基于VLM凸优化的网络直播视频场景图生成

李文生[1],张 菁[1,2*],王艺晓[1],卓 力[1,2]
(1. 北京工业大学信息科学技术学院,北京 100124;2. 北京工业大学计算智能与智能系统北京市重点实验室,北京 100124)

摘 要: 网络直播视频平台凭借庞大的主播群体、海量的内容供给以及极高的日活跃用户规模,已经成为当下数字内容传播、社交互动与商业转化的核心载体。然而直播内容的实时动态性和不可预测性,为网络内容监管带来严峻挑战。视频场景图作为一种能够刻画视频中对象、属性及行为关系的结构化表示方式,通过在时空维度上构建“对象—关系—行为”的语义网络,可实现视频内容的结构化表征。近年来,视觉语言模型(Visual-Language Models,VLMs)在跨模态特征语义理解与复杂场景推理方面展现出显著优势,为直播视频场景图生成提供了新的技术支撑。值得注意的是,VLM虽能提升复杂直播场景的语义解析精度,但仍需克服直播视频特征分布规律不易挖掘的瓶颈问题。在VLM模型训练过程中,凸函数优化对驱动模型收敛至全局最优解至关重要,提出了一种基于VLM凸优化的网络直播视频场景图生成方法(VLM-based Convex Optimization for Scene Graph Generation,VCO-SGG)。该方法构建VLM近似凸优化架构,通过优化对象语义及其关联关系的特征空间几何结构,缩小特征分布差异,缓解VLM模型在训练过程中的收敛震荡问题;同时,构建动态原型记忆模块,通过参数化记忆机制增强对视频帧间关键语义元素持续性与关联性的记忆能力;此外,提出特征联合与关系筛选策略,在线识别并过滤场景图中由动态变化产生的冗余对象索引,实现场景图的动态生成与更新。实验结果表明,该方法在自建直播视频数据集BJUT-LGSD上R@10与mR@10分别提升至55.41%与34.82%;在公开数据集Mini Charades和Mini Action Genome上R@10和mR@10分别达到48.19%/28.02%、43.42%/26.02%;推理速度保持在22.36 FPS,较现有对比方法更具竞争力,表明了其可以胜任直播视频场景图的生成任务。

关键词: 网络直播视频;场景图生成;视觉语言模型;凸优化;动态原型记忆;特征联合与关系筛选
基金项目: 国家自然科学基金(No.61971016,No.62471013);北京市自然科学基金(No.KZ201910005007)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2026)02-0544-18
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250586

Scene Graph Generation of Livestreaming Video via VLM Convex Optimization

LI Wensheng[1], ZHANG Jing[1,2*], WANG Yixiao[1], ZHUO Li[1,2]
(1. *School of Information Science and Technology*, *Beijing University of Technology*, *Beijing* 100124, *China*;
2. *Beijing Key Laboratory of Computational Intelligence and Intelligent System*, *Beijing University of Technology*, *Beijing* 100124, *China*)

Abstract: Livestreaming video platforms have become an important medium for digital content dissemination, social interaction, and commercial activities. This is largely due to their large number of streamers, massive content supply, and extremely high daily active user base. However, the real-time and unpredictable nature of livestreaming content poses serious challenges for online content supervision and regulation. Video scene graphs provide a structured representation for video understanding. They describe objects, attributes, and behavioral relationships within videos. By constructing a semantic network of “object-relation-action” in the spatiotemporal domain, video scene graphs enable structured modeling of video content. In recent years, vision-language models (VLMs) have shown strong capabilities in cross-modal semantic understanding and complex scene reasoning. These advantages provide new technical support for livestreaming video scene graph generation. Although VLMs can significantly improve semantic parsing accuracy in complex livestreaming scenarios, they still face an important challenge. Specifically, it is difficult to effectively capture the feature distribution patterns of livestreaming videos. Convex optimization plays an important role in training VLMs. It helps guide the model to converge toward a global optimal solution. Based on this observation, this paper proposes a VLM-based convex optimization for scene graph generation (VCO-SGG). The method constructs a VLM-based approximately convex optimization framework that constrains the geometric structure of the feature space for object semantics and their relationships, reducing feature distribution discrepancies and mitigating convergence oscillations during VLM training. A dynamic prototypical memory module is in-

收稿日期:2025-07-02;录用日期:2026-02-24;责任编辑:朱梅玉
*通信作者:张菁

troduced, employing a parametric memory mechanism to strengthen the memory of key semantic elements' continuity and correlations across video frames. Furthermore, a feature association and relation filtering strategy is proposed to identify and filter redundant object indices online, which are generated in the scene graph due to dynamic changes, thereby enabling dynamic generation and updating the scene graph. Experimental results demonstrate that our method achieves improvements of R@10 and mR@10 reaching 55.41% and 34.82%, on the self-built livestreaming video dataset BJUT-LGSD, respectively. In the publicly available datasets Mini Charades and Mini Action Genome datasets, R@10 and mR@10 are further improved to 48.19%/28.02% and 43.42%/26.02%, respectively, and the inference speed is 22.36 FPS. Overall, the results demonstrate greater competitiveness than other methods, indicating its capability to handle the task of generating scene graphs for livestreaming videos.

Keywords: livestreaming video; scene graph generation; vision-language models; convex optimization; dynamic prototype memory; feature association and relation filtering strategy

Foundation Item(s): National Natural Science Foundation of China (No.61971016, No.62471013); Beijing Natural Science Foundation (No.KZ201910005007)

0 引言

随着社交媒体和专业直播平台的爆发式发展,网络直播(livestreaming)已成为数字内容传播体系的核心载体,日均活跃用户规模呈指数级增长。在此背景下,直播内容的实时监管愈发迫切,亟需通过智能化技术手段构建高效、规范化的内容检测体系[1-2]。动态场景图生成(Dynamic Scene Graph Generation, DSGG)作为视频语义理解的关键技术,通过识别出对象及其行为关系实现直播场景的结构化表征,为优化直播内容监督提供了有潜力的技术手段。将场景图生成技术引入直播视频领域,不仅是计算机视觉向动态场景的自然延伸,也具有重要的科学与产业价值。与传统视频不同,直播的核心价值在于实时交互与语义关联。例如在电商直播中,仅识别"主播"和"商品"无法揭示带货逻辑,而场景图生成能够提取<主播-试用-商品>等关系三元组,进而理解深层交互意图。在内容审核方面,相比基于关键词或物体检测的方法,场景图生成能依据"关系"进行违规行为区分(如<人-攻击-人> vs <人-拥抱-人>),显著降低误判率。此外,场景图生成的结构化输出支持基于关系的细粒度检索(如检索"所有展示鞋子的片段"),可用于自动化生成直播精彩集锦,提升长视频分发效率。这种结构化语义理解对于解析直播叙事逻辑、捕捉关键事件至关重要,是实现全面视频理解的关键一步。因此,开展直播视频场景图生成研究,既是对高动态视觉理解理论的重要拓展,也是推动直播行业向智能化升级的关键技术路径。最近的视觉大语言模型(Vision-Language Models, VLMs)在各种计算机视觉任务中表现优异,利用其跨模态语义融合能力,有望显著提升复杂直播场景的语义解析精度,为场景图生成任务提供新的技术途径。VLM预训练模型一般利用凸优化(convex optimization)方法,将视频数据集上场景图生成任务转化为模型优化问题,然而,网络直播视频场景图生成将面临以下挑战。

以图1为例,直播过程中往往会有一些不可预测的内容,如图1(a)中行人突然闯入镜头范围,这种不可预测性会导致凸函数优化过程中不易收敛至全局最优解。在VLM模型训练过程中,凸函数优化对驱动模型收敛至全局最优解至关重要。然而,直播视频特征分布模态偏移问题,会导致凸优化过程中梯度迭代路径剧烈震荡,引发收敛不稳定现象。现有VLM凸优化通过假设模型参数服从特征先验分布来稳定优化过程[3],其依赖数学假设构建参数分布,面对内容不可预测的直播视频时,常出现预设分布与真实特征分布的显著偏差。针对这一特性,可以采用自适应分布建模机制构建凸函数初始参数分布,以克服VLM凸优化的梯度迭代路径震荡引发的收敛不稳定问题。

其次,相比普通视频,直播视频具有较强的在线互动性,如图1(b)中主播需要时常观察弹幕评论与观众交流,时而停顿,因而导致时序不够连贯性,即出现非连续跳转,因此时序信息建模成为关键挑战。记忆智能体通过实时更新记忆库建模视频时序信息,实现在当前帧特征中补充历史信息[4-5]。但在复杂直播场景中,固定的记忆更新策略难以适应突发互动导致的内容突变,影响场景图生成对象间动作关系的准确性。为此,可以根据历史帧特征提炼原型索引来降低直播互动对建模时序逻辑关系的影响,以提升场景图生成准确性。

随走随播是直播视频的特有属性,而直播中的内容及数量会无规律波动,如图1(c)的商铺和行人随着镜头的移动不断变化,在关系建模过程中容易引入噪声,导致特征联合过程易受冗余信息干扰,影响了场景图生成的鲁棒性。现有研究中通过预设查询(Query)作为对象索引,来定位关键对象,但固定的Query设计无法适应对象数量的动态变化,在高密度对象场景中会出现特征遗漏[6-8]。为此,本方法根据

(a) 内容不可预测 (a) Unpredictable content

(b) 缺乏时序连贯性 (b) Lack of temporal coherence

(c) 冗余信息干扰 (c) Interference from redundant information

图1 直播视频样例展示

Figure 1 Examples of livestreaming videos

直播视频中对象数量的随机变化筛选冗余的Query，以缓解动态对象及其关系建模的影响。

综上，本研究提出一种基于VLM凸优化的网络直播视频场景图生成方法（VLM-based Convex Optimization for Scene Graph Generation，VCO-SGG）。首先构建融合VLM与凸函数优化策略的专用架构，解决预训练数据与直播视频的特征分布差异问题；其次，设计动态原型记忆，捕获直播视频的时序上下文依赖关系；最后，通过特征联合与关系筛选策略，结合基于凸优化目标的对象关系优化机制，强化场景中有效对象的关联建模，优化场景图的生成。因此，本文的主要创新点总结如下：

（1）针对直播视频与预训练数据在场景结构、对象类型上的特征分布差异，提出专用VLM大模型凸优化架构。通过构建自适应特征建模体系，实现对象间语义关联、运动轨迹等关键特征的建模，克服因数据分布差异导致的模型适配难题。

（2）鉴于直播视频时序逻辑具有跳跃性的特点，设计动态原型记忆，强化对视频帧中关键元素的记忆关联，通过构建代表性记忆原型库，对当前帧的上下文特征提取提供指导优化，有效建模直播视频的时序逻辑关系。

（3）针对直播过程中对象状态的无规律波动，提出特征联合与关系筛选策略，高效识别并过滤冗余对象索引。通过降低冗余信息的干扰，提升模型对场景图的生成能力，确保模型在复杂多变的直播场景中仍具备良好的稳健性。

