观主导的动作奠定坚实基础。

(3)DMEF w/o GL:除去GSE、LCE专家的模型。该模型展现出了SDE的性能,准确率达65.81%(+0.71%),有力证明了显式建模帧间复杂依赖关系的重要性。DMEF采用的Transformer架构配合残差连接,使模型能深入理解视频中的动态演变过程,可同时确保训练过程的稳定性,有效规避了小样本场景下的过拟合风险。

(3)DMEF w/o GS:除去GSE、SDE专家的模型。该模型取得了65.58%的准确率,相较于基线模型提升0.48%,虽然幅度不大,但证明了LCE成功扮演了所设计的"判别性时刻探测器"角色。相较于处理整体场景或连续动态的专家而言,LCE通过轻量而高效的线性注意力机制,实现了对长视频序列的快速扫描,能自动定位并聚焦"投篮出手""扣篮触框"等转瞬即逝的关键帧。为模型提供了高信噪比的决策依据,在面对外观相似但核心动作瞬间迥异的细粒度类别(如"拿起某物"与"放下某物")时,这种聚焦能力成为了正确分类的关键。

当3个专家通过自适应门控网络协同工作时,模型性能达到了最高的66.34%,融合后的性能不仅显著超越基线,还超越了任何单一或双专家组合所能达到的性能上限,为DMEF框架提供了最有力的实证支持。实验表明,3个专家并非冗余设计,可分别从静态场景理解、连续动态建模和关键瞬间定位等3个正交且互补的认知维度来解析视频时序信息。

同时,为了探究式(14)中"主干固定+辅助自适应"权重设计机制的必要性,在UCF101数据集k=4的小样本设置下,对不同的专家融合策略进行消融研究,实验结果如表5所示。其中,3种融合策略分别为:①固定权重,分别将Y_g、Y_l权重设置为固定参数0.5;②全部门控,3个专家输出均经自适应门控加权;③部分门控(本文策略),即仅对辅助专家进行门控,主干专家固定权重。

Table 5 Adaptive gating ablation experiment results

表5 自适应门控消融实验结果

融合策略	权重机制	公式表示	准确率/%
固定权重	固定	$\bar{Y}_s + 0.5\odot Y_g + 0.5\odot Y_l$	93.2
全部门控	全部自适应	$g_s\odot\bar{Y}_s + g_g\odot Y_g + g_l\odot Y_l$	93.0
部分门控	主干固定+辅助自适应	$\bar{Y}_s + g_g\odot Y_g + g_l\odot Y_l$	**93.5**

实验表明,全部门控策略的性能最低(93.0%),相较于本文方法下降0.5%,证明了在小样本场景下,对主干专家SDE输出固定权重,可为模型训练提供稳定的梯度通路,避免了因门控网络初期不稳定而导致的优化震荡。同时,固定权重策略(93.2%)的性能虽优于全部门控,但仍低于部分门控的策略(93.5%),说明自适应门控机制能依据视频内容动态调整各辅助专家的贡献权重,相较于经验性固定权重的特征融合而言更精准,进一步体现了"主干固定+辅助自适应"设计原则的优越性。

综上,消融实验从多个专家协同建模的必要性和自适应调度器设计的合理性两个维度,共同验证了所提框架的有效性。

3.5 可视化实验

为了深入探究DMEF的优势体现在哪些动作类别上,在HMDB51数据集上进行了类别分析。图3直观比较了DMEF与基线方法ViFi-CLIP在24个代表性动作类别上的分类准确率。由此可见,DMEF的性能增益在以下3类动作中表现得尤为显著:

(1)快速瞬变动作:hit,fencing等。此类动作的判别性信息高度集中于极短的关键瞬间,得益于SDE通过注意力机制精准定位并放大了"击剑刺中""挥棒击打"等动作的决胜时刻。相较于ViFi-CLIP所采用的相对简单的时序融合策略,所提模型对此类短暂但关键的信号响应更及时。

(2)连续时序动作:run,walk等。此类动作由一系列连贯的姿势变化构成,DMEF通过Transformer显式建模帧与帧之间的动态演变关系,能更好地理解"起跳—腾空—落地""攀爬"的完整过程,相较于ViFi-CLIP的时序建模能力更强,对动作的连贯性理解更深。

(3)静态姿态动作:sit,stand等。DMEF同样保持了微弱的优势,证明了GSE能有效地守护CLIP的静态先验,且门控网络在处理此类视频时,倾向于依赖GSE提供稳定且准确的表征,避免了复杂时序模块可能引入的噪声,体现了DMEF框架的鲁棒性。

