E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP18;TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2026)04-1156-16

论文引用格式:Zhao L J, Liu X T, Wu Q B and Meng F M. 2026. From association to refinement: scalable chain-of-thought-guided few-shot continual teaching behavior recognition. Journal of Image and Graphics, 31(4):1156-1171(赵珞君, 刘小同, 吴庆波, 孟凡满. 2026. 从联想到凝练:可伸缩思维链引导的少样本连续教学行为识别. 中国图象图形学报, 31(4):1156-1171)[DOI:10.11834/jig.250327]

从联想到凝练:可伸缩思维链引导的少样本连续教学行为识别

赵珞君,刘小同,吴庆波*,孟凡满
电子科技大学信息与通信工程学院,成都　611731

摘　要:**目的**　教学行为识别在智慧课堂领域有着广泛的应用,但在实际教学场景中,随着各种教学改革的推进,新型的教学行为会不断衍生出来。同时由于标注的成本问题,新型教学行为的标注样本量相当有限。在这样的情况下,如何保证模型具有少样本学习能力,成为该任务的主要挑战。现有的少样本连续学习算法大多基于预训练的视觉语言模型(如CLIP(contrastive language-image pre-training)),通过微调主干网络进行图像与文本特征的匹配。然而,这些研究往往忽略了如"听讲"、"写字"等行为标签本身就含有丰富的语义信息。为此,提出了一种可伸缩思维链引导(scalable chain-of-thought-guided,SCOTG)的少样本连续教学行为识别算法。**方法**　具体而言,首先通过思维链生成有关行为标签的详细描述性文本,对行为标签的语义进行扩展挖掘,之后提取(主,谓,宾)结构的三元组知识表示,凝练结构化知识,从而更精准地反映出行为中的关键实体和关系,帮助模型更好地理解识别行为动作。SCOTG算法设计了多层次跨模态匹配机制,将不同层次的三元组文本特征与图像的多层视觉特征进行相似度匹配计算。与传统方法相比,SCOTG算法冻结了预训练视觉语言模型的主干网络,只对行为标签进行伸缩处理,通过提示学习训练视觉语言模型,降低了计算复杂度。**结果**　实验在具有32个行为类别的教室场景图像数据集ARIC(activity recognition in classroom)上与7种方法进行了比较,在3-way 5-shot任务设置下,相比于性能第2的模型,在所有任务中平均准确率提升了1.98%,最后任务中的平均准确率提升了1.36%。**结论**　提出的SCOTG算法有效提高了模型对于教学行为的理解,增强了模型在少样本场景下对新型教学行为的识别能力。代码已开源至https://github.com/2002zlj/scotg。

关键词:教学行为识别;少样本连续学习;思维链;大语言模型(LLM);视觉语言模型(VLM);多层次跨模态匹配

From association to refinement: scalable chain-of-thought-guided few-shot continual teaching behavior recognition

Zhao Luojun, Liu Xiaotong, Wu Qingbo*, Meng Fanman
School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu　611731, China

Abstract: **Objective** Teaching behavior recognition has a wide range of applications in the field of smart classrooms; in particular, the state of students can be analyzed in real time, and teachers can be provided with accurate feedback on teach-

收稿日期:2025-07-16;**修回日期**:2025-11-05;**预印本日期**:2025-11-12

*通信作者:吴庆波　qbwu@uestc.edu.cn

基金项目:科技创新2030——"新一代人工智能"重大项目(2021ZD0112001);国家自然科学基金项目(62271119);四川省自然科学基金项目(2025ZNSFSC0475)

Supported by:Science and Technology Innovation 2030—"New Generation Artificial Intelligence" Major Project(2021ZD0112001);National Natural Science Foundation of China (62271119);Natural Science Foundation of Sichuan Province,China(2025ZNSFSC0475)

ing behaviors to help them adjust the teaching rhythm and improve the teaching methods. In this way, the efficiency of classroom interactions and the quality of knowledge transfer are enhanced. However, in the actual teaching scenario, the promotion of various teaching reforms has resulted in the emergence of new interactive and technology-integrated teaching behaviors in the smart classroom, such as collaborative group discussions using tablets and students displaying results on whiteboards. These new teaching behaviors are different from the traditional teaching behaviors. Moreover, the annotation sample size is quite limited because of the cost of annotation. In such a situation, ensuring that the model has the ability of few-shot continual learning, i. e. , to achieve accurate recognition of new teaching behaviors without catastrophic forgetting of old teaching behaviors, becomes the main challenge of this task. Most of the existing few-sample continuous learning algorithms are based on pretrained visual language models (e. g. , contrastive language-image pre-training (CLIP)), which are fine-tuned to match image and text features through a fine-tuning network. However, these studies often ignore the fact that behavioral labels such as "listening", "writing" and "using a tablet" contain rich semantic information in themselves. For example, the label "writing" contains multilayered semantic information: From the basic semantics, it refers to "the action of a person who holds a writing instrument in his/her hand to record symbols on a specific carrier", which includes the sequential action logic of "hold-move-leave a mark". From the behavioral entity, it involves the student subject, the objects such as pen, notebook, tablet, and the interactions between them, e. g. , "student—with a pen—on a notebook". From the scene perception, it refers to the student writing on a desk in a classroom scene, which is a rich presentation of the deep details in the behavior. However, most existing algorithms simply match "writing" as a single text label with image features, thereby failing to parse out the action logic of the underlying semantics, the interaction relationship of behavioral entities, and the contextual associations of scene perception. Moreover, these existing algorithms have a weak understanding of the deep connotations of the behavior, thereby leading to difficulties in understanding the behavioral actions of the model. Thus, the recognition robustness of a scene with few samples is affected. To this end, a scalable chain-of-thought-guided (SCOTG) algorithm for continuous teaching behavior recognition with few samples is proposed.

Method We generate detailed descriptive text about the behavior labels through the chain of thought, expand the semantics of the behavior labels for mining, and extract the structure (subject, predicate, and object) of the ternary knowledge representation to condense the structured knowledge. In this way, the key entities and relationships in the behavior can be accurately reflected, and the model can deeply understand and identify the behavioral actions. In the stage of "association", the three-stage Q&A of the chain of thought is adopted. The first is the basic semantic level, which allows the big language model to identify the subject of the behavior and the location of the occurrence by adding the necessary nouns. This level also converts the behavior label into a basic sentence. The semantic extension at the entity interaction level follows. It further emphasizes the interactivity of the behavior by introducing other visual entities related to the action. The final and third is the semantic extension at the scene-aware level, which allows LLMs to generate detailed scene descriptions, capture the dynamic changes and reactions of the subjects in the scene, and enrich the details of the behaviors at a fine granularity. In the "refinement" stage, the large language model is used to extract the triples in the sentences, condense the knowledge, remove redundant information, and make it highly structured. Meanwhile, the concise knowledge of the triples makes being handled by the text encoder of CLIP effective. In addition to the scaling of text labels, the SCOTG algorithm designs a multilevel cross-modal matching mechanism to calculate the similarity matching between different levels of ternary textual features and multilayered visual features of the image. This algorithm also employs a hierarchical weighting strategy to set different weights according to the contribution of each layer of features in the behavior recognition task and obtain the total similarity that is finally used for classification, which takes full advantage of the knowledge of the ternary group on the textual side of the hierarchical semantic structure of the text-side triad (e. g. , base action triad, entity interaction triad, and scene perception triad) and the cross-modal associations of the image-side multilayer visual features (e. g. , low-layer texture features, mid-layer part features, and high-level scene features). Among them, the matching of the base action triad with low-level visual features can capture fine-grained action details such as "grip-move". The matching of the entity interaction triad with mid-level visual features can strengthen the entity relationship recognition of "subject-interaction-object" and the matching of the scene. The matching of the association triad with the high-level visual features can correlate with the contextual logic of the behavior. The model can understand the nature of the behavior from semantic to visual dimen-

sions through this kind of multilevel accurate alignment. Thus, the model can effectively make up for the defect of the traditional single-feature matching, which is insufficient for capturing the deeper associations. Compared with traditional methods, the SCOTG algorithm freezes the backbone network for pretraining the visual language model, scales only the behavioral labels, and reduces the computational complexity by training the visual language model through prompt learning. **Result** Experiments were conducted on a classroom scene image dataset with 32 behavioral categories, with seven methods being used for comparison. Compared with the model with the second-highest performance in the three-way five-shot task setting, the SCOTG algorithm showed an average accuracy improvement of 1.98% in all tasks, and 1.36% in the final task. Under the three-way three-shot task setting, the average accuracy in all tasks of the SCOTG algorithm improved by 1.03% compared with that of the model with the second-highest performance. Under the three-way one-shot task setting, the average accuracy of the SCOTG algorithm is improved by 0.78% in all tasks compared with the model with the second-highest performance. **Conclusion** In this study, we propose an SCOTG algorithm for continuous teaching behavior recognition with few samples and design a multilevel cross-modal matching mechanism. Experimental results show that the proposed SCOTG algorithm effectively improves the model's understanding of teaching behaviors and enhances the model's ability to recognize novel teaching behaviors in sampleless scenarios. The code is available at https://github.com/2002zlj/scotg.