1 相关工作

根据本文基于VLM的直播视频场景图生成的研究任务，相关工作将聚焦场景图生成方法、动态原型记忆智能体、特征联合三个方面。

1.1 场景图生成方法

在训练场景图生成模型中，提升泛化性能与鲁棒性需要使参数分布适应目标数据特征。现有的研究主要遵循两类优化范式：正则化方法与凸优化方法。正则化方法通过在损失函数中引入惩罚项来软性约束参数更新。许多工作通过对梯度或权重施加约束来抑制噪声影响，例如Kim等人[9]采用的动量阈值约束与Li等人[10]提出的动态权重约束。在防止过拟合方面，正则化技术也被广泛采用，如Zheng等人[11]构建的动态类别平衡正则项，旨在缓解训练向头部类别偏移的问题。这类方法虽在一定程度上提升了性能，但其约束往往依赖于训练集的静态统计特性。当应用于直播视频等内容不可预测的场景时，这类静态约束容易演变为模型偏见，反而制约其泛化能力。

相比之下，凸优化方法通过数学变换，在局部将非凸问题转化为凸子问题，并利用凸优化的几何收敛约束来引导优化。该方法侧重于从优化景观的几何结构上实现分布适配，而非直接压缩参数的表达空间，有助于在拟合目标分布的同时，还能保持模型的泛化能力。在VLM模型训练中直接实现全局凸优化较为困难，一些研究尝试通过结构设计逼近凸优化。例如，Dong等人[12]采用分组协同学习，将语义空间分解为局部子空间进行协同优化；Zheng等人[13]提出的基于原型的嵌入网络（Prototype-based Embedding Network，PENet）则利用语义原型锚定特征分布以规避全局寻优的复杂性。这两种思路均存在局限，前者对子空间划分较为敏感，容易损失细粒度信息，从而限制模型的泛化与迁移能力；后者依赖于静态统计先验，在面对直播视频等动态数据时易出现优化震荡与收敛不稳的问题。

因此，针对直播视频内容不可预测、特征分布实时流动的特点，发展一种能够自适应感知特征模态偏移并动态形成凸优化子空间的机制，对于保障模型泛化能力实现精准分布适配，显得尤为必要。本研究将探索此类自适应凸优化路径，以平衡模型对特定场景的适应能力与对未知场景的泛化能力。

1.2 记忆智能体

记忆智能体能够从时间维度感知内在联系和依赖关系,近些年被用于视频理解中,主要分为在线处理的记忆机制、层次化事件记忆机制。其中,在线处理记忆机制将关键信息存储在记忆库中,以便后续的高效访问和推理,如Zhang等人[14]提出的面向长视频流的实时理解(Real-Time understanding for long Video Streams, Flash-VStream)通过显式筛选关键帧来降低显存消耗;层次化方法则侧重于长时理解,如Cheng等人[15]通过划分事件边界并构建局部-全局双层记忆库来减少信息冗余。尽管现有记忆机制能够减少信息冗余,却难以适应直播场景的动态变化,面临着"更新灵活性"与"事件边界模糊"的双重挑战。固定的更新策略无法及时响应由弹幕互动引发的语义跳变;同时,直播流缺乏清晰的事件边界,导致基于预定义段落或场景划分的记忆更新机制容易失效,造成记忆内容与当前语境错位。因此,构建一种能够基于语义原型动态更新并能融合弹幕-视觉跨模态时序逻辑的记忆模块,是提升直播场景理解能力的关键。

1.3 特征联合

在特征联合方法中,常利用Query为模型提供高级描述和对象感知,如Tu等人[16]致力于通过多模态Query增强表示能力。受限于这类方法设置Query方式的灵活性不足问题,基于Query筛选的特征联合机制被提出来去除冗余信息。在特征选择与关联建模上,Hu等人[17]采用高斯滤波器进行空间筛选,Lu等人[18]构建了跨实例信息传递图,而Kim等人[19]提出的分组查询特化及质量感知多重分配(Groupwise query SPEciAlization and Quality-aware multi-assignment, SpeaQ)策略则依据统计频率对目标进行分组训练。现有方法虽然在标准数据集上表现良好,但其查询机制往往无法根据实时语义语境动态调整关注焦点,导致模型要么保留过多冗余背景特征,要么遗漏弹幕所关注的细粒度目标。与已有方法不同,本文提出了一种特征联合与关系筛选策略来筛选冗余的对象,从而加强弹幕信息与视觉内容的关联性,以便生成更鲁棒的场景图。

2 方法

2.1 问题定义

动态场景图生成:通过从视频序列的目标帧中,精准检测和解析视觉关系,形成场景图这一结构化数据表征。场景图的节点对应视频帧中的对象实例,由类别标签和实例间的语义关系共同定义,可以视为物体对象s、预测类别p、人物对象o信息构成的三元组形式列表$\langle s,p,o\rangle$。在之前的工作中[20],场景图生成的方式是在每个帧上建模联合概率f:

$$f\left(\langle s,p,o\rangle|\boldsymbol{V}\right)=f\left(p|s,o\right)f\left(s,o|\boldsymbol{R}\right)f\left(\boldsymbol{R}|\boldsymbol{V}\right) \tag{1}$$

其中:$\boldsymbol{V}$表示输入的视频帧序列;$\boldsymbol{R}=\{s,o\}$表示对象的检测结果。

在视频场景图生成任务中,为提升帧间一致性,在时间维度上引入状态信息$\boldsymbol{T}$:

$$f\left(\langle s,p,o\rangle|\boldsymbol{V}\right)=f\left(p|s,o\right)f\left(s,o|\boldsymbol{T}\right)f\left(\boldsymbol{T}|\boldsymbol{R}\right)f\left(\boldsymbol{R}|\boldsymbol{V}\right) \tag{2}$$

为了优化针对网络直播视频的场景图生成性能,本文基于VLM凸优化理论计算并更新联合概率f中各变量的梯度参数集合$\boldsymbol{\Theta}=\left\{\boldsymbol{\theta}_{ij}\right\}_{i,j=1}^{N}$。在凸优化过程中,需要累积每个参数$\boldsymbol{\theta}_{ij}$的梯度来计算损失函数在区间$N$的梯度累积情况$\boldsymbol{\Gamma}_{ij}$:

$$\boldsymbol{\Gamma}_{ij}=\frac{1}{N}\sum_{n=1}^{N}\nabla\boldsymbol{\theta}_{ij}\mathcal{L}\left(\boldsymbol{\Theta},\mathcal{B}_n\right) \tag{3}$$

其中:$\mathcal{B}_n$代表第n个最小批量数据;$\nabla\boldsymbol{\theta}_{ij}\mathcal{L}\left(\boldsymbol{\Theta},\mathcal{B}_n\right)$是损失函数相对于小批量数据$\boldsymbol{\theta}_{ij}$的梯度。

在累积N个批次的梯度后,梯度信息可被看作一种信号,以精细的方式确定各要素参数的重要性。这一操作包括按绝对值对每个累积梯度$\boldsymbol{\Gamma}_{ij}$的分量$\boldsymbol{g}_{ij}$进行排序,并选择一个预定义的最高百分位数k计算阈值$\boldsymbol{\varphi}_k$,从而屏蔽梯度$\mathcal{M}\left(\boldsymbol{\Gamma}_{ij},k\right)$:

$$\mathcal{M}\left(\boldsymbol{\Gamma}_{ij},k\right)=\left\{\boldsymbol{g}_{ij}|\boldsymbol{g}_{ij}\in\boldsymbol{\Gamma}_{ij},|\boldsymbol{g}_{ij}\geqslant\boldsymbol{\varphi}_k\right\} \tag{4}$$

随后利用掩蔽梯度$\mathcal{M}\left(\boldsymbol{\Gamma}_{ij},k\right)$更新参数$\boldsymbol{\theta}_{ij}$:

$$\boldsymbol{\theta}_{ij}^{(t+1)}=\boldsymbol{\theta}_{ij}^{(t)}-\eta\cdot\mathcal{M}\left(\boldsymbol{\Gamma}_{ij},k\right) \tag{5}$$

其中:η是学习率;t表示当前训练步骤。

2.2 整体结构

本文提出的VCO-SGG方法通过视觉/文本特征编码、动态原型记忆模块、特征联合与关系筛选策略,对直播视频中的多模态信息进行统一表征,并利用VLM凸优化架构对模型参数进行联合优化,以提升场景图生成准确性。整体框架如图2所示,其核心由以下三个部分构成。

(1)VLM凸优化架构

该部分采用ResNet与VLM编码器,分别提取视频帧的视觉特征及伴随文本(如弹幕评论)的语言特征,形成包含场景空间与语义信息的初始表示。基于局部凸近似理论,本文在优化过程中将整体非凸问题转化为一系列局部凸子空间内的逼近求解。通过对编码器、动态原型记忆模块、特征联合与关系筛选策略进行联合凸优化,模型能够在各子空间内稳定调整参数分布,从而提升生成结果的准确性与鲁棒性。

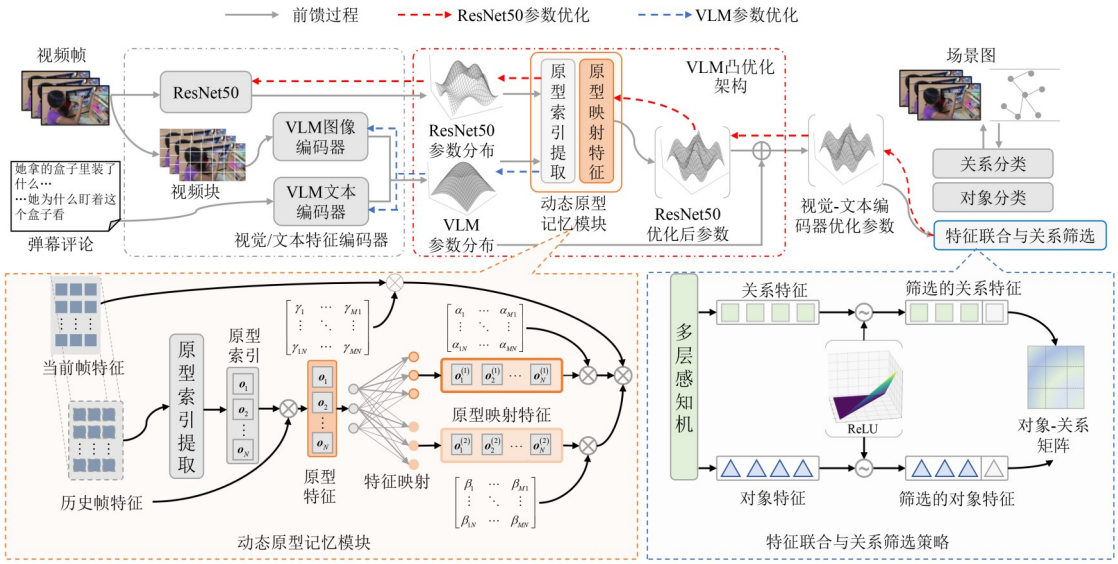


图2 所提场景图生成模型的整体结构

Figure 2 The overall structure of proposed scene graph generation model

(2)动态原型记忆模块

为有效捕捉直播场景中的时序上下文,本模块接收编码后的特征,并通过原型索引与特征映射机制,自适应地聚合历史帧中的关键信息。该设计增强了对内容突变与时序跳变的建模能力,为关系识别提供了连贯的上下文依据。

(3)特征联合与关系筛选策略

该模块负责对场景中对象及其关系进行关联建模。通过引入基于ReLU的门控机制,对关系表征中的冗余项进行动态筛选,抑制无效或弱相关的交互,从而提升对象分类与关系预测的判别性。最终,筛选后的对象与关系表示被统一送入凸优化框架,参与梯度计算以及模型参数的更新。