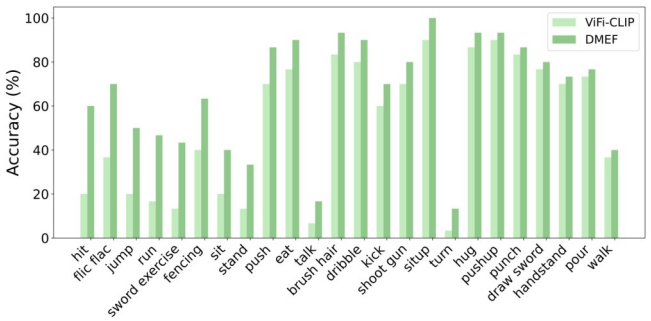


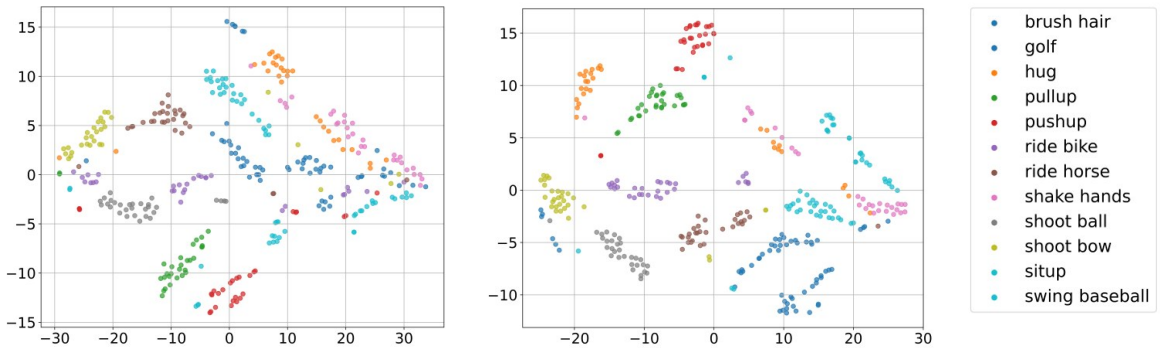
Fig. 3 Classification accuracy of ViFi-CLIP and DMEF on HMDB-51 (24 categories)

图3 ViFi-CLIP与DMEF在HMDB-51(24个类别)上的分类准确率

为了从表征层面探究DMEF的优越性,使用t-SNE工具可视化ViFi-CLIP、DMEF在HMDB-51数据集前12个类别上的特征嵌入分布,为DMEF的有效性提供直观且有力的证据,如图4所示。由此可见,DMEF所生成的特征簇呈现出了更高的类内聚合度与更清晰的类间分离度。例如,shake hands、hug类别在ViFi-CLIP的特征空间中存在显著重叠,而在DMEF中则被明确地划分开来,一个边界清晰、结构紧凑的特征空间,意味着模型学习到了各类别最本质

的判别特征。因此，当模型面对同一数据分布但未曾见过的新类时，这些新类样本将更有机会落入特征空间中恰当的区域内，从而被正确分类。

这种特征质量的提升的原因为：①SDE通过建模帧间动态，学习到了与动作演变过程相关的鲁棒特征，即使同一动作的起始姿态不同，其整体动态模式也能在特征空间中被拉近；②LCE通过聚焦最具判别性的帧，过滤了视频中的冗余信息，生成了纯度更高的类别表征，促进了类内聚合；③GSE确保模型不会丢失与场景和物体相关的基础语义信息，为特征提供了稳定的语义锚点；④门控网络自适应地融合上述多元信息，使每个类别的最终表征都汇聚了最相关的静态、动态和关键判别信息，在特征空间中形成了界限分明的簇。



（a）　ViFi-CLIP　　（b）　DMEF　　（c）　label

Fig. 4　Feature distribution of ViFi-CLIP and DMEF on HMDB-51 (4-shot, first 12 classes)

图4　ViFi-CLIP与DMEF在HMDB-51(4-shot，前12类)上的特征分布

4　结语

针对现有方法普遍采用单一时序建模范式所导致的特征粒度受限、泛化性能不足及解释性不强等问题。本文提出动态多专家融合框架DMEF，实现了小样本场景下视频动作识别的准确分类与泛化迁移。首先，构建“全局—短程—长程”三专家结构，对视频时序不同尺度进行动态建模；其次，引入基于轻量门控网络的自适应调度器，依据视频内容动态激活、融合各专家特征。

实验表明，DMEF在HMDB51、UCF101和SSv2数据集的小样本设置(k=8)实验中，分别取得68.6%，94.7%和10.4%的准确率，显著优于基线模型。尽管，DMEF框架在多个数据集上性能良好，但计算效率仍有待提升。此外，当前框架未对文本端进行提示增强设计，未来将引入文本端Prompt机制，以进一步提升跨模态特征对齐质量，并将模型拓展至开放词汇视频理解等更广泛的任务中。

参考文献：

[1] ZHU S W, YE B L, WU W M. Review and simulation of short-term traffic flow prediction methods based on deep learning[J]. Software Guide, 2024, 23(2): 182-193.
朱仕威，叶宝林，吴维敏. 基于深度学习的短时交通流预测方法综述与仿真研究[J]. 软件导刊，2024，23(2)：182-193.

[2] RADFORD A, KIM J W, 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.