Key words: teaching behavior recognition; few shot continual learning; chain-of-thought; large language model(LLM); vision-language model(VLM); multilevel cross-modal matching

0 引 言

课堂教学是教育过程的核心环节,其质量直接影响着学生的学习成效。深入理解课堂中师生互动与行为模式,对于优化教学策略、提升教学效率具有至关重要的意义。教学行为识别(Liu等,2025)旨在通过分析师生在教学活动中展现出的姿态、动作和交互等行为模式,助力教学质量提升与智能课堂构建,在教育领域发挥着重要的作用。传统课堂中,教学行为识别主要依赖人工观察记录。这种方法耗费大量人力物力、效率低下,难以覆盖大规模的课堂场景,无法实时、客观以及全面地反映课堂教学动态,不能及时地提供教学反馈。因此,寻求自动化、智能化的教学行为识别方法成为提升课堂分析效率与准确性的迫切需求。

随着计算机视觉和人工智能的飞速发展,基于深度学习的教学行为识别算法取得了显著突破(Jia和He,2024)。这些算法能够自动、高效地从课堂视频或图像数据中提取师生姿态、动作轨迹、交互状态等高维特征,并识别出特定的教学行为(Gu和Li,2022)。这种技术极大地克服了人工观察的局限性,为实现大规模、实时化和客观精准的课堂教学行为分析,为教学反馈优化与智能课堂构建提供了强有力的技术支撑。

在教学行为识别领域,主要通过对监控视频或图像数据的分析,实现对特定行为的识别与分类(王帅琛,2022)。早期研究中,Simonyan和Zisserman(2014)提出了经典的双流网络(two-stream network)。该网络利用两个并行的卷积神经网络分支,分别提取视频帧的空间外观特征(RGB图像)和运动时序特征(光流图像),并通过融合这两种时空特征显著提升了行为识别效果。为进一步处理长视频序列,Wang等人(2016)引入分段与稀疏化采样的思想,构建了时间分段网络(temporal segment network, TSN),有效聚合了长时范围内的视频信息。Donahue等人(2015)则提出了长期循环卷积网络(long-term recurrent convolutional network, LRCN),其策略是首先使用卷积神经网络提取单帧图像的RGB特征,再将序列化的特征输入长短期记忆网络(long short-term memory, LSTM)单元(Hochreiter和Schmidhuber,1997),以捕获行为序列中的长程上下文依赖信息,进而完成行为分类。为更有效地建模时空关系,Shi等人(2015)设计了卷积长短时记忆网络(convolutional LSTM, ConvLSTM),该网络将传统LSTM门结构中的矩阵乘法替换为卷积运算,使其能够直接从多维数据(如图像序列)中学习基础的空间特征表示,并用于行为识别。此外,Tran等人(2015)提出的3D卷积网络(C3D)进一步提升了动作特征的提取能力,其核心在于使用三维卷积核直接在视频的时空维度上进行卷积操作,能够更加自然和高效地同时捕获视频中的空间信息和时间动态。在此基

础上，谭等泰等人(2020)采用双流 3D 卷积网络架构，分别提取视频的细节特征和整体特征，利用融合后的特征进行行为识别。近年来，随着 ViT(vision Transformer)在计算机视觉领域的广泛应用，王腾等人(2025)提出了一种融合全局与局部特征的两阶段 ViT行为识别算法，提升了行为识别的鲁棒性。

针对具体教学场景，廖鹏等人(2018)对 VGG(Visual Geometry Group)网络进行了改进，用于识别学生个体的站立、举手和听讲等动作。秦道影(2019)则结合了在ImageNet大规模数据集上预训练的图像分类网络和迁移学习技术，有效识别了教学环境中的多种常见行为。徐家臻等人(2020)提出了一种基于 Boosting 算法和卷积神经网络(convolutional neural network，CNN)的方法，专注于从单帧静态课堂场景图像中提取人体骨架信息，并据此实现学生行为的自动识别。

同时，大语言模型(large language model，LLM)的创新应用也在推动着行为识别技术的进步。Li 等人(2022)率先提出跨动作语义提示框架(bridge-prompt)，通过将教学视频中相邻动作标签重构为集成文本提示，显著提升复杂动作序列的理解能力。Yuan 等人(2022)进一步设计骨骼—文本联合编码模型(SkeletonCLIP)，利用 CLIP(contrastive language-image pre-training)提取动作语义特征并替代传统分类器，提升了识别准确率。而Leng等人(2023)开发的虚拟传感器框架(inertial measurement unit generative pre-trained Transformer，IMUGPT)则突破数据瓶颈，以ChatGPT(chat generative pre-trained Transformer)生成多样化活动文本描述，经 T2M-GPT(text-to-motion generative pre-trained Transformer)模型合成 3D 动作序列并转换为虚拟IMU(inertial measurement unit)数据，在减少真实数据依赖的同时提升模型泛化性。

然而，真实课堂场景的复杂性对现有技术提出了严峻挑战。如图 1 所示，一方面，随着教学改革的推进，新型教学行为不断涌现；另一方面，由于标注的成本问题，标注的样本量通常极其有限。这就要求模型必须具备少样本连续学习能力，能够在只有少量标注样本的条件下，持续学习增量任务中新出现的教学行为类别，并在此过程中保持对基类任务中已学旧类别的识别性能。

图 1　少样本连续教学行为识别任务示例

Fig. 1　Example of few shot continue teaching behavior recognition task

少样本连续学习通常包含一个基类任务和一系列增量任务。在基类任务阶段，模型利用充足的训练数据学习大量基础类别。而在随后的增量任务阶段，模型仅能利用极少量样本(通常每类仅 1～5 个)学习新类别。这种数据极度稀缺的设定带来了两大核心挑战：1)对新类过拟合的风险；2)对已学习旧类的灾难性遗忘(付浩 等，2025)。为了应对这些挑战，研究者们提出了多种方法。早期工作中，Tao 等人(2020)提出了 TOPIC(topology-preserving knowledge incrementer)框架，利用神经气体(neural gas，NG)网络学习类别特征空间的拓扑结构进行知识表示。Yang 等人(2021)设计了可学习的扩展和压缩网络(learnable expansionand-compression network，LEC-Net)，通过选择性扩展网络节点增强特征表示

能力，并引入模型正则化减少特征迁移干扰。Agarwal等人（2022）则提出了少样本增量学习生成对抗网络（few-shot incremental learning GAN，FSILGAN），利用GAN（generative adversarial network）合成旧类样本以缓解遗忘，其架构包含预训练的特征提取器、生成器、鉴别器和语义投影模块。Cheraghian等人（2021）提出了一种基于语义感知知识蒸馏的方法，将类别语义信息融入蒸馏过程，旨在更有效地保留旧类知识。

利用大规模预训练视觉语言模型（vision-language model，VLM），特别是CLIP（Radford等，2021），进行少样本连续学习展现出显著优势。D’Alessandro等人（2023）提出的CPE-CLIP（continual parameter-efficient contrastive language-image pre-training）是此类方法的典型代表，其核心思想在于充分挖掘并利用CLIP模型在海量图文数据上预训练所得的强大跨模态对齐知识。CLIP所蕴含的丰富视觉概念与语义理解能力，为新类别的识别提供了坚实的先验知识基础，有效缓解了增量任务中因样本稀缺导致的过拟合风险。但是，利用预训练的视觉语言模型进行少样本连续教学行为识别时，存在一个问题：现有视觉语言模型（如CLIP）对行为标签与其对应视频帧图像的匹配能力较弱。具体而言，模型对动词本身的语义细节及其所表征的动态变化捕捉不足。究其原因，行为识别中的原始标签通常为单个动词或简短动词短语，难以充分表征复杂动作的完整语义内涵，往往无法有效传达动作执行过程中涉及的主体、客体及其交互关系等关键信息。这种标签表示的局限性，限制了模型对行为动作的理解和泛化能力。

为了解决这个问题，本文提出了一种基于可伸缩思维链引导的少样本连续教学行为识别算法。该方法在冻结预训练视觉语言模型CLIP主干网络的基础上，仅通过提示学习与参数正则化进行轻量微调，并创新性地引入大语言模型（LLM）的思维链（chain-of-thought，CoT）机制（Wei等，2022）。该方法的核心在于对原始行为标签（通常为行为类别中未扩展的简洁动词或短语，如“writing”、“taking bag”）进行可伸缩的语义深度扩展。具体而言，它逐步生成包含丰富上下文信息的多层次描述性文本，从基础语义描述（如“A student writing in a notebook”）逐层递进至场景感知描述（如“A student writing in a crowded classroom，surrounded by other students，with a teacher lecturing on stage”）。之后，对大语言模型输入特定提示词，系统性地从扩展文本中提取（主语，谓语，宾语）结构的三元组知识表示。这一步骤使得文本信息更加凝练和结构化，突出行为中的关键实体及其相互关系。这种“先扩展语义广度（伸），后凝练核心结构（缩）”的可伸缩思维链引导机制，有效地增强了模型对教学行为的表征能力，并显著提升了在少样本条件下的识别性能。最后，本文方法建立了多层次跨模态匹配机制。该机制将图像的多层级视觉特征与前述生成的多层次文本三元组特征进行跨模态对齐，计算各层次特征对间的余弦相似度。融合（加权求和）这些层次化相似度得分后，将其作为分类得分（logits）输入至softmax函数，最终输出行为类别预测结果。