2.3 VLM凸优化

VLM凸优化架构在场景图生成中作用过程如图3所示。在前馈过程,视频帧与弹幕评论分别经由视觉编码器与文本编码器进行特征提取,并通过动态原型记忆模块增强时空语义建模能力,随后经过特征联合与关系筛选,输出结构化的场景图表示。在凸优化过程,所生成的场景图用于对模型参数分布进行反向优化。具体而言,首先在动态原型记忆模块以及特征联合与关系筛选模块中进行梯度计算与累积,以量化各模块对场景图生成结果的贡献度。随后,基于累积的梯度信息更新编码器的参数分布,从而在凸优化约束下实现参数的稳定收敛与整体性能的持续提升。相比于全参微调和低秩子空间方法,本文所采用的凸优化策略在模型对特定场景的适应能力与对未知场景的泛化能力之间取得了更好的平衡,如图4所示。常规全参微调(图4(a))直接利用直播数据更新VLM参数,虽然有效调整了特征分布,但计算开销过大,在实际部署中难以应用。基于低秩子空间的方法(图4(b))尝试通过低维约束来稳定训练过程,在计算成本与分布调整效果之间取得了一定平衡,但其强行降维易导致参数分布僵化,难以充分拟合直播视频中复杂多变的非平稳特征。

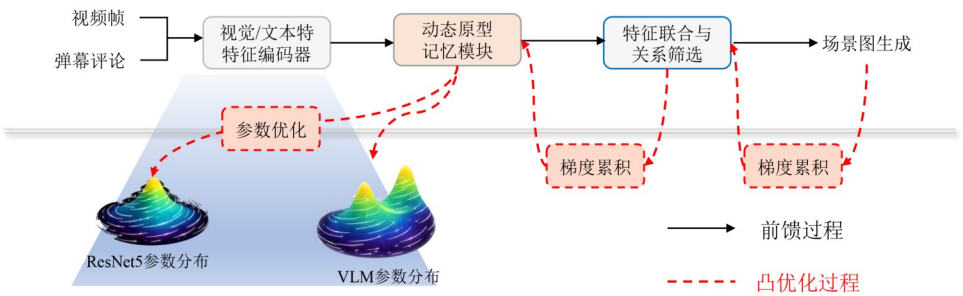
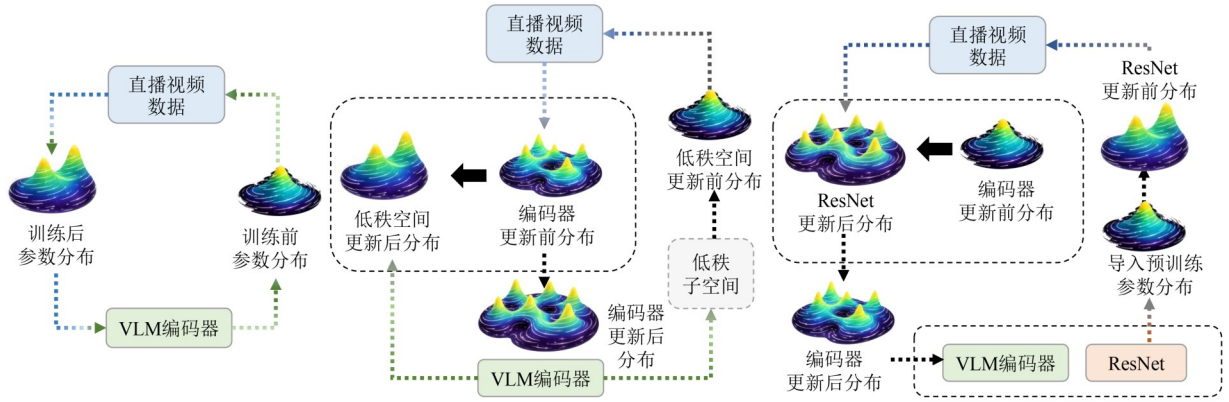


图3 VLM凸优化的作用过程

Figure 3 The process of VLM convex optimization

相比之下,本文方法[图4(c)]引入了凸优化策略,将优化路径约束在预训练分布的局部邻域内,并利用梯度掩码剔除负曲率方向的参数更新。该方法在适应新数据的同时,能确保参数流形在局部范围内始终保持平滑、良态的凸结构,从而实现稳定而高效的特征分布适配。



(a) 全参微调优化方法　(b) 低秩子空间优化方法　(c) 凸优化方法

(a) Full-parameter fine-tuning optimization　(b) Low-rank subspace optimization　(c) Convex optimization

图4　VLM优化方法比较

Figure 4　Comparison of VLM optimization methods

2.3.1　基于VLM的特征提取

使用ResNet和CLIP编码器提取帧特征 $\boldsymbol{F}_{\mathrm{V}}$ 和文本特征 $\boldsymbol{F}_{\mathrm{T}}$,融合这两种模态的特征以建模对象信息。鉴于不同模态间语义的不兼容性,直接进行融合存在困难,因此本文引入多头注意力机制提取初始特征 $\tilde{\boldsymbol{F}}$:

$$\tilde{\boldsymbol{F}}=\mathrm{softmax}\left(\frac{\mathcal{F}\left(\boldsymbol{F}_{\mathrm{V}}\right)\cdot\boldsymbol{W}_{\mathrm{V}}\left(\boldsymbol{W}_{\mathrm{T}}\cdot\mathrm{FC}\left(\boldsymbol{F}_{\mathrm{T}}\right)\right)^{\mathrm{T}}}{\sqrt{d/2}}\right)\mathrm{FC}\left(\boldsymbol{F}_{\mathrm{T}}\right)\cdot\boldsymbol{W}_{\mathrm{T}}+\mathcal{F}\left(\boldsymbol{F}_{\mathrm{V}}\right)\tag{6}$$

其中:$\mathcal{F}$ 和FC分别表示展平和全连接层操作;$\boldsymbol{W}_{\mathrm{V}}$ 和 $\boldsymbol{W}_{\mathrm{T}}$ 代表视频帧和文本特征的映射矩阵;d 为通道数。

将 $\tilde{\boldsymbol{F}}$ 初始特征融合为视觉-文本特征 $\boldsymbol{F}_{\mathrm{ini}}$:

$$\boldsymbol{F}_{\mathrm{ini}}=\mathrm{MLP}\left(\mathrm{LN}\left(\tilde{\boldsymbol{F}}\right)\right)+\tilde{\boldsymbol{F}}\tag{7}$$

融合后的多模态信息用于建模目标的形态与位置信息。在此过程中,对视觉-文本特征进行非线性映射SiLU,得到视觉-文本映射特征 $\tilde{\boldsymbol{F}}_{\mathrm{io}}$:

$$\tilde{\boldsymbol{F}}_{\mathrm{io}}=\mathrm{SiLU}\left(\mathrm{BN}\left(\mathrm{Conv}\left(\boldsymbol{F}_{\mathrm{ini}}\right)\right)\right)\tag{8}$$

随后,结合自注意力和凸优化桥接矩阵,获得初始凸优化特征 $\boldsymbol{F}_{\mathrm{io}}$:

$$\boldsymbol{F}_{\mathrm{io}}=\mathrm{MLP}\left(\mathrm{LN}\left(\mathrm{SA}\left(\mathrm{LN}\left(\tilde{\boldsymbol{F}}_{\mathrm{io}}\right)\right)+\tilde{\boldsymbol{F}}_{\mathrm{io}}\right)\right)\tag{9}$$

接着,将该特征转化为展平特征 $\tilde{\boldsymbol{F}}$,并进一步生成视觉-文本特征:

$$\tilde{\boldsymbol{F}}_{\mathrm{o}}=\mathcal{F}\left(\mathcal{R}\left(\boldsymbol{F}_{\mathrm{io}}\right)\odot\begin{bmatrix}1&0\\0&0\end{bmatrix}\right)\tag{10}$$

其中:$\mathcal{R}$ 表示反展平操作;$\odot$ 为平滑加权操作。

最后,利用权重矩阵 $\boldsymbol{Q}$ 优化特征中的对象信息及相互关联,得到视觉-文本特征 $\boldsymbol{F}_{\mathrm{o}}$:

$$\boldsymbol{F}_{\mathrm{o}}=\mathrm{BN}\left(\mathrm{softmax}\left(\frac{\boldsymbol{Q}\cdot\left(\tilde{\boldsymbol{F}}_{\mathrm{o}}\right)^{\mathrm{T}}}{\sqrt{d/2}}\right)\tilde{\boldsymbol{F}}_{\mathrm{o}}+\tilde{\boldsymbol{F}}_{\mathrm{o}}\right)\tag{11}$$

2.3.2　凸函数优化特征分布

在凸函数优化特征分布的过程中,首先采用凸优化机制从整体上优化特征分布,然后应用梯度更新适配直播视频的特征分布。

(1)凸函数优化

利用凸优化机制将视觉-文本特征隐式投影到一个凸约束流形上可得到凸优化特征。凸优化特征满足局部凸条件,可用于解决直播视频数据中由负曲率方向和非凸结构引起的优化不稳定问题。为使其满足凸优化条件,定义目标凸约束流形 $\mathcal{M}$ 为

$$\mathcal{M}=\left\{\boldsymbol{g}_{ij}\middle|\boldsymbol{g}_{ij}\in\boldsymbol{\Gamma}_{ij},|\boldsymbol{g}_{ij}\geqslant\boldsymbol{\varphi}\right\}\tag{12}$$

其中:$\boldsymbol{\Gamma}_{ij}$、$\boldsymbol{g}_{ij}$ 分别为累积梯度及其分量。通过求解如下约束优化问题,将 $\boldsymbol{F}_{\mathrm{o}}$ 投影至 $\mathcal{M}$:

$$\boldsymbol{F}_{\mathrm{o}}=\arg\min_{\boldsymbol{F}_{\mathcal{M}}\in\mathcal{M}}\left\|\boldsymbol{F}_{\mathcal{M}}-\boldsymbol{F}_{\mathrm{ini}}\right\|^{2}\tag{13}$$

其中:$\boldsymbol{F}_{\mathcal{M}}$ 是符合约束 $\mathcal{M}$ 的特征。

该投影操作沿几何最短路径将游离于流形外的特征拉回至凸区域,从而得到满足局部凸性条件的凸优化特征 $\boldsymbol{F}_{\mathrm{o}}$,为后续优化提供稳定的特征基础。整体流程如算法1所示。

(2)凸函数优化直播视频特征

为了适配直播视频特征,采用调整ResNet模型的

算法1 凸优化机制

输入：

- N: 步长
- η: 学习率
- $\theta^{(0)}$: 初始模型参数
- x_i, y_i: 特征值与标签

输出：更新后模型参数$\theta^{(T)}$

for t in $0 \rightarrow T-1$ **do**

$\boldsymbol{\Gamma} \leftarrow 0$

for n in $1 \rightarrow N$ **do**

$\boldsymbol{\Gamma} \leftarrow \boldsymbol{\Gamma} + \nabla_{\theta}\mathcal{L}\left(\boldsymbol{\theta}^{(t)}, \boldsymbol{O}^{(t)}; \boldsymbol{D}\right)$

end for

$\boldsymbol{\Gamma} \leftarrow (1/N)\boldsymbol{\Gamma}$

$\boldsymbol{\varphi}_k \leftarrow$ Percentile threshold based on $\boldsymbol{\Gamma}$