[3] LI Y, WU S, XU C, et al. A review of semantic text similarity calculation methods[J]. Software Guide, 2024, 23(11): 1-11.
李莹，伍胜，徐聪，等. 语义文本相似度计算方法研究综述[J]. 软件导刊，2024，23(11)：1-11.

[4] JIA C, YANG Y, XIA Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]// Proceedings of the 38th International Conference on Machine Learning, 2021: 4904-4916.

[5] 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.

[6] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 4015-4026.

[7] ZOU X, YANG J, ZHANG H, et al. Segment everything everywhere all at once[J]. Advances in Neural Information Processing Systems, 2023, 36: 19769-19782.

[8] LI L H, ZHANG P, ZHANG H, et al. Grounded language-image pre-training[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 10965-10975.

[9] MINDERER M, GRITSENKO A, STONE A, et al. Simple open-vocabulary object detection[C]// European Conference on Computer Vision, 2022: 728-755.

[10] WANG M M, XING J Z, LIU Y. Actionclip: a new paradigm for video action recognition[DB/OL]. https://arxiv.org/abs/2109.08472.

[11] JING X, YANG G, CHU J. An empirical study of excitation and aggregation design adaptions in clip4clip for video-text retrieval[J]. Neurocomputing, 2024, 596: 127905.

[12] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[DB/OL]. https://arxiv.org/abs/1706.03762.

[14] BERTASIUS G, WANG H, TORRESANI L. Is space-time attention all you need for video understanding?[DB/OL]. https://arxiv.org/abs/2102.05095.

[15] RASHEED H, KHATTAK M U, MAAZ M, et al. Fine-tuned clip mod-

els are efficient video learners[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 6545-6554.

[16] JU C, HAN T, ZHENG K, et al. Prompting visual-language models for efficient video understanding[C]// European Conference on Computer Vision, 2022: 105-124.

[17] LI J, TANG S, ZHU L, et al. Adaptive hierarchical graph reasoning with semantic coherence for video-and-language inference[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 1867-1877.

[18] YUAN L, HE Z, WANG Q, et al. Advancing human motion recognition with skeletonclip++: weighted video feature integration and enhanced contrastive sample discrimination[J]. Sensors, 2024, 24(4): 1189.

[19] LEPIKHIN D, LEE H, XU Y, et al. Gshard: scaling giant models with conditional computation and automatic sharding[DB/OL]. https://arxiv.org/abs/2006.16668.

[20] FEDUS W, ZOPH B, SHAZEER N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity[J]. Journal of Machine Learning Research, 2022, 23(1): 5232-5270.

[21] PARK K, KIM Y, KIM D, et al. Resilient sensor fusion under adverse sensor failures via multi-modal expert fusion[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025: 6720-6729.

[22] CAI W, LIU Q, WANG Y. SPMTrack: spatio-temporal parameter-efficient fine-tuning with mixture of experts for scalable visual tracking[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025: 16871-16881.

[23] WANG Z, LIU L, WANG L, et al. METransformer: radiology report generation by transformer with multiple learnable expert tokens[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 11558-11567.

[24] LI Z, HU Y, YIN B, et al. Multi-token enhancing for vision representation learning[DB/OL]. https://arxiv.org/abs/2411.15787.

[25] HORNIK K, STINCHCOMBE M, WHITE H. Multilayer feedforward networks are universal approximators[J]. Neural Networks, 1989, 2(5): 359-366.

[26] KUEHNE H, JHUANG H, GARROTE E, et al. Hmdb: a large video database for human motion recognition[C]// 2011 International Conference on Computer Vision, 2011: 2556-2563.

[27] SOOMRO K, ZAMIR A, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild[DB/OL]. https://arxiv.org/abs/1212.0402.

[28] GOYAL R, KAHOU S E, 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: 5843-5851.

[29] NI B, PENG H, CHEN M, et al. Expanding language-image pretrained models for general video recognition[C]// European Conference on Computer Vision, 2022: 1-18.

[30] LIN W, KARLINSKY L, SHVETSOVA N, et al. Match, expand and improve: unsupervised finetuning for zero-shot action recognition with language knowledge[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 2839-2850.

[31] WANG H, LIU F, JIAO L, et al. Vilt-clip: video and language tuning clip with multimodal prompt learning and scenario-guided optimization[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2024: 5390-5400.

[32] KIM M, HAN D, KIM T, et al. Leveraging temporal contextualization for video action recognition[C]// European Conference on Computer Vision, 2024: 74-91.

[33] AHMAD S, CHANDA S, RAWAT Y S. T2L: efficient zero-shot action recognition with temporal token learning[EB/OL]. https://openreview.net/forum?id=WvgoxpGpuU.

[34] SCHIAPPA M C, BIYANI N, KAMTAM P, et al. A large-scale robustness analysis of video action recognition models[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 14698-14708.

[35] QUAN Z, CHEN J, DEGUCHI D, et al. Semantic matters: a constrained approach for zero-shot video action recognition[J]. Pattern Recognition, 2025, 162: 111402.

(责任编辑:刘嘉文)