本文的主要贡献如下：1）创新性地将可伸缩思维链机制引入教学行为识别任务。该机制引导大语言模型（LLM）逐步挖掘原始行为标签中蕴含的深层动词语义，并将其凝练为结构化的（主语，谓语，宾语）三元组知识。这一过程显著增强了标签的语义表达，同时有效滤除了冗余信息，使模型能够精准聚焦于与行为相关的关键实体及其动作关系，从而大幅提升了文本模态表征的可解释性与类别区分能力。2）设计了基于预训练CLIP模型的多层次跨模态匹配框架。该框架同步提取图像与文本特征，并创新性地构建了层次化对齐机制：将图像不同层级的视觉特征与多层次文本三元组特征进行跨模态匹配，计算各层相似度后加权融合，最终完成识别。该方法有效提升了模型对复杂教学行为的识别精度与泛化性能。3）在具有32个行为类别的教室场景图像数据集ARIC（activity recognition in classroom）（Xu等，2024；Chen等，2025）上进行了充分的实验验证，实验结果表明，本文算法优于现有基线，取得了先进的性能。

1 本文算法

1.1 问题定义

少样本连续教学行为识别任务的设置定义如下：给定一个带有行为类别标签的训练集流$\boldsymbol{D}_0$，$\boldsymbol{D}_1$，…，$\boldsymbol{D}_T$，其中，$\boldsymbol{D}_t=\{(x_{i,t},y_{i,t})\}_{i=1}^{N_t^D}$，$N_t^D$是任务$t$的训练集样

本数，$x_{i,t}$和$y_{i,t}$分别是训练图像和其对应的行为类别标签。在T个任务中，$\boldsymbol{D}_0$表示基类的训练集，$\boldsymbol{D}_t(t>0)$则表示新类的数据集，在少样本连续行为识别中，基类训练的数据集规模较大，样本量充足，与之相反的是新类训练的样本量很少。对于训练集$\boldsymbol{D}_t$的类别标签集$\boldsymbol{C}_t$，同样设置如下：1）每个任务中样本的类别标签不重合，即满足$\forall t_1, t_2, t_1 \neq t_2, \boldsymbol{C}_{t_1} \cap \boldsymbol{C}_{t_2} = \varnothing$；2）基类训练集的类别数量要远大于新类训练阶段的类别数，即满足当$t>0$时，$|\boldsymbol{C}_0| > |\boldsymbol{C}_t|, N_0^D > N_t^D$；3）在新类训练阶段，每个任务的训练数据集大小相同，类别数和样本数保持一致，即满足$\forall t_1 > 0, t_2 > 0, t_1 \neq t_2, |\boldsymbol{C}_{t_1}| = |\boldsymbol{C}_{t_2}|, N_{t_1}^D = N_{t_2}^D$。

1.2 算法整体流程

真实世界教室场景中的行为类别通常以流式数据形式呈现，且各类别样本数量高度不均衡。同时，现有预训练视觉语言模型对动词的理解能力有限，难以有效识别复杂行为。为解决上述问题，本文提出了一种基于可伸缩思维链引导的少样本连续教学行为识别算法SCOTG，其流程如图2所示。

SCOTG算法采用预训练的视觉语言模型CLIP作为基础架构，充分利用其强大的跨模态对齐能力。CLIP的核心优势在于能够将视觉内容（图像/视频帧）与语义信息（文本描述）映射到统一的特征空间，并通过计算视觉特征嵌入与文本特征嵌入之间的余弦相似度实现高效的图像识别与分类。为了适应真实教室场景中流式、类别不均衡且需要持续学习新行为的挑战，并提升模型对复杂行为（尤其是动词理解）的识别能力，SCOTG算法在CLIP基础上引入了3个关键模块协同工作。

提示学习模块（详见第2.3节）通过在输入空间引入少量可学习的提示向量，分别引导视觉编码器和文本编码器关注任务相关的特征，实现高效的模型适配，避免了传统微调整个主干网络的高昂计算代价。

基于可伸缩思维链的行为标签语义挖掘模块（详见第2.4节）则针对CLIP对动词理解较弱及处理长文本能力的不足，利用大型语言模型（LLM）进行深度语义推理与结构化三元组关系抽取，将原始行为标签（如“使用电脑”）转化为蕴含丰富上下文关联和明确行为关系的高质量文本描述，极大地增强了模型对行为本质的理解。

最后，多层跨模态匹配模块（详见第2.5节）负责利用视觉编码器提取的多层次图像特征与经过语义增强后的文本特征进行跨模态相似度计算与融合，实现最终的行为识别分类。该模块能够有效捕捉行为在不同视觉层次上的表征，并与增强的语义信息进行精准匹配，从而提升整体识别性能。

1.3 提示学习模块

提示学习的核心在于仅优化少量的可学习提示向量，而非调整庞大的主干网络参数，在显著降低计算复杂度和存储需求的同时，能实现与全参微调相当的性能（Liu等，2022）。形式上，假设预训练模型参数为θ_0，可学习提示向量参数为φ，则训练目标可表示为

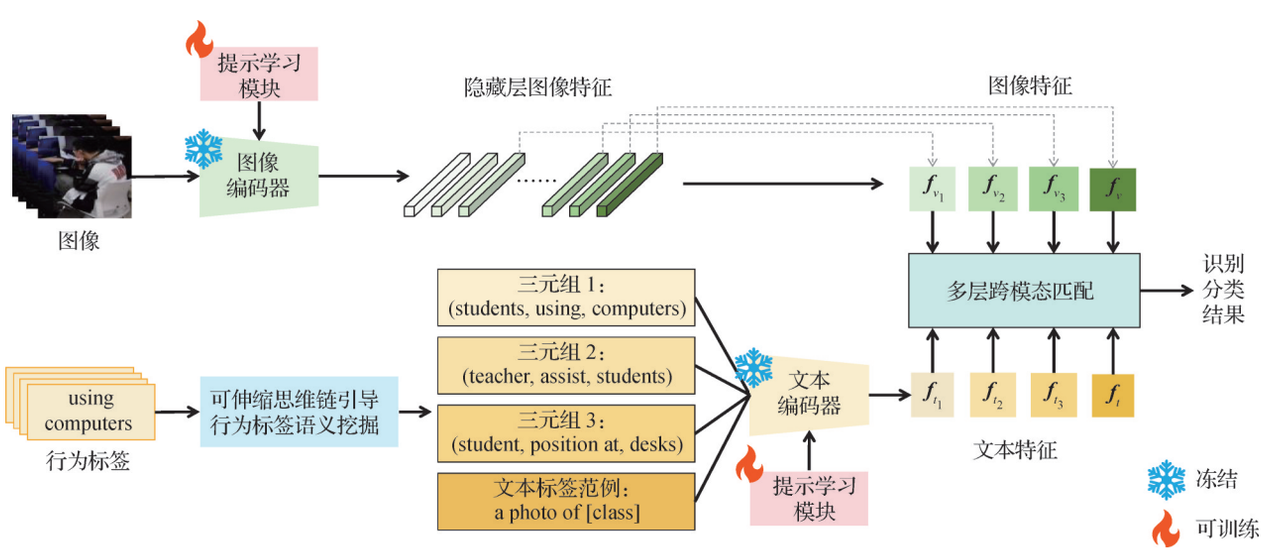

图2 基于可伸缩思维链引导的少样本连续教学行为识别算法

Fig. 2 The overview of scalable-chain-of-thought-guided few-shot continual teaching behavior recognition algorithm

$$\min_{\varphi} L(f(x;\theta_0,\varphi),y), \quad \text{s.t. } \theta_0 \text{ 固定} \tag{1}$$

与全参微调($\min_{\theta_0} L(f(x;\theta_0),y)$)相比,该约束优化问题在参数空间上显著收缩,仅在提示向量子空间中进行梯度更新,从理论上减少了优化过程的自由度和过拟合风险。

此外,通过冻结预训练主干网络,仅对小尺寸可学习提示向量进行微调的方式,能够有效缓解少样本连续学习场景下的灾难性遗忘问题,因为该策略既保留了主干网络已习得的通用特征提取能力,又通过局部参数更新实现了新知识的学习,从而在知识保留与增量学习间取得平衡。SCOTG算法的提示学习模块结构如图3所示。

具体而言,在文本编码器输入层引入可学习上下文提示向量$\boldsymbol{P}_t \in \mathbf{R}^{L\times d_t}$,与原始的文本嵌入向量$\boldsymbol{E}_t$连接在一起,构建成文本编码器的输入$\boldsymbol{I}_t=\{\boldsymbol{P}_t,\boldsymbol{E}_t\}=\{\boldsymbol{p}_{1t},\boldsymbol{p}_{2t},\cdots,\boldsymbol{p}_{Lt},\boldsymbol{E}_t\}$,其中$\boldsymbol{p}_{it}\in\mathbf{R}^{1\times d_t}$,$1\leqslant i\leqslant L$。$L$是文本提示向量的长度,$d_t$是CLIP文本编码器的嵌入向量维度。图像编码器的提示向量通过一个跨模态的线性投影层$\boldsymbol{f}_{\text{proj}}$生成,具体为

$$\boldsymbol{P}_v=\boldsymbol{f}_{\text{proj}}(\boldsymbol{P}_t) \tag{2}$$

通过将文本提示映射到图像嵌入空间,增强了图像与文本模态的跨模态一致性,从而更好地支持多模态少样本类别增量学习。获得的图像提示向量与图像嵌入向量$\boldsymbol{E}_v$连接在一起,构建成图像编码器的输入$\boldsymbol{I}_v=\{\boldsymbol{P}_v,\boldsymbol{E}_v\}=\{\boldsymbol{p}_{1v},\boldsymbol{p}_{2v},\cdots,\boldsymbol{p}_{Lv},\boldsymbol{E}_v\}$,其中$\boldsymbol{P}_v\in\mathbf{R}^{L\times d_v}$,$\boldsymbol{p}_{iv}\in\mathbf{R}^{1\times d_v}$,$1\leqslant i\leqslant L$,$L$是图像提示向量的长度,$d_v$是CLIP图像编码器的嵌入向量维度。