$(\boldsymbol{\Gamma}, k) \leftarrow \{\boldsymbol{g}_{ij} \mid \boldsymbol{g}_{ij} \in \boldsymbol{\Gamma}_{ij}, |\boldsymbol{g}_{ij}| \geqslant \boldsymbol{\varphi}_k\}$

for $\boldsymbol{\theta}_{ij} \in \boldsymbol{\Theta}$ **do**

if $|\boldsymbol{\Gamma}_{ij}| \geqslant \boldsymbol{\varphi}_k$ **then**

$\tilde{\boldsymbol{\theta}}_{ij}^{(t)} \leftarrow \mathcal{P}_{\mathcal{C}}(\boldsymbol{\theta}_{ij}^{(t)})$

$\boldsymbol{\theta}_{ij}^{(t+1)} \leftarrow \tilde{\boldsymbol{\theta}}_{ij}^{(t)} - \eta \times \boldsymbol{\Gamma}_{ij}$

end if

end for

end for

Return $\boldsymbol{\theta}^{(T)}$

梯度更新方法。假设数据集$\boldsymbol{D}$上各变量的梯度参数集合$\boldsymbol{\Theta}=\left\{\boldsymbol{\theta}_{ij}\right\}_{i,j=1}^{N}$，则损失函数$\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{D})$为

$$\mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{D}) = \frac{1}{n}\sum_{i=1}^{n}\ell\left(\boldsymbol{\Theta}; \left(\boldsymbol{x}_i, \boldsymbol{y}_i\right)\right) \tag{14}$$

其中：ℓ表示单个数据点的损失；$\boldsymbol{x}_i$和$\boldsymbol{y}_i$分别表示输入特征和相应标签。

为辨别单个参数$\boldsymbol{\theta}_{ij}$对损失函数$\mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{D})$的影响，在保持所有其他参数不变的情况下去除这些参数的影响，即剔除参数$\boldsymbol{\theta}_{ij}$受直播视频数据分布的影响，通过损失$\Delta\mathcal{L}_{ij}(\boldsymbol{\Theta}; \boldsymbol{D})$的变化进行量化。

$$\Delta\mathcal{L}_{ij}(\boldsymbol{\Theta}; \boldsymbol{D}) = \mathcal{L}(\boldsymbol{I}\odot\boldsymbol{\Theta}; \boldsymbol{D}) - \mathcal{L}\left(\left(\boldsymbol{I}-\mathcal{E}_{ij}\right)\odot\boldsymbol{\Theta}; \boldsymbol{D}\right) \tag{15}$$

其中：$\boldsymbol{I}$表示同一矩阵；$\boldsymbol{\varepsilon}_{ij}$与$\boldsymbol{\Theta}$具有相同维度的指示矩阵，除$(i,j)$元素为1外，其他元素均为0。

在评估更新每个参数产生的损失$\Delta\mathcal{L}_{ij}$时，对当前参数向量$\boldsymbol{\Theta}$进行一阶泰勒级数展开，其近似值由损失函数关于$\boldsymbol{\theta}_{ij}$的梯度表示：

$$\Delta\mathcal{L}_{ij}(\boldsymbol{\Theta}; \boldsymbol{D}) \approx \nabla\boldsymbol{\theta}_{ij}\mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{D}) \cdot \left(-\boldsymbol{\theta}_{ij}\right) \tag{16}$$

其中：$\nabla\boldsymbol{\theta}_{ij}\mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{D})$表示损失函数$\mathcal{L}$关于参数$\boldsymbol{\theta}_{ij}$的偏导数。

为确保目标函数在直播视频特征学习中可以被转化为凸形式，本文通过对参数空间中$\boldsymbol{\Theta}=\left\{\boldsymbol{\theta}_{ij}\right\}_{i,j=1}^{N}$的海森矩阵$\mathcal{H}(\boldsymbol{\theta})$施加约束来规范损失曲面的局部几何结构。

具体而言，损失函数$\mathcal{L}(\boldsymbol{\theta})$在参数$\boldsymbol{\theta}_0$附近的二阶泰勒展开为

$$\mathcal{L}(\boldsymbol{\theta}) \approx \mathcal{L}(\boldsymbol{\theta}_0) + \nabla\mathcal{L}(\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathcal{H}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0) \tag{17}$$

其中：二次项$(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathcal{H}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$直接主导了局部曲面的形状。当约束海森矩阵半正定（即$\mathcal{H}(\boldsymbol{\theta})\geqslant 0$）时，该领域为凸区域。

为处理海森矩阵的不定性，本文引入梯度累积量$\boldsymbol{\Gamma}$来追踪参数更新方向的历史信息，其定义如下：

$$\boldsymbol{\Gamma}^{(t+1)} = \beta\cdot\boldsymbol{\Gamma}^{(t)} + \left(1-\beta\right)\cdot\nabla\mathcal{L}\left(\boldsymbol{\theta}^{(t)}\right) \tag{18}$$

其中：β为可学习参数。梯度累积量$\boldsymbol{\Gamma}$的稳定性反映了损失曲面主曲率方向的性质。基于这一显著性度量，本文构建二进制掩码矩阵$\boldsymbol{M}=\left\{\boldsymbol{M}_{ij}\right\}_{i,j=1}^{N}$，其元素定义如下：

$$\boldsymbol{M}_{ij} = \begin{cases} 1, & \text{if } \left|\boldsymbol{\Gamma}_{ij}\right| \geqslant \tau \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

其中：τ为预设的显著性阈值。此掩码机制确保仅对累积梯度幅度较大（即曲率方向稳定且倾向于正定）的参数子集进行更新，从而将优化过程限制在海森矩阵半正定的子空间内，从数学上保证了问题的局部凸性。

当VLM模型参数被加入掩码之后，模型参数就会被固定。随着训练轮数的不断加深，梯度变化中的显著性度量会逐渐累积到ResNet模型参数中。这使得该模型会逐渐与VLM模型参数互补，形成符合直播视频数据的模型参数分布，从而提升模型的准确率。

（3）凸优化收敛性与稳定性

为了从数学上严格证明模型能收敛至全局最优解，本文分析了Lipschitz连续梯度假设对于目标函数的适用性，并基于该假设分析了模型的收敛边界。

首先，介绍Lipschitz连续梯度假设的适用性。深度学习优化中Lipschitz条件的破坏，常源于非光滑激活函数（如ReLU）梯度的累积效应。为此，本方法通过差异化架构设计规避了这一问题。

特征提取骨干网络。采用SiLU激活函数替代ReLU。SiLU处处连续可微，从源头保证了特征映射的二阶光滑性，切断了非光滑算子的累积路径，使主干网络满足Lipschitz条件。

特征联合与关系筛选策略。引入的ReLU并非用于传统的非线性激活，而是与Sigmoid配合构成一个稀疏门控，其数学本质是一个软阈值算子。该算子仅作用于网络的浅层末端，用于过滤负值冗余索引，稀疏化特性有效抑制了梯度范数的膨胀，而不会破坏深

层的光滑性。

综上，通过骨干网络的光滑化设计与门控模块的受限使用，确保了目标函数在经掩码筛选后的有效优化子空间内满足广义Lipschitz连续梯度条件。

其次，介绍基于Lipschitz条件的收敛性分析。Lipschitz连续性为损失曲面的曲率提供了上界，从而为优化过程引入了关键的几何约束。基于此，本文可引用优化理论中的下降引理来严格推导收敛边界。

具体而言，对于满足Lipschitz条件的损失函数$\mathcal{L}(\boldsymbol{\theta})$，其单步损失下降可近似为

$$\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) \leqslant \mathcal{L}(\boldsymbol{\theta}^{(t)}) - \left\| \boldsymbol{M}_t \nabla \mathcal{L}\left(\boldsymbol{\theta}^{(t)}\right) \right\|^2 \tag{20}$$

其中：$\boldsymbol{M}_t$表示第t步更新的二进制掩码矩阵。由于损失$\mathcal{L}(\boldsymbol{\theta})$有下界，对上述不等式从$t=(1,\infty)$求和可证明，梯度范数序列$\sum_{t=1}^{\infty}\left\| \boldsymbol{M}_t \nabla \mathcal{L}\left(\boldsymbol{\theta}^{(t)}\right) \right\|^2$收敛，从而必有$\left\| \boldsymbol{M}_t \nabla \mathcal{L}\left(\boldsymbol{\theta}^{(t)}\right) \right\|^2 \to 0$。这从理论上确保了优化过程必然收敛至一个梯度投影为零的临界点。

更关键的是，掩码矩阵$\boldsymbol{M}$将优化严格限制在海森矩阵$\mathcal{H}(\boldsymbol{\theta})$半正定的凸子空间内。结合凸优化的基本定理，在凸区域中，任何临界点即为全局最小点，可以断定，该收敛终点正是目标函数的全局最优解。因此，模型具备收敛至全局最优的理论保证。

最后，介绍稳定性分析。本文通过梯度累积机制主动管理这一方差。具体而言，定义累积梯度变量$\boldsymbol{\Gamma}=\frac{1}{N}\sum_{n=1}^{N}\nabla\mathcal{L}_n(\boldsymbol{\theta})$，$N$为累积步数。根据中心极限定理，该估计量的方差满足：

$$\mathrm{Var}(\boldsymbol{\Gamma}) \approx \frac{\sigma^2}{N} \tag{21}$$

其中：σ^2为单步梯度的方差。这意味着，累积梯度估计的标准差（即震荡幅度的主要能量）被压缩了$1/\sqrt{N}$倍。

因此，本文的梯度累积操作在数学上等价于对优化路径的噪声进行了低通滤波。通过选择适当的累积步数N，可以将更新轨迹的震荡幅度系统性地降低，从而在复杂的直播视频数据流中，依然确保优化路径的平滑与稳定，从而保障模型可靠、鲁棒地收敛。

2.4　动态原型记忆模块

动态原型记忆模块主要用于建模非剪裁网络直播视频中的时序关系。由于此类视频的历史帧特征往往包含较多冗余信息，直接使用时序建模效果受限。为此，本模块通过强化对关键历史信息的关联与利用，借助历史帧特征引导当前帧的特征的提取，从而更有效地捕捉帧间的时序逻辑。具体流程如图5所示：首先根据历史帧的特征分布生成原型索引，进而提取能够保留关键信息的原型映射特征；随后将该特征与当前帧的映射特征进行融合，得到增强后的记忆融合特征。

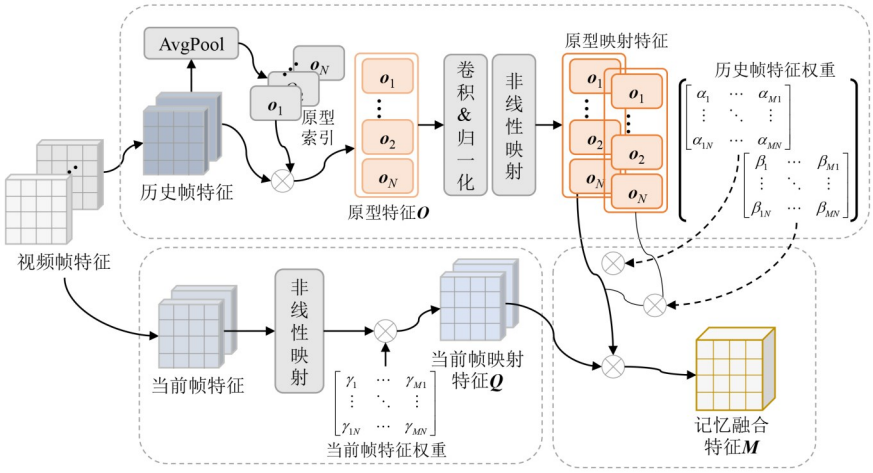


图5　动态原型记忆体流程

Figure 5　The process of dynamic prototype memory

具体步骤如下。

步骤1：原型特征生成

利用全局平均池化Avgpool生成原型索引，并抽取历史帧特征中的关键信息得到原型特征$\boldsymbol{O}=[\boldsymbol{o}_1 \cdots \boldsymbol{o}_N]$：

$$\boldsymbol{O}=\mathrm{Avgpool}\left(\boldsymbol{F}_{\mathrm{o}}\right) \tag{22}$$