同时,为缓解遗忘,提示学习模块使用梯度缩放正则化,采用了基于已见类别数量的动态缩放因子,具体为

$$\alpha_t=\frac{|\boldsymbol{C}_t|}{\sum_{\tau=0}^{t}|\boldsymbol{C}_\tau|} \tag{3}$$

式中,$|\boldsymbol{C}_t|$是当前任务的类别数,$\sum_{\tau=0}^{t}|\boldsymbol{C}_\tau|$是模型已见的类别总数,通过缩放因子$\alpha_t$动态调整可学习参数的更新率,让已见类别越多,新任务的参数更新率越慢,平衡新知识学习与旧知识保留。具体而言,利用α_t对可学习的提示向量$\boldsymbol{P}_t$和线性投影层$\boldsymbol{f}_{\text{proj}}$的参数的梯度进行缩放,具体为

$$\frac{\partial L_t}{\partial \theta_{P_t}}=\alpha_t\cdot\frac{\partial L_{t-1}}{\partial \theta_{P_t}}(t\geqslant 1) \tag{4}$$

$$\frac{\partial L_t}{\partial \theta_{f_{\text{proj}}}}=\alpha_t\cdot\frac{\partial L_{t-1}}{\partial \theta_{f_{\text{proj}}}}(t\geqslant 1) \tag{5}$$

式中,L_t为任务t阶段的损失函数。

1.4 基于可伸缩思维链引导的行为标签语义挖掘

在教学行为识别研究中,原始的行为标签通常表现为单一的动词或简短的动词短语。然而,当前主流的预训练视觉语言模型大多在包含大量名词性概念的图像分类数据集上进行训练,对动词性行为语义的理解能力存在局限。值得注意的是,这些看似简单的行为标签本身蕴含着丰富的、多层次的语

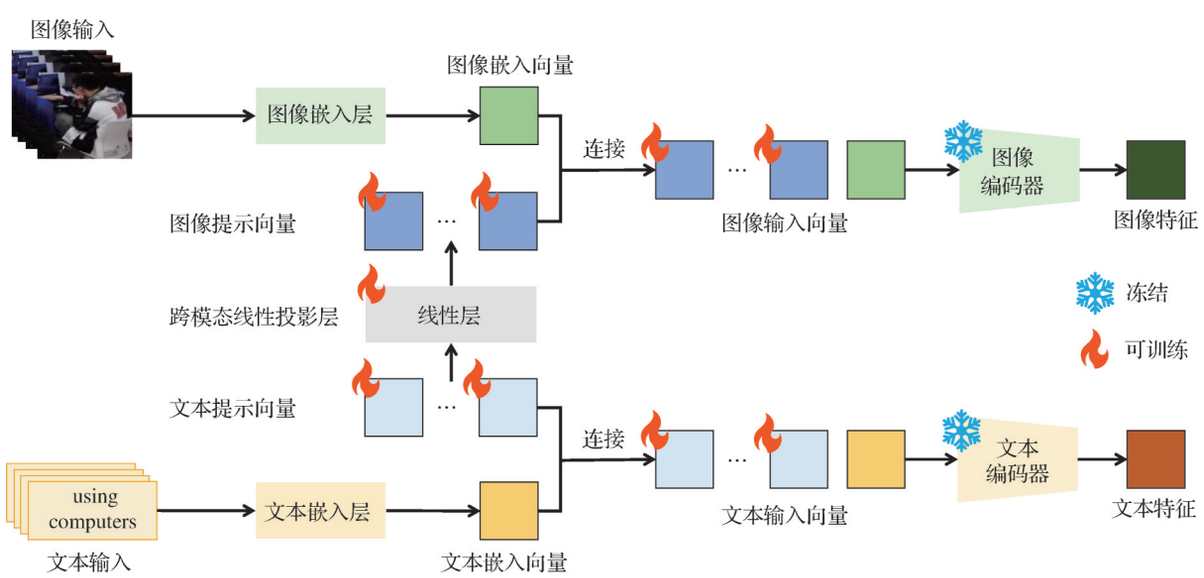

图3 提示学习模块

Fig. 3 prompt learning module

义信息。

以“写字”这一行为标签为例，从基础语义看，是指“某人通过手部握持书写工具在特定载体上进行符号记录的动作”，包含“握持—移动—留痕”的连续动作逻辑；从行为实体看，涉及学生主体，钢笔、笔记本和平板等客体，以及它们之间的相互作用关系，如“学生，用钢笔，在笔记本上”；从场景感知看，“写字”行为往往与特定的情境（如教室环境）、空间位置（如书桌）紧密关联，这些上下文细节共同构成了行为的完整表示。然而，现有算法多将“写字”这类标签作为单一文本符号与图像特征进行直接的浅层匹配。这种方法未能有效解析和利用行为标签内在的基础动作逻辑、实体间的交互关系以及场景上下文关联，严重限制了模型对行为深层内涵的捕捉与理解。这种理解的欠缺不仅导致模型对行为动作本身的辨识能力不足，更在少样本场景下显著削弱了模型的识别鲁棒性和泛化能力。

针对上述挑战，特别是预训练视觉语言模型（如CLIP）在动词理解能力薄弱、复杂行为表征受限以及长文本处理效率低下 3 个关键问题，提出的 SCOTG 算法创新性地引入大语言模型的强大语义生成与推理能力。SCOTG 的核心在于设计了一种可伸缩的思维链引导机制，旨在对原始行为标签进行深度语义挖掘与结构化扩展。其中“可伸缩”特性体现为动态调节语义信息的扩展（伸）与提炼（缩）。在需要丰富上下文时，能够扩展生成包含动作逻辑、实体交互和场景细节的多层次描述，在需要去除冗余信息时，则能提炼出最关键的语义要素。具体流程如图 4 所示，这种灵活的动态调节过程使得模型能够更精准、更高效地捕捉行为的本质特征，从而显著提升复杂行为的表征能力和少样本场景下的识别性能。

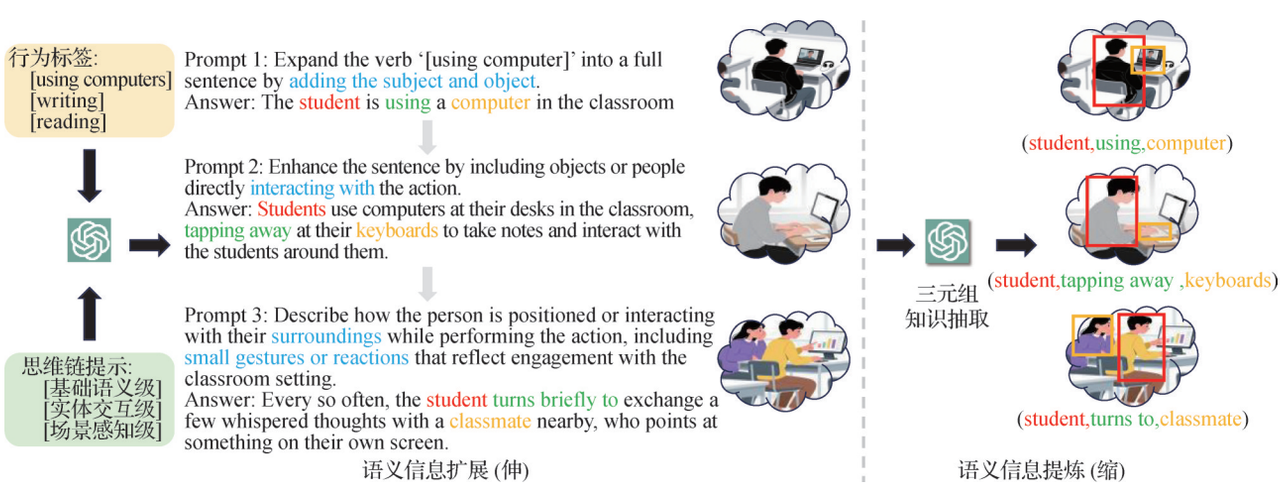

图 4　基于可伸缩思维链引导的行为标签语义挖掘流程

Fig. 4　The pipeline of chain-of-thought-guided behavior label semantic mining

为深度挖掘行为标签蕴含的多层次语义信息并构建丰富的上下文关联，SCOTG 首先采用基于思维链原理的三段式结构化问答流程，引导大语言模型生成针对原始行为标签的详尽描述性文本，实现语义的层次化扩展。

具体而言，在基础语义层级，设计结构化提示模板，要求大语言模型通过补充必要的行为主体、作用对象及空间位置等关键名词性成分，将简略的行为标签转化为一个语义完整的基础陈述句，从而明确动作的核心执行逻辑（加粗字体表示提示词中的重点核心内容），具体流程如下

$prompt_1$ = “Expand the verb ‘[ ]’ into a full sentence by **adding the subject and object.**”

$$text_1 = LLM\left(prompt_1\right) \tag{6}$$

随后，在实体交互层级，通过特定的提示词设计，引导大语言模型识别并引入与该行为密切相关的其他视觉实体（如工具、伴随物、受作用对象等），着重刻画主体与客体之间以及不同客体之间的动态交互关系（如“使用”、“操作”、“影响”），强化行为表征中的交互性特征，具体流程如下

$prompt_2$ = “Enhance the sentence by including objects or people directly **interacting with** the action.”

$$text_2 = LLM(prompt_2) \tag{7}$$

最终，在场景感知层级，进一步通过提示词驱动大语言模型生成包含细粒度环境要素与动态细节的场景描写，捕捉行为发生时的具体环境、空间布局、背景元素，以及主体在该场景下的细微动作变化或反应，实现对行为上下文更全面、更细致的语义补充，具体流程如下

$prompt_3$ = “Describe how the person is positioned or interacting with their **surroundings** while performing the action, including **small gestures or reactions** that reflect engagement with the classroom setting.”

$$text_3 = LLM(prompt_3) \tag{8}$$

在通过上述流程获取涵盖基础语义、实体交互及场景感知3个维度的多层次动词语义知识文本$text_1$、$text_2$、$text_3$后，SCOTG进一步利用大语言模型强大的信息抽取与结构化能力，从生成的描述性文本中精准提取核心语义单元，并将其凝练为（主，谓，宾）形式的结构化三元组知识，即$triple = (S, V, O)$。这个知识凝练过程有效滤除了原始描述文本中的冗余修饰信息和次要细节，将复杂的自然语言描述转化为高度结构化、语义明确的知识表示，具体流程如下

$$triple_1 = (S_1, V_1, O_1) = LLM(text_1) \tag{9}$$

$$triple_2 = (S_2, V_2, O_2) = LLM(text_2) \tag{10}$$

$$triple_3 = (S_3, V_3, O_3) = LLM(text_3) \tag{11}$$

这种三元组形式的凝练知识不仅具有表达简洁、关系清晰的优点，更重要的是，其高度结构化的特性显著降低了语义的复杂性，使之与预训练视觉语言模型CLIP的文本编码器处理机制更为兼容。通过规避对冗长、非结构化描述文本的直接处理，三元组知识能够更高效地被CLIP文本编码器编码为具有强判别性的特征向量，从而有效缓解CLIP模型在处理复杂长文本语义时可能面临的效率瓶颈与信息损失问题，为后续的跨模态匹配与行为识别奠定坚实基础。

1.5 多层跨模态匹配模块

SCOTG从图像编码器的不同隐藏层提取层级化图像特征$\boldsymbol{f}_{v_1}, \boldsymbol{f}_{v_2}, \boldsymbol{f}_{v_3}, \boldsymbol{f}_v$，并同步获取文本标签及其经由前述机制生成的多层次语义三元组的文本特征$\boldsymbol{f}_{t_1}, \boldsymbol{f}_{t_2}, \boldsymbol{f}_{t_3}, \boldsymbol{f}_t$，在此基础上设计了多层次跨模态匹配机制，旨在实现图像模态与文本模态在语义层级上的精细化对齐。该机制的核心在于，分别计算图像侧不同层级的图像特征与文本侧相应层次的语义扩展特征之间的余弦相似度，具体为

$$cosine\_sim(\boldsymbol{f}_v, \boldsymbol{f}_t) = \frac{\boldsymbol{f}_v \cdot \boldsymbol{f}_t}{\|\boldsymbol{f}_v\| \times \|\boldsymbol{f}_t\|} \tag{12}$$

具体而言，基础层视觉特征$\boldsymbol{f}_{v_1}$将与基础语义级文本特征$\boldsymbol{f}_{t_1}$进行匹配；中层视觉特征$\boldsymbol{f}_{v_2}$将与实体交互级文本特征$\boldsymbol{f}_{t_2}$进行匹配；高层视觉特征$\boldsymbol{f}_{v_3}$将与场景感知级文本特征$\boldsymbol{f}_{t_3}$进行匹配，而全局视觉特征$\boldsymbol{f}_v$则与原始的行为标签对应的文本特征$\boldsymbol{f}_t$进行匹配。这种分层、分级的匹配策略旨在精确捕捉不同抽象层级上的语义关联。

考虑到不同层级的图像特征和不同层次的语义扩展三元组文本特征对于最终行为识别任务的判别性贡献存在显著差异，SCOTG进一步引入层次加权融合策略。根据各层特征在行为识别任务中的贡献度设置不同的权重，得到最终用于分类的总相似度，计算式为

$$sim = cosine\_sim(\boldsymbol{f}_v, \boldsymbol{f}_t) + \alpha \sum_{i=1}^{3} cosine\_sim(\boldsymbol{f}_{v_i}, \boldsymbol{f}_{t_i}) \tag{13}$$

式中，$cosine_sim$表示余弦相似度的计算，α是用来平衡不同特征相似度影响的超参数。

在得到图像和所有类别文本特征的最终相似度后，将其作为分类得分（logits）输入至softmax函数，最终输出行为类别预测结果$\hat{y}$，具体为

$$\hat{y} = softmax(sim) \tag{14}$$

值得注意的是，从理论角度来看，SCOTG的跨模态匹配过程可以理解为一种最近邻分类（1-NN分类）机制：对于每个图像样本，其预测类别依赖于与文本特征空间中最相似样本的标签。定义最近邻标签与真实标签一致的概率为p，对于1-NN分类，其错误率可直接表示为$\varepsilon_{NN} = 1 - p$。依据Cover和Hart（1967）提出的最近邻误差上下界理论，1-NN分类器的错误率ε_{NN}满足$\varepsilon_{bayes} \leqslant \varepsilon_{NN} \leqslant 2\varepsilon_{bayes}$，其中$\varepsilon_{bayes} = 1 - P(y^*|x)$表示贝叶斯最优分类器的错误率，$y^* = \arg\max_y P(y|x)$为输入样本$x$的最优预测标签。$\varepsilon_{bayes}$是所有分类器可达到的理论误差下界，任何算法的错误率都无法低于该值。而对于输入x和它的最近邻x_{NN}，当训练样本量充分大时，x和x_{NN}的距离接近于0，则$P(y^*|x) = P(y^*|x_{NN})$，那么1-NN分类在x上的错误率可以表示为

$$\begin{aligned}\varepsilon_{\mathrm{NN}} &= P(y^*|x)(1-P(y^*|x_{\mathrm{NN}}))+\\ &P(y^*|x_{\mathrm{NN}})(1-P(y^*|x))=\\ &2P(y^*|x)(1-P(y^*|x))\leqslant\\ &2(1-P(y^*|x))=2\varepsilon_{\mathrm{bayes}}\end{aligned}\tag{15}$$

因此，1-NN 分类的错误率下界为贝叶斯最优错误率，上界为其两倍，且随着最近邻一致性提高，实际错误率将趋近该下界。

基于上述理论可知，提高最近邻标签一致性 p 能有效降低 1-NN 分类的错误率。SCOTG 通过将多层次视觉特征与不同层次的文本三元组相匹配，使得图像样本在特征空间中更容易与正确类别的文本样本靠近，如图 5 所示。这样，正确类别的文本样本更可能成为最近邻，而错误类别的干扰样本被有效区分，从而提高了最近邻标签一致概率 p，降低了 1-NN 的错误率 $\varepsilon_{\mathrm{NN}}$，使模型的错误率更接近贝叶斯最优下界。

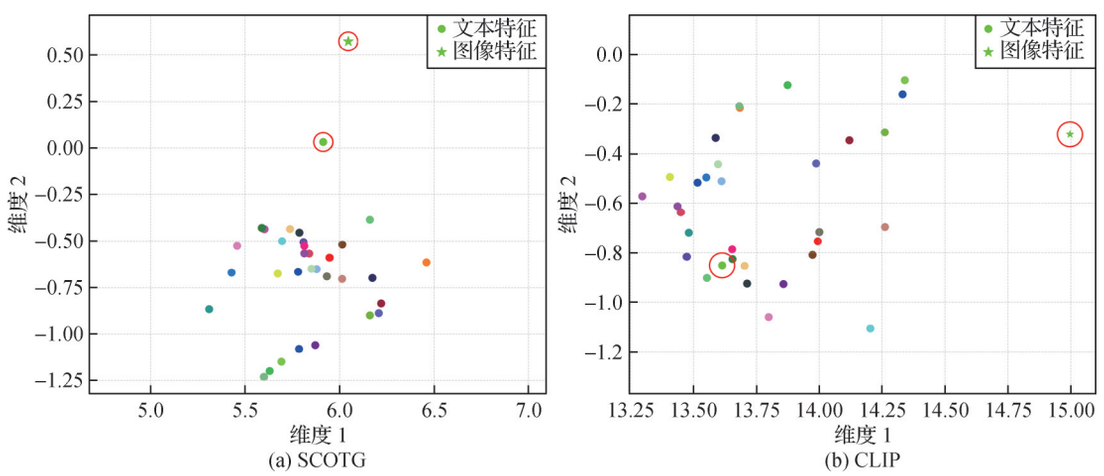

图 5　SCOTG 与 CLIP 的图文特征分布对比

Fig. 5　Comparison of text-image feature distribution between SCOTG and CLIP ((a) SCOTG; (b) CLIP)