该操作可在不增加计算开销的前提下，初始化具有丰富对象信息的原型索引向量，相较于初始化为空矩阵的方法，能够提供更充分的语义基础。

步骤2：原型特征映射

为了将历史帧特征的语义信息补充至当前帧特征中，将原型特征 $\boldsymbol{O}$ 映射为“键”信息 $\boldsymbol{O}_{\mathrm{K}}$ 以及“值”信息 $\boldsymbol{O}_{\mathrm{V}}$。

$$\boldsymbol{O}_{\mathrm{K}}=\begin{bmatrix}\alpha_{11} & \cdots & \alpha_{M1}\\ \vdots & \ddots & \vdots\\ \alpha_{1N} & \cdots & \alpha_{MN}\end{bmatrix}\cdot\boldsymbol{O},\boldsymbol{O}_{\mathrm{V}}=\begin{bmatrix}\beta_{11} & \cdots & \beta_{M1}\\ \vdots & \ddots & \vdots\\ \beta_{1N} & \cdots & \beta_{MN}\end{bmatrix}\cdot\boldsymbol{O}\tag{23}$$

其中：$[\alpha_{11}\cdots\alpha_{MN}]$ 和 $[\beta_{11}\cdots\beta_{MN}]$ 为历史帧特征权重。基于此，在后续的特征融合过程中能够通过历史帧特征权重将关键帧特征中的语义信息传递到当前帧特征中。

步骤3：当前帧映射特征提取

为避免融合过程破坏当前帧原有的语义结构，通过线性映射获取当前帧特征的映射表示 $\boldsymbol{Q}$：

$$\boldsymbol{Q}=\mathrm{Linear}\left(\begin{bmatrix}\gamma_{1} & \cdots & \gamma_{M1}\\ \vdots & \ddots & \vdots\\ \gamma_{1N} & \cdots & \gamma_{MN}\end{bmatrix}\cdot\boldsymbol{F}\right)\tag{24}$$

其中：Linear为线性变换。

步骤4：语义融合

在获得映射特征 $\boldsymbol{Q}$、键 $\boldsymbol{O}_{\mathrm{K}}$、值 $\boldsymbol{O}_{\mathrm{V}}$ 之后，通过缩放点积注意力机制进行特征融合，得到融合特征 $\boldsymbol{M}$：

$$\boldsymbol{M}=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{O}_{\mathrm{K}}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{O}_{\mathrm{V}}\tag{25}$$

其中：d_k 为特征维度，缩放因子 $\sqrt{d_k}$ 用于稳定梯度。该操作将关键帧中的语义信息有选择地补充到当前帧中，从而增强时间维度上的语义一致性。

步骤5：模型参数更新

引入动态原型记忆模块后，在凸优化框架下更新模型参数 $\boldsymbol{\theta}$ 时，需考虑原型特征 $\boldsymbol{O}$ 在直播数据集 $\boldsymbol{D}$ 上的累积梯度值 $\boldsymbol{\Gamma}$：

$$\boldsymbol{\Gamma}=\frac{1}{N}\sum_{n=1}^{N}\nabla\mathcal{L}\left(\boldsymbol{\theta}^{(t)},\boldsymbol{O}^{(t)};\boldsymbol{D}\right)\tag{26}$$

其中：$\nabla\mathcal{L}(\cdot)$ 为当前批次的梯度值。

通过上述步骤，可高效地提取关键帧特征，并借助交叉注意力机制实现历史帧与当前帧语义信息的融合，从而提升视频特征表示的质量与一致性。

2.5 场景图生成

由于网络直播视频中元素的类型和数量变化不规律，固定数量的对象索引中不可避免地会包含冗余项，可能对元素间关联关系的准确性产生负面影响。本节介绍对象索引的去冗余方法，并阐释如何基于筛选后的对象索引实现场景图的生成。

2.5.1 特征联合与关系筛选策略

本部分采用基于ReLU结构的简洁设计，该设计主要基于以下三个方面的考虑，以实现与整体凸优化框架的针对性适配。

（1）语义保真度：该模块位于网络末端，其输入特征已由前序VLM凸优化架构中的主干网络充分处理，语义表征相对完备。引入复杂的筛选机制易引入额外归纳偏置，干扰现有语义信息；而ReLU作为无参数的激活算子，能以最小代价保留核心语义。

（2）优化连续性：本文核心在于构建稳定的凸优化训练框架。复杂模块的非线性映射可能加剧梯度震荡，与凸优化目标冲突。为最大限度保持优化过程的Lipschitz连续性，此处采用单一的ReLU算子，确保凸优化约束可平滑延伸至网络末端。

（3）直播场景特性：该模块用于解决直播视频中对象数量波动引起的查询冗余问题。ReLU在数学上等效于一个阈值门控，能够以极低计算成本过滤低置信度查询，从而自适应地筛选关键特征。

经综合权衡特征保护、模块间兼容性与去冗余效能后，最终选择以简单的ReLU构建特征关联机制，从而以最小计算代价实现稳健的场景图生成。

具体而言，采用ReLU函数将负值对象索引置零，仅保留正值索引，以更精准地构建元素间的关联关系。去除冗余对象索引后，借助检测头获取对象的位置和类别信息，并结合关系联合头获取对象间的相互关系，生成场景图。该过程有效过滤了冗余信息，并确保了场景图生成中元素间关系的准确性和一致性。

令对象索引 $\boldsymbol{H}=\left[\boldsymbol{h}_1\,\boldsymbol{h}_2\cdots\boldsymbol{h}_n\right]$，其中 $\boldsymbol{h}_n$ 表示第 n 个对象的索引值。

（1）使用sigmoid函数对特征向量进行非线性映射：

$$\boldsymbol{H}=\mathrm{sigmoid}\left(\mathrm{Linear}\left(\boldsymbol{H}\right)\right)\tag{27}$$

（2）为了确保元素间关联的准确性，对映射后的对象索引执行ReLU操作，以去除冗余项，得到筛选索引 $\boldsymbol{H}'$：

$$\boldsymbol{H}'=\mathrm{ReLU}\left(\boldsymbol{H}\right)\tag{28}$$

此操作将负值索引置零，仅保留正值索引。

（3）引入特征联合与关系筛选策略后，在模型参数 $\boldsymbol{\theta}$ 的更新过程中需增加基于ReLU的凸约束投影：

$$\tilde{\boldsymbol{\theta}}_{ij}^{(t)}=\mathcal{P}_{\mathcal{C}}\left(\boldsymbol{\theta}^{(t)}\right)\tag{29}$$

其中：$\mathcal{P}_{\mathcal{C}}(\cdot)$ 表示基于ReLU的凸约束投影，在梯度优化时自动过滤负值特征。

（4）结合累积梯度值 $\boldsymbol{\Gamma}$ 和投影后的参数 $\tilde{\boldsymbol{\theta}}$，可得到更新后的参数：

$$\boldsymbol{\theta}_{ij}^{(t+1)}=\tilde{\boldsymbol{\theta}}_{ij}^{(t)}-\eta\times\boldsymbol{\Gamma}_{ij}\tag{30}$$

其中：η 为学习率，累积梯度值 $\boldsymbol{\Gamma}_{ij}\in\mathcal{M}(\boldsymbol{\Gamma},k)$ 需满足凸约束流形 $\mathcal{M}$ 的定义。

2.5.2 基于索引的场景图生成

在获得去冗余后的对象索引 $\boldsymbol{H}'$ 后,通过目标检测头获取每个对象的边界位置 $\boldsymbol{R}_{\mathrm{o}}$:

$$\boldsymbol{R}_{\mathrm{o}}=\mathrm{softmax}\left(\mathrm{FC}\left(\boldsymbol{H}'\right)\right) \tag{31}$$

同时,通过关系分类头获取元素间的相互联系表示 $\boldsymbol{R}_{\mathrm{e}}$:

$$\boldsymbol{R}_{\mathrm{e}}=\mathrm{softmax}\left(\boldsymbol{H}'\right) \tag{32}$$

结合筛选索引 $\boldsymbol{H}'$、边界框位置 $\boldsymbol{R}_{\mathrm{o}}$ 以及关系表示 $\boldsymbol{R}_{\mathrm{e}}$,生成场景图 $\boldsymbol{G}$:

$$\boldsymbol{G}=\mathrm{SceneGraph}\left(\boldsymbol{H}',\boldsymbol{R}_{\mathrm{o}},\boldsymbol{R}_{\mathrm{e}}\right) \tag{33}$$

这一过程不仅有效过滤了冗余信息,还确保了场景图生成过程中元素间关系的准确性和一致性。

3 实验结果与分析

3.1 实验设置

3.1.1 数据集

为验证所提VCO-SGG方法在直播视频场景图生成中的有效性和优越性,本文收集并构建了一个直播视频数据集BJUT-LSGD。同时,为测试方法的泛化性能,选取两个公开数据集的精简版本Mini Charades与Mini Action Genom进行实验,精简版本在满足各对比方法部署要求的同时提升了实验效率。各数据集具体信息如下。

BJUT-LSGD数据集:包括126个训练视频和40个测试视频,每个视频含20~30帧,分辨率为1 280×720。为避免过拟合,对训练数据进行包括旋转、裁剪、平移、翻转在内的数据增强。增强后数据规模扩展至503训练视频和161测试视频,输入分辨率统一调整为600×384。

Mini Charades数据集:包含3 090个训练视频和1 224个测试视频,每段视频包含10~20帧,帧分辨率约为500×300。数据集中主要标注了人体、物体的位置尺寸及其之间的交互活动。

Mini Action Genome数据集:包含3 884个训练视频和1 538个测试视频,每段包含10~20帧,帧分辨率约为500×300。该数据同样关注人体、物体的空间位置及其相互关系。

3.1.2 评价指标

实验采用Recall(R@10/R@20/R@50)、median Recall(mR@10/mR@20/mR@50)、F-measure(F@10/F@20/F@50)来评估各方法在数据集上的场景图生成性能。其中,mR@k表示在预测的前k个关系中,能够正确召回的真值关系的中位数比例。

R@k指标衡量的是Recall的前k个预测结果中,与正确召回的真值关系的比例,Recall计算公式如下:

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{34}$$

其中:TP、FN表示模型正确预测为正类和负类的样本数量。

mR@k指标衡量的是在Median Recall的前k预测结果中,能够正确召回的真值关系的中位数比例,计算步骤如下:

$$\mathrm{MedRecall}=\mathcal{M}\left(\mathrm{Recall}\right) \tag{35}$$

其中:$\mathcal{M}$ 表示取中位数。

F-measure指标是Precision和Recall的调和平均值,用于衡量模型的综合性能。

$$\mathrm{F-measure}=2\times\frac{\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{36}$$

3.1.3 实现细节

实验平台为Ubuntu20.04操作系统,NVIDIA3090 GPU加速,内存24 GB,实现框架使用Python3.8和PyTorch 1.9.1与CUDA11.4和cuDNN8.4.1,AdamW优化器对模型进行优化。为了保证实验的公平性,实验中各方法的比较在相同配置下进行:保持batch size设置为8,学习率1×10^{-4}、训练迭代12轮。为了验证所提方法的有效性和先进性,选取的对比方法均为相似配置可运行的主流算法,包括基于原型的嵌入网络(Prototype-based Embedding Network,PENet)[13]、分组协同学习方法(Group Collaborative Learning,GCL)[12]、双重偏执预测器(Dual-Biased Predictor,DBiased-P)[21]、双分支混合学习网络(Dual-branch Hybrid Learning network,DHL)[11]、噪声标签修正与训练策略(NoIsy label CorrEction and Sample Training strategy,NICEST)[10]、分组查询特化及质量感知多重分配(Groupwise query SPEciAlization and Quality-aware multi-assignment,SpeaQ)[19]、基于二分图网络的场景图生成方法(Graph Generation using Bipartite Graph Network,DG-BGN)[18]、单阶段端对端的动态场景图生成方法(One-stage End-to-end Dynamic scene graph generation,OED)[20]、自适应自训练的场景图生成方法(adaptive Self-Training framework for fine-grained Scene Graph Generation,ST-SGG)[9]。本文共设计了7组实验,包含与其他主流方法的对比、场景图生成主观结果展示、热力图分析、混淆矩阵分析、不同特征提取模型的影响、消融研究,以全面说明本方法对网络直播视频场景图生成的有效性和先进性。

3.2 与其他主流方法的性能比较

3.2.1 精度对比

为了验证所提VCO-SGG在直播视频场景图生成任务上的有效性,首先在自建的BJUT-LSGD数据集上与主流方法进行性能比较。如表1所示,本方法在

R@10、R@20 和 R@50 指标上均取得最优结果，分别达到了 55.41%、68.11% 和 76.85%，较次优方法（OED 与 ST-SGG）提升 0.29%、0.5% 与 1.04%。此外，本方法在 mR@10、mR@20 和 mR@50 也优于 OED 和 DG-BGN，分别达到 34.82%、42.43% 和 49.39%。这一优势主要得益于本文构建的 VLM 凸优化架构，相比 ST-SGG、OED、DG-BGN 等采用的基准 VLM 架构，能够更稳定地收敛至全局最优，从而更准确地建模对象间的关联关系。在综合性能方面，本方法的 F@10、F@20 和 F@50 分别为 40.91%、47.76% 和 54.18%，高于 NICEST（36.82%、42.06%、47.34%）。这是由于 NICEST 采用经典的时间-空间特征提取与融合策略，难以抽取出场景关键元素的抽象表示作为建模基准，而本方法通过原型记忆机制向关键帧中补充了准确且稳定的时序信息。上述结果表明，本方法能够有效抑制直播视频中冗余对象的干扰，显著提升场景图生成质量。

表 1 在公开数据集 BJUT-LGSD 上的性能比较 单位:%

Table 1 Performance comparison on the BJUT-LGSD dataset unit:%

方法	精度↑								
	R@10	R@20	R@50	mR@10	mR@20	mR@50	F@10	F@20	F@50
PENet[13]	46.70	55.49	65.00	24.23	32.42	39.94	34.93	40.08	43.82
GCL[12]	46.91	55.85	65.56	24.79	33.06	39.82	35.03	40.22	43.85
DBiased-P[21]	47.05	58.14	66.23	27.07	35.39	40.62	35.68	41.02	44.57
DHL[11]	49.94	58.66	68.05	27.62	34.13	40.68	35.98	42.33	44.63
NICEST[10]	52.63	62.24	70.54	28.38	36.83	43.15	36.82	42.06	47.34
SpeaQ[19]	53.19	64.03	73.54	30.21	38.05	45.36	38.04	44.05	49.76
DG-BGN[18]	54.77	65.68	74.34	33.12	41.38	47.36	39.66	46.03	51.95
OED[20]	55.12	67.32	74.91	33.78	41.25	48.10	39.90	46.59	52.77
ST-SGG[9]	55.10	67.61	75.81	34.46	42.30	48.52	40.46	47.26	53.23
VCO-SGG(所提方法)	**55.41**	**68.11**	**76.85**	**34.82**	**42.43**	**49.39**	**40.91**	**47.76**	**54.18**

注：粗体代表最优，下划线代表次优。

在 Mini Charades 数据集上的对比结果如表 2 所示。本方法在 R@10、R@20 和 R@50 上分别达到了 48.19%、56.67% 和 66.55%，均优于其他方法，较次优的 ST-SGG 高出 0.71%、0.78% 和 0.54%。这同样得益于所提凸优化架构带来的稳定收敛特性。同时，本方法的 mR@10、mR@20 和 mR@50 指标也取得了最高值（28.02%、35.81%、42.64%），相比次优方法（OED 与 DG-BGN）至少高出 0.73%、0.69% 和 0.22%。该优势源于本文方法引入的原型记忆模块，其增强了场景关键元素的表征能力和连续建模。在 F@10、F@20 和 F@50 上，本方法达到 34.48%、40.79% 和 46.78%，略优于 ST-SGG（34.12%、40.68% 和 46.54%）。不同于 ST-SGG 采用的静态特征联合方式，本方法通过特征联合与关系筛选策略有效剔除了冗余索引特征，降低了对象数量波动对生成过程的影响。

表 3 展示了 Mini Action Genome 数据集上的实验结果。本方法在 R@10、R@20 和 R@50 上分别达到 43.42%、53.12% 和 63.06%，较次优的 ST-SGG 提升 0.45%、0.43% 和 0.79%，再次体现了凸优化设计的优势。在 mR@10、mR@20 和 mR@50 上，本方法亦取得最优结果（26.02%、33.81% 和 40.88%），相比 OED 和 ST-SGG 至少高出 0.95%、0.47% 和 0.69%，进一步验证了原型记忆模块在特征表征和连续建模方面的作用。此外，本方法的 F@10、F@20 和 F@50 分别为 34.44%、40.14% 和 44.74%，均优于 ST-SGG（34.07%、39.47%、44.61%）。ST-SGG 在特征联合过程使用固定数量的查询来表征视频对象，而本方法通过 ReLU 机制筛选冗余查询，从而获得了更好的泛化性能。

3.2.2 效率分析

为了分析关于直播视频场景图生成的效率，本实验比较并展示了不同方法在各个数据集上的每秒可以执行的浮点运算次数（Giga Floating Point Operations Per Second，GFLOPs）、参数量（Parameter）和每秒帧数（Frame Per Second，FPS）对比。如图 6 所示，本方法的 GFLOPs 和 Parameter 分别达到了 42.74×10^9 和 63.37×10^6，相比精度次优的 ST-SGG 降低了 3.9×10^9 和 114.2×10^6。这是因为 ST-SGG 需要计算每个类别概率的置信度，而本方法的 VLM 凸优化框架根据类别平均概率来调整特征分布，获得了更高的计算效率。此外，虽然本方法的 GFLOPs 和 Parameter 相比效率最高的 NICEST 高了 6.81×10^9 和 22.54×10^6，但是准确率 R@10、mR@10 和 F@10 也至少提高了 2.78%、2.95% 和 4.09%。这是因为 NICEST 采用 KNN 筛选冗余特征，该方法相比特征关联具有更低的复杂度，然而 KNN 难以随环境变化更新筛选范围，限制了场景图生成的准确性。总体而言，所提方法在精度-效率方面取得了更好的平衡。

表2　在公开数据集Mini Charades上的性能比较　　单位:%

Table 2　Performance Comparison on the Mini Charades Dataset　　unit:%

方法	精度↑								
	R@10	R@20	R@50	mR@10	mR@20	mR@50	F@10	F@20	F@50
PENet[13]	36.38	47.37	56.32	18.22	26.67	34.21	25.01	31.73	37.66
GCL[12]	38.04	48.84	56.89	19.06	27.32	34.89	25.51	32.31	38.28
DBiased-P[21]	40.06	50.81	58.46	19.12	27.51	34.94	25.52	32.22	38.33
DHL[11]	40.46	48.71	60.59	22.61	30.03	37.07	28.10	34.76	40.67
NICEST[10]	43.82	54.57	63.17	25.07	33.37	38.96	30.28	37.06	42.75
SpeaQ[19]	43.85	54.73	63.86	26.02	34.36	40.68	31.87	38.72	44.63
DG-BGN[18]	45.43	56.18	64.35	26.54	34.80	41.32	32.89	39.59	45.33
OED[20]	45.93	56.37	65.12	26.84	33.44	41.67	33.71	40.40	45.72
ST-SGG[9]	47.48	55.89	66.01	27.29	35.12	42.42	34.12	40.68	46.54
VCO-SGG(所提方法)	**48.19**	**56.67**	**66.55**	**28.02**	**35.81**	**42.64**	**34.48**	**40.79**	**46.78**

注:粗体代表最优,下划线代表次优。

表3　不同方法在Mini Action Genome的性能比较　　单位:%

Table 3　Performance Comparison on the Mini Action Genome Dataset　　unit:%

方法	精度↑								
	R@10	R@20	R@50	mR@10	mR@20	mR@50	F@10	F@20	F@50
PENet[13]	30.84	41.83	51.63	17.81	25.97	32.28	19.16	25.86	36.18
GCL[12]	32.39	43.22	51.83	18.55	26.96	34.43	23.41	30.09	31.95
DBiased-P[21]	32.89	43.08	52.57	20.08	26.75	33.23	24.16	30.89	36.80
DHL[11]	35.63	46.16	52.71	21.11	27.66	33.27	24.77	31.68	37.85
NICEST[10]	36.08	44.42	55.94	21.28	29.55	37.07	26.23	33.04	39.09
SpeaQ[19]	40.39	51.02	59.17	21.52	29.89	36.91	27.41	34.17	40.09
DG-BGN[18]	41.20	51.94	60.83	24.04	32.57	39.59	32.94	38.27	42.89
OED[20]	42.89	52.58	61.78	24.27	32.45	40.19	33.83	39.11	43.57
ST-SGG[9]	42.97	52.69	62.27	25.07	33.34	40.00	34.07	39.47	44.61
VCO-SGG(所提方法)	**43.42**	**53.12**	**63.06**	**26.02**	**33.81**	**40.88**	**34.44**	**40.14**	**44.74**