最后，为实现模型的端到端优化，采用标准的交叉熵损失函数(cross-entropy loss)作为模型的整体优化目标。设样本的真实行为类别通过独热编码表示为 y，则交叉熵损失函数定义为

$$L=-\sum_{i=1}^{N}y_i\log(\hat{y}_i)\tag{16}$$

式中，N 为训练批次中的样本数量，y_i 和 $\hat{y}_i$ 分别表示第 i 个样本的真实标签与预测概率。该损失函数驱动模型学习最大化目标行为类别的相似度得分，同时最小化非目标类别的得分，从而有效地引导模型进行端到端的优化训练，提升模型在复杂行为识别任务上的判别精度与泛化能力。

2　实　验

2.1　实验设置

2.1.1　数据集

本文采用的教学行为识别图像数据集 ARIC (activity recognition in classroom)基于真实课堂监控场景构建(Xu 等，2024；Chen 等，2025)，涵盖 32 类课堂教学行为，如听讲(listening)、教学(teaching)、使用电脑(using computers)等。训练集中各类别的样本数量分布如图 6 所示。此外，为了验证使用大语言模型进行三元组提取的正确性，选用 WebNLG (web natural language generation)数据集(Gardent 等，2017)。该数据集包含从简单到复杂不同层级的三元组样本，且每个三元组均附带人工标注的自然语言描述文本，为三元组提取结果的正确性验证提供了明确的参照标准。

2.1.2　训练设置

本文采用 CLIP-ViT-B/16 作为主干网络，并在训练期间冻结其参数，所有实验框架均基于 PyTorch 实现。在基类训练阶段，模型学习包含 20 个行为类别的数据，批尺寸(batch size)和训练轮次(epoch)分别设置为 32 和 5。在新类训练阶段，采用少样本连续学习范式，分别评估了 3-way 5-shot、3-way 3-shot 及 3-way 1-shot 这 3 种设置(即每个任务引入 3 个新类别，每个类别分别提供 5、3 或 1 个样本)，该阶段批尺寸和训练轮次统一设置为 5。所有训练阶段均使用带动量的随机梯度下降(stochastic gradient

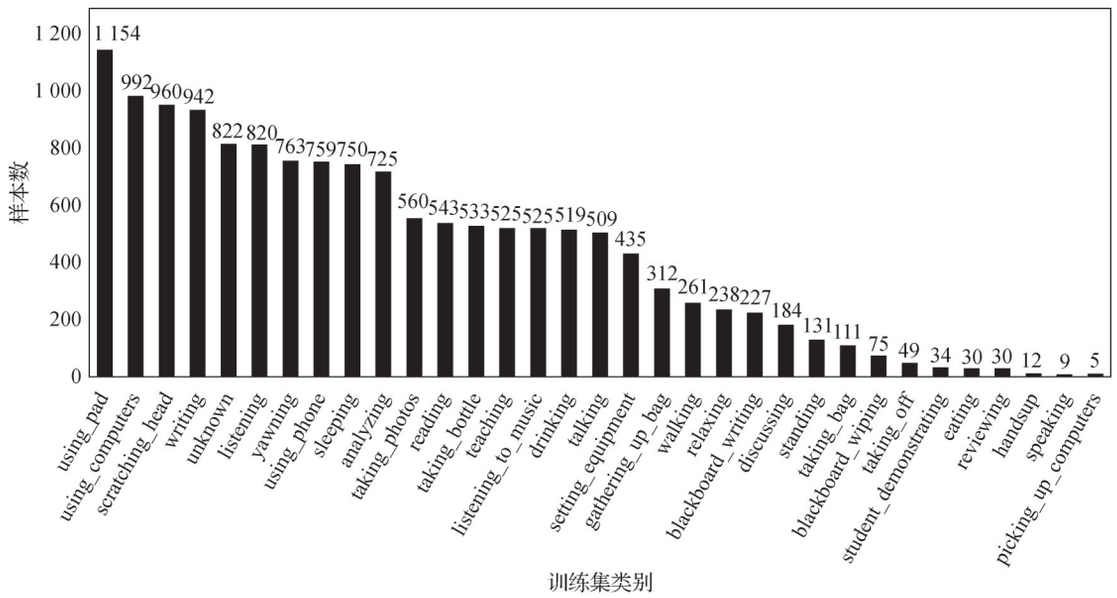

图6　教室场景数据集的训练集类别样本数分布

Fig. 6　Distribution of class samples in the train set of classroom scene dataset

descent, SGD)优化器,基础学习率(learning rate)设为 0.003 25,权重衰减(weight decay)系数为 1×10^{-5};学习率调度策略综合应用了余弦退火(cosine annealing)与热身重启(warmup restarts),且交叉熵损失函数引入了标签平滑(label smoothing)技术。对于多层特征的跨模态匹配阶段,本文利用验证集进行超参数α与视觉特征层数的选择,最终使用$\alpha = 0.1$,$L_1 = 6$,$L_2 = 9$,$L_3 = 11$,其中L_1、L_2和L_3这3层的视觉特征分别对应$\boldsymbol{f}_{v_1}$、$\boldsymbol{f}_{v_2}$和$\boldsymbol{f}_{v_3}$,最后一层输出的视觉特征即为$\boldsymbol{f}_v$。本文使用GPT-4-turbo模型(Open AI等,2023)对动词标签进行语义挖掘和三元组知识提取。

2.1.3　评估指标

对于少样本连续行为识别,主要有两个评估指标,一个是所有任务的平均识别准确率$\mathrm{ACC}_{\mathrm{avg}}$,另一个是最后一个任务的平均识别准确率$\mathrm{ACC}_{\mathrm{last}}$,分别展现出了模型的总体性能和缓解灾难性遗忘的能力。对于三元组提取,主要有精确率(precision)、召回率(recall)和F1值(F1-score)3个核心指标。

2.2　大语言模型提取三元组实验结果

为验证大语言模型在三元组提取任务中的性能,本实验采用GPT-4-turbo模型,在WebNLG数据集的验证集子集上开展三元组提取实验。实验中,模型输入为WebNLG数据集中的自然语言描述文本,输出为模型自动识别并结构化的三元组,提取指标结果如表1所示。

表1　大语言模型提取三元组实验结果

Table 1　Results of extracting triples from LLM

指标	结果/%
precision	80.13
recall	86.12
F1-score	81.91

从实验结果来看,GPT-4-turbo在WebNLG数据集的三元组提取任务中表现出较强的性能,其精确率(precision)、召回率(recall)和F1值(F1-score)都在80%以上,这一结果证明了大语言模型进行三元组提取的结果具备高度正确性,能够为后续任务提供可靠支撑。

2.3　少样本连续教学行为识别对比实验结果

本文在教室场景数据集ARIC上进行了少样本连续教学行为识别的实验,并与当前先进算法进行对比。在3-way 5-shot的实验设置下,本文方法显著优于基线模型,如表2所示,其中ACC_t表示任务t的识别准确率,$t = 0$表示基类任务,$t > 0$表示增量任务。为评估预训练CLIP模型的迁移能力,首先测试了其零样本(zero-shot)识别性能。如表2所示,CLIP模型在每个任务中的零样本识别准确率只有24.67% ± 4.33%,表明其语义对齐能力难以适应教学行为识别需求。为了进一步探究预训练CLIP模型在少样本连续教学行为识别任务中的适应性,对其主干网络进行了端到端的微调(fine-tuning)实验。

结果显示，虽然直接微调可以在基类任务达到较高精度77.79%，但随着增量任务的进行，模型的性能急速退化，最后任务的平均准确率相对于基类任务下降了15.21%，存在严重灾难性遗忘问题。相比较而言，本文算法最后任务的平均准确率相对于基类任务仅仅下降了6.11%，有效缓解了遗忘现象。

此外，如表2所示，与对比算法相比，在基类和新类训练的所有任务中，本文算法在教室场景数据集上都表现出了最好的性能，与最优的基线结果相比，本文算法的ACC_{avg}提升了1.98%，这展示了SCOTG算法对于提升模型对行为的理解识别能力有很大的帮助，同时，SCOTG算法在ACC_{last}上提升了1.36%，表明本文算法可以有效解决新类学习中的灾难性遗忘问题。

表2　3-way 5-shot设置下的对比实验结果

Table 2　Comparison experiment results of 3-way 5-shot setting

/%

算法	ACC_0	ACC_1	ACC_2	ACC_3	ACC_4	ACC_{avg}	ACC_{last}
CLIP-零样本	28.45	26.82	26.13	21.63	20.34	24.67	20.34
CLIP-微调	**77.79**	73.18	70.73	68.88	62.58	70.63	62.58
L2P（Wang等，2022b）	68.14	64.67	63.08	62.51	62.33	64.15	62.33
Dualprompt（Wang等，2022a）	71.19	67.56	65.98	65.50	65.29	67.10	65.29
Privilege（Park等，2024）	68.79	65.29	62.34	55.87	55.21	61.50	55.21
ZSCL（Zheng等，2023）	74.00	69.21	65.83	64.70	61.60	67.07	61.60
CPE-CLIP（D'Alessandro等，2023）	75.61	71.27	69.38	69.14	69.32	70.94	69.32
SCOTG（本文）	76.79	**74.09**	**71.84**	**71.23**	**70.68**	**72.92**	**70.68**

注：加粗字体表示各列最优结果。

本文SCOTG算法冻结了预训练CLIP模型的主干网络，只采用提示学习的方法训练轻量化的提示（prompt）向量，相比于微调全部的主干网络，所需要的计算资源大大降低，模型参数量对比如表3所示。

表3　模型参数量对比

Table 3　Comparison of model parameter quantity

模型	参数量/M
CLIP-微调	149.62
SCOTG（本文）	**0.41**

注：加粗字体表示最优结果。

实验进一步分析每个任务中的平均准确率，结果如图7所示。可以看到，无论是在基类任务还是增量任务阶段，SCOTG算法的识别准确率都要高于其他对比算法，而且随着任务的递进，SCOTG算法的性能衰减率也明显低于其他算法，这说明了SCOTG算法在帮助模型准确理解行为动作的同时，展现了优秀的知识迁移能力，有效平衡了新旧类别的识别，缓解了灾难性遗忘问题。

本文同样在3-way 3-shot和3-way 1-shot设置下进行了实验，结果如表4和表5所示。实验结果表明，与现有最优基线方法相比，SCOTG算法在3-way 3-shot设置中将平均识别准确率ACC_{avg}提升了1.03%；即使在样本数量更为匮乏的3-way 1-shot条件下，SCOTG算法依然保持了性能优势，平均识别

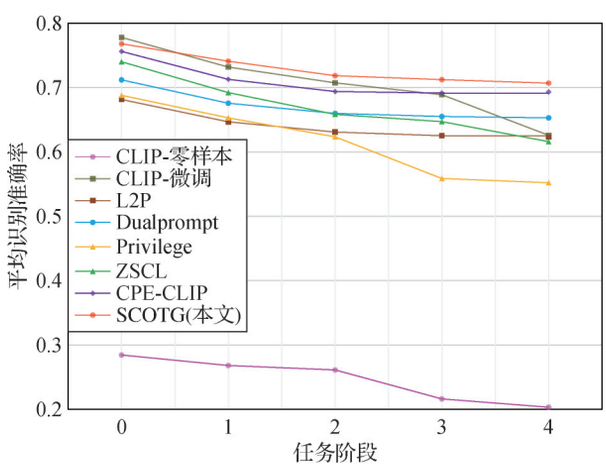

图7　3-way 5-shot设置下每个任务中的识别准确率对比

Fig. 7　Comparison of recognition accuracy in each task of 3-way 5-shot setting

表4　3-way 3-shot设置下的对比实验结果

Table 4　Comparison experiment results of 3-way 3-shot setting

/%

算法	ACC_0	ACC_1	ACC_2	ACC_3	ACC_4	ACC_{avg}	ACC_{last}
CLIP-零样本	28.45	26.82	26.13	21.63	20.34	24.67	20.34
CLIP-微调	**83.05**	**76.85**	65.87	59.10	45.12	66.00	45.12
L2P(Wang等,2022b)	68.14	64.74	63.17	62.59	62.35	64.20	62.35
Dualprompt(Wang等,2022a)	71.19	67.51	65.96	65.41	65.39	67.09	65.39
Privilege(Park等,2024)	68.79	65.78	61.78	52.55	47.59	59.30	47.59
ZSCL(Zheng等,2023)	75.03	70.63	67.31	66.17	63.41	68.51	63.41
CPE-CLIP(D'Alessandro等,2023)	75.47	71.80	70.34	69.78	**69.63**	71.41	**69.63**
SCOTG(本文)	77.17	73.75	**71.30**	**70.68**	69.28	**72.44**	69.28

注:加粗字体表示各列最优结果。

表5　3-way 1-shot设置下的对比实验结果

Table 5　Comparison experiment results of 3-way 1-shot setting

/%

算法	ACC_0	ACC_1	ACC_2	ACC_3	ACC_4	ACC_{avg}	ACC_{last}
CLIP-零样本	28.45	26.82	26.13	21.63	20.34	24.67	20.34
CLIP-微调	**83.39**	71.69	54.28	32.53	37.73	55.92	37.73
L2P(Wang等,2022b)	68.14	64.61	63.02	62.39	62.27	64.09	62.27
Dualprompt(Wang等,2022a)	71.19	67.54	65.94	65.47	65.26	67.08	65.26
Privilege(Park等,2024)	68.79	62.02	55.55	46.11	42.03	54.90	42.03
ZSCL(Zheng等,2023)	74.12	69.25	66.22	64.94	61.82	67.27	61.82
CPE-CLIP(D'Alessandro等,2023)	76.71	72.41	69.91	69.25	66.78	71.01	66.78
SCOTG(本文)	77.25	**73.68**	**70.36**	**69.78**	**67.89**	**71.79**	**67.89**

注:加粗字体表示各列最优结果。

准确率ACC_{avg}提升了0.78%。这些结果充分证明了SCOTG算法在不同样本规模的小样本学习任务中均能带来稳定且具有竞争力的性能提升,凸显了其良好的泛化能力和鲁棒性。

2.4　消融实验结果

为了评估本文算法中各组件的影响,设计系统的消融实验,这些消融实验都是在3-way 5-shot的少样本连续学习设置下进行的,实验结果如表6所示。首先,不通过思维链逐步扩展行为标签,而是只让LLM对行为标签进行单次浅层语义扩展,然后提取扩展上下文中的三元组,与视觉特征做跨模态匹配,得到识别结果。如表6所示,与本文完整算法相比,不使用思维链的简化算法在ACC_{avg}上降低了2.23%,在ACC_{last}上更是显著下降了5.22%,这说明单一层次的语义扩展没有充分挖掘行为标签中隐含的动词语义信息和上下文关联信息,难以真正帮助模型理解动作和减少灾难性遗忘。之后,为了验证使用三元组关系抽取进行知识凝练这一步骤的有效性,实验去掉这一组件,只通过思维链进行行为标签的联想扩充,实验结果如表6所示,与完整算法相比,缺少知识凝练的简化算法在ACC_{avg}和ACC_{last}上都略有降低,分别损失了0.48%和0.68%的性能,这是因为三元组知识抽取这一操作可以帮助去除行为标签上下文中的冗余信息,得到更精炼的结构化知识,让模型更准确地理解行为细节。此外,本文验证了所提出的多层次跨模态匹配机制的有效性,在对行为标签进行联想——凝练这一可伸缩操作的基础上,实验让所有的文本特征都只与最后一层的视

觉特征做跨模态匹配，与使用多层特征的完整算法相比，如表 6 所示，这一简化算法在 ACC_{avg} 降低了 1.57%，在 ACC_{last} 上更是损失了 4.09% 的性能。实验结果表明，本文设计的多层次跨模态匹配机制可以在不同尺度上捕获行为细节，增强模型对行为动作的识别能力，让模型更好地识别复杂的行为细节。

表 6　消融实验结果

Table 6　Results of ablation experiment

/%

模型	ACC_{avg}	ACC_{last}
不使用思维链	70.69	65.46
不提取三元组	72.44	70.00
单层特征	71.35	66.59
SCOTG（本文）	**72.92**	**70.68**

注：加粗字体表示各列最优结果。

最后，为探究不同大语言模型基线对实验结果的影响，分别选取 GPT-3.5-turbo、GPT-4 和 GPT-4-turbo 这 3 个版本的模型进行对比实验。实验结果如表 7 所示。分析可知，相较于使用 GPT-3.5-turbo 作为基线模型，采用 GPT-4-turbo 时，平均识别准确率提升了 1.59%。结果表明，LLM 生成内容质量的改进，能够进一步提升 SCOTG 算法的识别性能。

表 7　不同大语言模型基线的消融实验结果

Table 7　Results of ablation experimental of different LLMs baselines

/%

模型	ACC_{avg}	ACC_{last}
GPT-3.5-turbo	71.33	68.61
GPT-4	71.99	69.58
GPT-4-turbo	**72.92**	**70.68**

注：加粗字体表示各列最优结果。

3　结　论

本文提出了一种基于可伸缩思维链引导的少样本连续教学行为识别算法 SCOTG，旨在解决教室场景中教学行为识别的关键挑战。针对该场景中普遍存在的各类行为样本数量严重不平衡问题，首先引入基于预训练视觉语言模型 CLIP 的少样本连续学习框架，有效缓解了数据匮乏带来的影响。进一步地，针对现有预训练视觉语言模型对行为动词理解能力不足、难以精准匹配图像与动词标签的问题，本文创新性地提出利用大语言模型（LLM），通过可伸缩思维链引导机制，深度挖掘行为标签隐含的多层次语义知识。该机制能够将复杂的行为语义凝练为结构化的（主，谓，宾）三元组知识。随后，算法通过将这些语义三元组分别与视觉主干网络提取的多层次视觉特征进行精细化跨模态匹配，实现了对教学行为的更准确理解和识别。在具有 32 个行为类别的教室场景图像数据集 ARIC 上进行的广泛实验表明，与当前先进的行为识别算法相比，SCOTG 算法在多个评估指标上均取得了优异的性能，充分验证了所提出方法在处理少样本连续教学行为识别任务上的有效性与优越性。

然而，本文方法也存在一定的局限性：SCOTG 对 LLM 的依赖性较重，LLM 在扩展和提炼行为标签语义知识过程中的表现（如生成内容的准确性、相关性与一致性）会显著影响最终的行为识别效果。未来的研究工作将重点聚焦于降低 LLM 引入的噪声干扰并提升模型鲁棒性。