注:粗体代表最优,下划线代表次优。

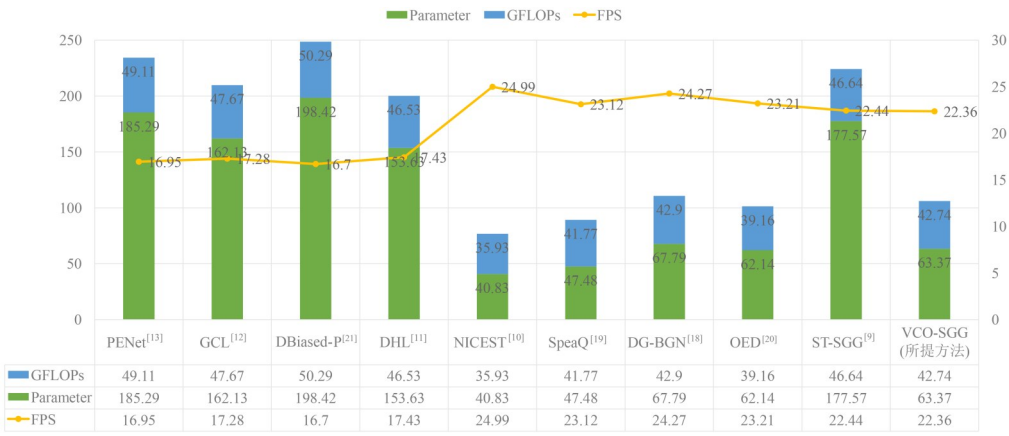


图6　不同方法在各数据集上的效率比较

Figure 6　Effective comparison with other methods

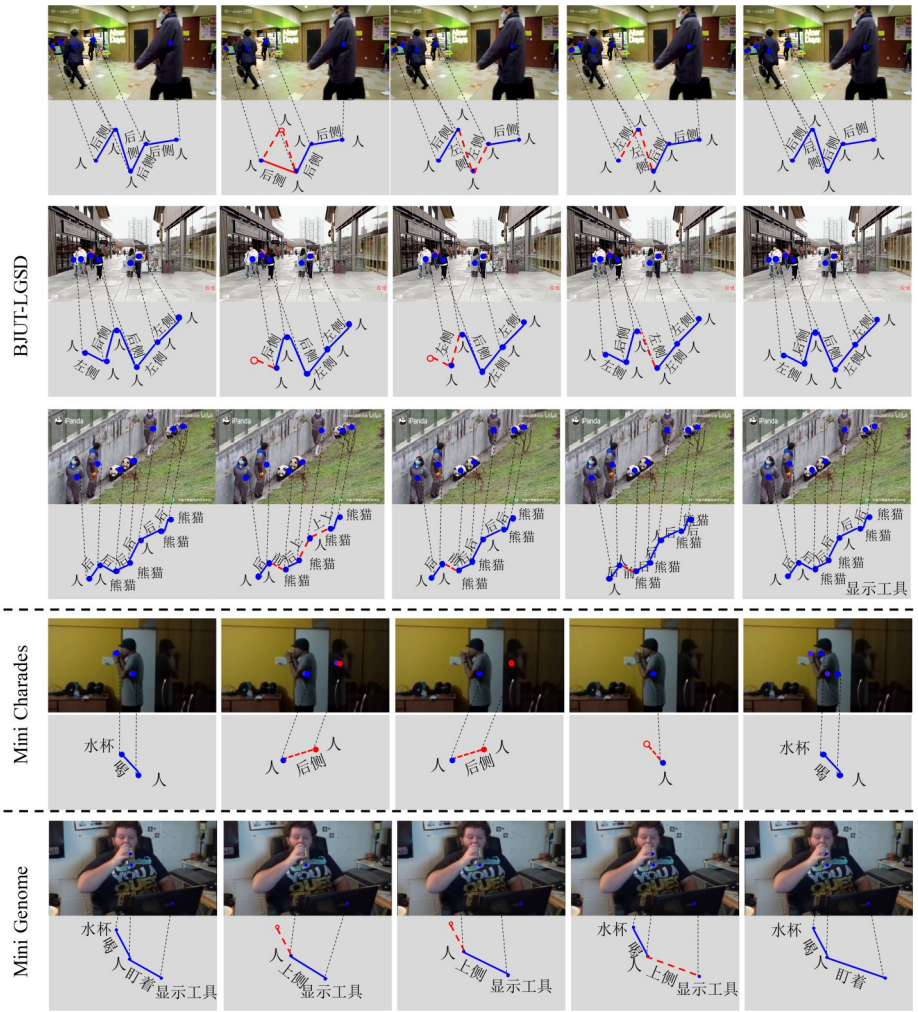
3.3　场景图生成主观结果

3.3.1　与其他方法比较

为直观展示所提VCO-SGG方法对直播视频时序不连贯、内容无规律波动等特性的适应能力,图7呈现了多组场景图的对比结果。

在直播场景数据集BJUT-LSGD中,所有方法在识

别小尺度、远距离目标(如行人)时均面临较大挑战,例如对于远处行人(第1行),DG-BGN容易丢失目标。但即使在目标相互遮挡的情况下(第2行),本方法仍能准确关注到最左侧的行人。此外,在目标密集场景下(第3行),本方法能够建立更清晰、完整的关联关系,而ST-SGG则出现目标漏检。以上结果表明,通过结合查询索引与激活函数来区分对象与背景的机制,有效提升了模型对复杂、不稳定视频流中的场景理解能力。总体而言,可视化结果直观说明,本方法通过凸函数优化自适应平滑特征分布波动,增强了对时序不连贯、内容无规律的直播视频的建模能力,从而生成更丰富、稳定的场景图。



(a) 原图 (b) DG-BGN (c) OED (d) ST-SGG (e) VCO-SGG(所提方法)

(a) Origin (b) DG-BGN (c) OED (d) ST-SGG (e) VCO-SGG (Proposed method)

图7 不同方法的主观结果比较

Figure 7 Visual results of different methods

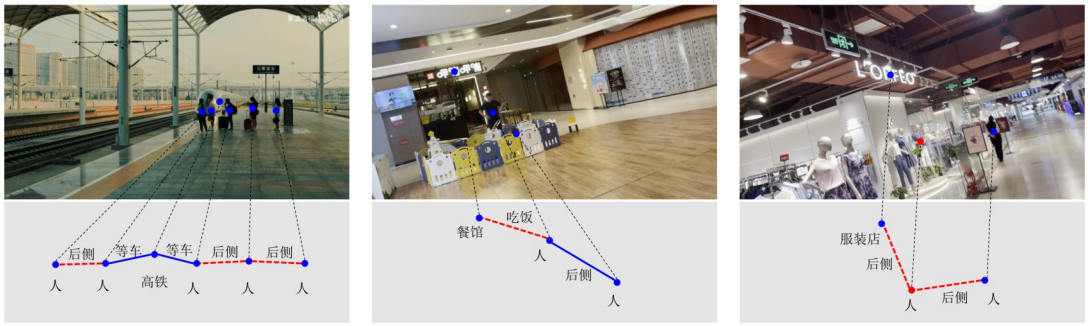
在Mini Charades数据集上,本方法能有效识别细粒度水杯等手持物,而对比方法(DG-BGN、OED、ST-SGG)则存在遗漏。面对镜子反射等干扰(第4行),当镜像完整时,OED与ST-SGG易将反射误判为独立对象,而本方法仍能保持稳定识别。这得益于所提出的VLM凸优化架构,缓解了不同视频数据的特征分布差异,从而在动态直播视频中学习到更具判别性和关联性的特征表示。

在Mini Action Genome数据集上的结果进一步验证了本方法的鲁棒性。在“边喝水边看屏”这类多对象交互的复杂瞬间(第5行),对比方法常将水杯或者显示器误判为背景,而本方法仍能捕捉对象及其关系,体现了其对内容动态波动的适应能力。

3.3.2 失败案例分析

分析一些失败案例有利于获知所提方法的局限性,为后续改进提供参考。如图8(a)中,五个人正在等车,但其中三个人的状态被误识为位于后侧,这表明模型在判定对象时通常基于最近的对象进行考虑。

在图8(b)中,商场场景中的人物后方出现了餐厅场景,因此被错误地认为他正在吃饭。此外,在服装店场景中[图8(c)],包含了顾客和模特模型,其中一部分模型被误识为顾客,反而干扰了判别结果。这些失败案例表明,对象间的相互干扰对场景分类性能有较大影响,也展示了本方法在处理复杂场景时的挑战和局限性。尽管本方法在许多情况下能够准确地生成场景图,但在某些特定场景中,如背景信息复杂或对象间存在相互干扰时,仍可能出现误判。



(a) 对象位置干扰　(b) 背景场地干扰　(c) 相似物体干扰

(a) Object position interference　(b) Background scene interference　(c) Similar object interference

图8　失败案例分析

Figure 8　Several failure cases

3.4　热力图分析

本实验进一步通过注意力热力图分析随走随播场景下的模型性能,从图9可以看出,本方法能够随着镜头移动动态调整关注区域。在图9(a)所示的公园直播场景中,即使镜头持续移动,模型仍能稳定追踪白衣男子及其周围行人。这得益于本文在时间维度引入的动态原型记忆模块:尽管视频流中对象随帧变化,但关键帧的历史信息可及时补充至当前帧,从而有效缓解模型因时序推移而遗忘重要的对象关联信息。图9(b)和(c)展示了固定镜头下车流场景的注意力变化,模型关注区域主要由车辆行驶轨迹所引导。例如在图9(b)中,当一辆白色汽车从右侧远处驶近时,本文方法的关注点从街道中心逐渐转移并集中至该车辆。而在车流密集场景中[图9(c)],模型不仅能优先捕捉体积较大的显著车辆,还能有效关注到远处分布零散的车辆。这一能力源于本文采用的特征联合与关系筛选策略,该策略通过激活函数剔除冗余的对象索引特征,抑制背景噪声对场景图生成的干扰,从而使模型更加聚焦于实例对象本身。

3.5　混淆矩阵分析

接下来,本实验通过展示不同数据集的混淆矩阵进一步评估所提VCO-SGG方法在各类别上的性能。图10(a)为BJUT-LSGD数据集上的结果,可见本方法的预测结果更多集中在对角线,误分类至非对角线类别的比例相对较低,表明其在自建直播数据上具有较高的精度与召回率。同时,主对角线上的高灰度值反映出模型在这些类别中分类准确性较高。这些结果整体验证了本方法所融合的VLM凸优化架构、动态原型记忆模块以及特征联合与关系筛选策略的有效性。

图10(b)为Mini Charades数据集上的混淆矩阵,本方法在“行走”等常见类别上表现出较高的识别准确性。这主要得益于VLM凸优化架构在特征提取过程中能有效建模语义关联,从而提升了模型在动态视频中的鲁棒性。

在图10(c)所示的Mini Action Genome数据集混淆矩阵上,本方法在“攀爬”“位于后侧”等空间关系类别上具有更优的判别精度。这一优势源于动态原型记忆模块能够充分挖掘视频中的动态场景信息,增强对关键元素的表征和连续建模能力。

3.6　不同特征提取模型的影响

本实验比较并分析了不同的特征提取模型对场景图生成性能的影响。如图11所示,仅使用ResNet50的mR@10、mR@20和mR@50指标分别达到了33.71%、41.19%和48.08%;在加入CLIP之后,mR@10、mR@20和mR@50分别提升至33.73%、41.20%、48.15%,验证了VLM凸优化框架对于直播视频场景图生成的有效性;当使用ResNet101时,模型性能得到了进一步提升并达到了最高的34.82%、42.43%、49.39%。可见,随着特征提取网络的加深,模型可以帮助网络学习到更丰富的特征信息,从而提高场景图生成的准确性。这得益于VLM模块建模了复杂场景下不同对象间的语义关联,同时丰富的场景信息使得模型获得更高的场景图生成效果。