参考文献（References）

Agarwal A, Banerjee B, Cuzzolin F and Chaudhuri S. 2022. Semantics-driven generative replay for few-shot class incremental learning// Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portugal: ACM: 5246-5254 [DOI: 10.1145/3503161.3548160]

Chen Y K, Qiu Z H, Meng F M, Li H L, Xu L F and Wu Q B. 2025. Leveraging pre-trained models for multimodal class-incremental learning under adaptive fusion//Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad, India: 1-5 [DOI: 10.1109/ICASSP49660.2025.10888210]

Cheraghian A, Rahman S, Fang P, Roy S K, Petersson L and Harandi M. 2021. Semantic-aware knowledge distillation for few-shot class-incremental learning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2534-2543 [DOI: 10.1109/CVPR46437.2021.00256]

Cover T and Hart P. 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13 (1): 21-27 [DOI: 10.1109/TIT.1967.1053964]

D'Alessandro M, Alonso A, Calabrés E and Galar M. 2023. Multimodal

parameter-efficient few-shot class incremental learning//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3385-3395 [DOI: 10.1109/ICCVW60793.2023.00364]

Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Darrell T, et al. 2015. Long-term recurrent convolutional networks for visual recognition and description//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 2625-2634 [DOI: 10.1109/CVPR.2015.7298878]

Fu H, Feng Q, Tu J H, Zhao H B, Zhang C, Du X, et al. 2025. Advances in incremental learning research. Journal of Image and Graphics, 30(6): 1690-1716 (付浩, 冯前, 涂嘉航, 赵涵斌, 张超, 杜歆, 等. 2025. 增量学习研究进展. 中国图象图形学报, 30(6): 1690-1716) [DOI: 10.11834/jig.240790]

Gardent C, Shimorina A, Narayan S and Perez-Beltrachini L. 2017. Creating training corpora for NLG micro-planners//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: ACL: 179-188 [DOI: 10.18653/v1/P17-1017]

Gu C H and Li Y X. 2022. Analysis of art classroom teaching behavior based on intelligent image recognition. Mobile Information Systems, 2022: #5736407 [DOI: 10.1155/2022/5736407]

Hochreiter S and Schmidhuber J. 1997. Long short-term memory. Neural Computation, 9(8): 1735-1780 [DOI: 10.1162/neco.1997.9.8.1735]

Jia Q and He J. 2024. Student behavior recognition in classroom based on deep learning. Applied Sciences, 14(17): #7981 [DOI: 10.3390/app14177981]

Leng Z K, Kwon H and Plötz T. 2023. Generating virtual on-body accelerometer data from virtual textual descriptions for human activity recognition//2023 ACM International Symposium on Wearable Computers. Cancun, Mexico: ACM: 39-43 [DOI: 10.1145/3594738.3611361]

Li M H, Chen L, Duan Y Q, Hu Z L, Feng J J, Zhou J, et al. 2022. Bridge-prompt: towards ordinal action understanding in instructional videos//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 19848-19857 [DOI: 10.1109/CVPR52688.2022.01926]

Liao P, Liu C M, Su H, Li Q F and Han Y J. 2018. Student classroom abnormal behavior detection and analysis system based on deep learning. Electronics World, (8): 97-98 (廖鹏, 刘宸铭, 苏航, 李启芳, 韩延巾. 2018. 基于深度学习的学生课堂异常行为检测与分析系统. 电子世界, (8): 97-98) [DOI: 10.19353/j.cnki.dzsj.2018.08.054]

Liu Q T, Jiang X Y and Jiang R Y. 2025. Classroom behavior recognition using computer vision: a systematic review. Sensors, 25(2): #373 [DOI: 10.3390/s25020373]

Liu X, Ji K X, Fu Y C, Tam W, Du Z X, Yang Z L, et al. 2022. P-tuning: prompt tuning can be comparable to fine-tuning across scales and tasks//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland: ACL: 61-68 [DOI: 10.18653/v1/2022.acl-short.8]

OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. 2023. Gpt-4 technical report [EB/OL]. [2025-07-01]. https://arxiv.org/pdf/2303.08774.pdf

Park K H, Song K and Park G M. 2024. Pre-trained vision and language transformers are few-shot incremental learners//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 23881-23890 [DOI: 10.1109/CVPR52733.2024.02254]

Qin D Y. 2019. Student Classroom Behavior Recognition Based on Deep Learning. Wuhan: Central China Normal University (秦道影. 2019. 基于深度学习的学生课堂行为识别. 武汉: 华中师范大学)

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 8748-8763

Shi X J, Chen Z R, Wang H, Yeung D Y, Wong W K and Woo W C. 2015. Convolutional LSTM network: a machine learning approach for precipitation nowcasting//Proceedings of the 29th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 802-810

Simonyan K and Zisserman A. 2014. Two-stream convolutional networks for action recognition in videos//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 568-576

Tan D T, Li S C, Chang W W and Li D L. 2020. Multi-feature fusion behavior recognition model. Journal of Image and Graphics, 25(12): 2541-2552 (谭等泰, 李世超, 常文文, 李登楼. 2020. 多特征融合的行为识别模型. 中国图象图形学报, 25(12): 2541-2552) [DOI: 10.11834/jig.190637]

Tao X Y, Hong X P, Chang X Y, Dong S L, Wei X and Gong Y H. 2020. Few-shot class-incremental learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 12180-12189 [DOI: 10.1109/CVPR42600.2020.01220]

Tran D, Bourdev L, Fergus R, Torresani L and Paluri M. 2015. Learning spatiotemporal features with 3D convolutional networks//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 4489-4497 [DOI: 10.1109/ICCV.2015.510]

Wang L M, Xiong Y J, Wang Z, Qiao Y, Lin D H, Tang X O, et al. 2016. Temporal segment networks: towards good practices for deep action recognition//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 20-36 [DOI: 10.1007/978-3-319-46484-8_2]

Wang S C, Huang Q, Zhang Y F, Li X, Nie Y Q and Luo G C. 2022. Review of action recognition based on multimodal data. Journal of Image and Graphics, 27(11): 3139-3159 (王帅琛, 黄倩, 张云飞, 李兴, 聂云清, 雒国萃. 2022. 多模态数据的行为识别综述. 中国图象图形学报, 27(11): 3139-3159) [DOI: 10.11834/jig.210786]

Wang T, Gao S B and Ren G. 2025. Two-stage vision transformer for fusing global and local features in distracted driving behavior recognition. Journal of Image and Graphics, 30(11): 3617-3633 (王腾, 高尚兵, 任刚. 2025. 融合全局与局部特征的两阶段 ViT 分心驾驶行为识别方法. 中国图象图形学报, 30(11): 3617-3633) [DOI: 10.11834/jig.240533]

Wang Z F, Zhang Z Z, Ebrahimi S, Sun R X, Zhang H, Lee C Y, et al. 2022a. DualPrompt: complementary prompting for rehearsal-free continual learning//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 631-648 [DOI: 10.1007/978-3-031-19809-0_36]

Wang Z F, Zhang Z Z, Lee C Y, Zhang H, Sun R X, Ren X Q, et al. 2022b. Learning to prompt for continual learning//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 139-149 [DOI: 10.1109/CVPR52688.2022.00024]

Wei J, Wang X Z, Schuurmans D, Bosma M, Ichter B, Xia F, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #180

Xu J Z, Deng W and Wei Y T. 2020. Automatic recognition of student's classroom behaviors based on human skeleton information extraction. Modern Educational Technology, 30(5): 108-113 (徐家臻, 邓伟, 魏艳涛. 2020. 基于人体骨架信息提取的学生课堂行为自动识别. 现代教育技术, 30(5): 108-113) [DOI: 10.3969/j.issn.1009-8097.2020.05.016]

Xu L F, Meng F M, Wu Q B, Pan L L, Qiu H Q, Wang L X, et al. 2024. ARIC: an activity recognition dataset in classroom surveillance images [EB/OL]. [2025-07-01]. https://arxiv.org/pdf/2410.12337.pdf

Yang B Y, Lin M B, Liu B H, Fu M Y, Liu C, Ji R R, et al. 2021. Learnable expansion-and-compression network for few-shot class-incremental learning [EB/OL]. [2025-07-01]. https://arxiv.org/pdf/2104.02281.pdf

Yuan L, He Z, Wang Q, Xu L Y and Ma X. 2022. SkeletonCLIP: recognizing skeleton-based human actions with text prompts//Proceedings of the 8th International Conference on Systems and Informatics (ICSAI). Kunming, China: IEEE: 1-6 [DOI: 10.1109/ICSAI57119.2022.10005459]

Zheng Z W, Ma M Y, Wang K, Qin Z H, Yue X Y and You Y. 2023. Preventing zero-shot transfer degradation in continual learning of vision-language models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 19068-19079 [DOI: 10.1109/ICCV51070.2023.01752]

作者简介

赵珞君，女，硕士研究生，主要研究方向为少样本连续学习行为识别。E-mail：202321011839@std.uestc.edu.cn

吴庆波，通信作者，男，教授，主要研究方向为视觉质量评价与增强、图像视频编码、多模态目标检测与分割。E-mail：qbwu@uestc.edu.cn

刘小同，女，硕士研究生，主要研究方向为多模态图像质量评估。E-mail：202321011828@std.uestc.edu.cn

孟凡满，男，教授，主要研究方向为智能图像分析和深度学习。E-mail：fmmeng@uestc.edu.cn