3.7　收敛性与稳定性分析

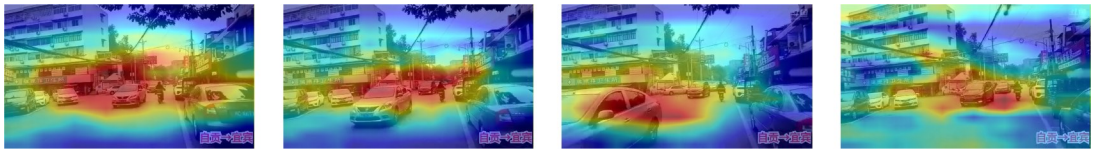
为了验证本方法的收敛性和稳定性,实验对比了

(a) 画面跳跃

(a) Frame jitter

(b) 对象无规律波动

(b) Irregular object fluctuation

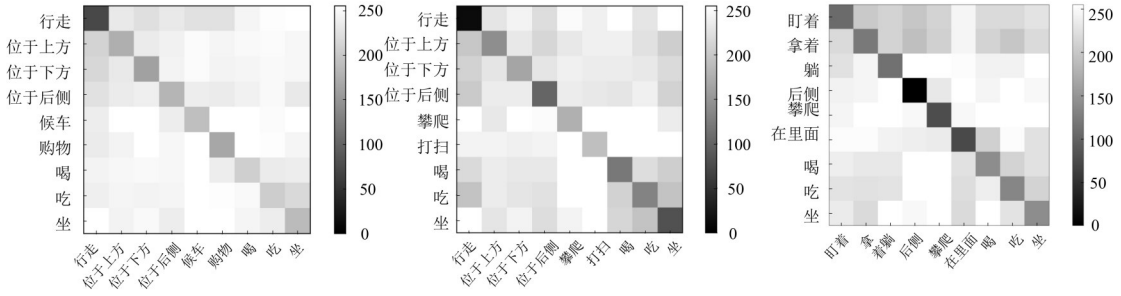


(c) 密集对象

(c) Dense objects

图9 场景图生成热力图

Figure 9 SGG heatmaps



(a) BJUT-LGSD数据集

(a) BJUT-LGSD dataset

(b) Mini Charades数据集

(b) Mini Charades dataset

(c) Mini Action Genome数据集

(c) Mini Action Genome dataset

图10 在不同数据集上的混淆矩阵

Figure 10 Confusion matrix on different datasets

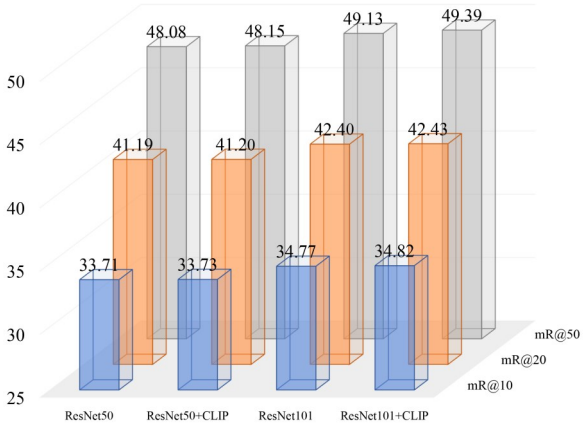


图11 不同特征提取策略对模型性能的影响

Figure 11 Feature extraction of different backbone

采用与不采用凸优化方法的损失曲线，结果如图12所示。使用凸优化方法相比基线具有更低的初始损失值，且在训练前期表现出更大的波动幅度，主要源于本方法中的ResNet和VLM初始特征分布差异所引起的性能震荡。随着训练轮数的增加，凸优化方法的训练损失逐渐降至基线以下。该现象表明ResNet与VLM的特质分布逐渐实现互补，由此可见，利用VLM进行凸优化架构设计，能有效缓解不同视频数据的特征分布差异，从而逐步增强模型训练的收敛性与稳定性。

3.8 消融研究

为了验证所提方法三个关键组件（VLM凸优化架构、动态原型记忆模块、特征联合与关系筛选策略）

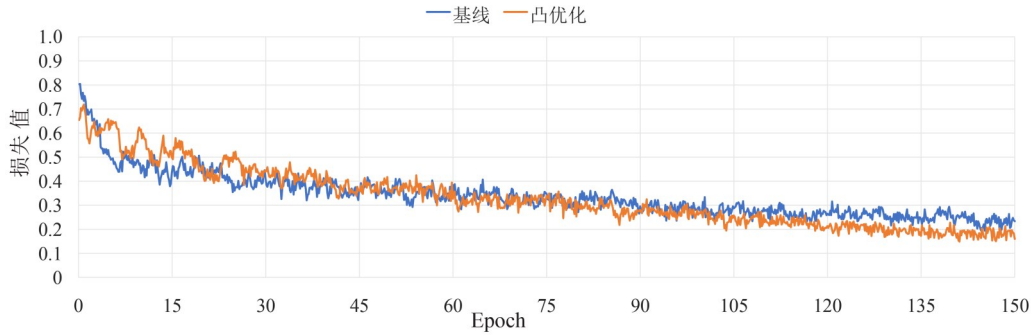


图12　凸优化方法的损失曲线

Figure 12　Loss curve for the convex optimization method

的贡献,本实验通过消融研究分析了不同组件对场景图生成性能的影响。如表4所示,在未使用任何模块的基线设置下,模型在mR@10、mR@20和mR@50指标分别达到33.78%、41.25%和48.10%。在此基础上,引入动态原型记忆模块(移除VLM)后,三项指标分别提升0.18%、0.34%和0.26%,表明该模块可通过补充关键帧中的对象信息,缓解不同目标间的相互干扰。进一步引入特征联合与关系筛选策略后,性能提升至34.35%、42.12%和49.02%,对应增幅为0.57%、0.87%和0.92%,说明该策略可有效抑制复杂背景的干扰。

表4　不同模块对场景图生成性能的影响　　单位:%

Table 4　Comparison with different components　　unit:%

模块			精度↑		
VLM凸优化架构	动态原型记忆	特征联合与关系筛选	mR@10	mR@20	mR@50
×	×	×	33.78	41.25	48.10
	√	×	33.96 (+0.18)	41.59 (+0.34)	48.36 (+0.26)
	√	√	34.35 (+0.57)	42.12 (+0.87)	49.02 (+0.92)
√	×	×	34.03 (+0.25)	41.54 (+0.29)	48.37 (+0.27)
	√	×	34.48 (+0.70)	42.06 (+0.81)	48.94 (+0.84)
	√	√	**34.82** (+1.04)	**42.43** (+1.18)	**49.39** (+1.29)

注:加粗表示最优性能,括号内绿色数字表示提升的性能幅度。

在单独引入VLM凸优化架构,模型在以上三项指标上分别达到34.03%、41.54%和48.37%,较基线提升0.25%、0.29%和0.27%。这一改进主要得益于该架构能够建立对象特征之间的关联,从而增强整体特征的表征能力。将VLM凸优化架构与动态原型记忆模块结合后,性能进一步上升至34.48%、42.06%和48.94%。在此基础上引入特征联合与关系筛选策略,最终在mR@10、mR@20和mR@50取得了最佳结果,分别为34.82%、42.43%和49.39%,相比基线增幅达1.04%、1.18%、1.29%。

上述实验结果充分表明,VLM凸优化架构、动态原型记忆模块、特征联合与关系筛选策略能有效提升模型在直播视频场景图生成任务上的性能,且三者表现出明显的互补性,共同推进了整体性能的提升。

4　结束语

本文提出了一种基于VLM凸优化架构的视频场景图生成方法。现有的视频场景图生成方法主要受到视频场景中元素复杂多样的影响,导致生成结果的准确性受到限制。与此不同,首先利用VLM提取帧特征,并结合凸优化策略解析特征中的对象,提升模型解析场景内各元素特征语义信息的能力;然后设计了动态原型记忆来提取和优化帧语义,增强模型对帧序列的表征能力;最后嵌入了特征联合与关系筛选策略,通过筛选与过滤冗余的对象来实现场景图生成,从而降低复杂元素信息对生成质量的干扰。通过在一个自建数据集和两个公共数据集上全面评估本方法的性能,本方法在BJUT-LGSD上的R@10与mR@10值达到了最高的55.41%、34.82%,并且在Mini Cha-

rades 和 Mini Action Genome 上分别达到 48.19%/28.02%、43.42%/26.02%。推理速度保持在 22.36 FPS，在对比方法中相对较优，说明对直播视频场景图生成的有效性和优越性。

在未来的工作中，将会继续对 VLM 结构进行优化，以提高视频场景图生成的准确性。具体而言，实验部分的失败案例表明，现有方法在处理对象间关联时过于依赖局部范围内的信息，而全面考虑所有对象两两之间的关系会导致计算量过大。因此，如何平衡效率和准确率间的关系是亟待解决的问题。后续可以尝试引入一个邻域筛选机制以适当补充网络提取的全局信息，提升网络对生成结果的准确率和稳定性。此外，基于激活函数来筛选对象索引虽然展现出了一定的效果，但是在筛选过程中会不可避免地清除索引中的有效信息。观察到 Transformer 中的分组策略可以在不增加计算量的基础上将完整的任务拆解为多个小任务，因此尝试将该策略引入特征联合与关系筛选过程中，实现冗余信息的精准去除。本方法针对直播环境中实时、动态、多变的特点展开研究，通过对复杂视频场景中对象及其关系的表征和识别，为直播平台提供更深层次的内容理解能力，进而为更好地管理和利用其海量的视频数据作出贡献。

参考文献

[1] 国家市场监督管理总局, 直播电商服务质量的信息监测与评价规范[R]. 2024.
State Administration for Market Regulation, Specification for information monitoring and evaluation of live streaming E-commerce service quality[R]. 2024. (in Chinese)

[2] 韩志冬, 胡升龙, 宋慧慧, 等. 运动提示引导自适应学习无监督视频目标分割[J]. 电子学报, 2025, 53(7): 2305-2323.
Han Zhidong, Hu Shenglong, Song Huihui, et al. Motion-prompts guided adaptive learning for unsupervised video object segmentatio[J]. Acta Electronica Sinica, 2025, 53(7): 2305-2323. (in Chinese)

[3] 杨静, 刘成城, 黄洁, 等. 联合时延-多普勒-角度的无源雷达目标定位凸优化算法[J]. 电子学报, 2024, 52(6): 2091-2102.
Yang Jing, Liu Chengcheng, Huang Jie, et al. Convex solution for target localization in passive MIMO radar using delay, Doppler and angle measurements[J]. Acta Electronica Sinica, 2024, 52(6): 2091-2102. (in Chinese)

[4] Jing Shuaiqi, Zhang Haonan, Zeng Pengpeng, et al. Memory-based augmentation network for video captioning[J]. IEEE Transactions on Multimedia, 2024, 26: 2367-2379.

[5] 林丽群, 暨书逸, 何嘉晨, 等. 基于感知和记忆的视频动态质量评价[J]. 电子学报, 2024, 52(11): 3727-3740.
Lin Liqun, Ji Shuyi, He Jiachen, et al. Research of video dynamic quality evaluation based on human perception and memory[J]. Acta Electronica Sinica, 2024, 52(11): 3727-3740. (in Chinese)

[6] Mishra D, Saha P, Zhao H, et al. TIER-LOC: Visual Query-based Video Clip Localization in fetal ultrasound videos with a multi-tier Transformer[J]. Medical Image Analysis, 2025, 103: 103611.

[7] Yang Jingkang, Yizhe Ang, Guo Zujin, et al. Panoptic scene graph generation[C]//2022 European Conference on Computer Vision. Cham: Springer, 2022: 178-196.

[8] Khandelwal A. FloCoDe: Unbiased dynamic scene graph generation with temporal consistency and correlation debiasing[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2024: 2516-2526.

[9] Kim K, Yoon K, In Y, et al. Adaptive self-training framework for fine-grained scene graph generation[C]//12th International Conference on Learning Representations. Vienna: ICLR, 2024.

[10] Li Lin, Xiao Jun, Shi Hanrong, et al. NICEST: Noisy label correction and training for robust scene graph generation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(10): 6873-6888.

[11] Zheng Chaofan, Gao Lianli, Xinyu Lyu, et al. Dual-branch hybrid learning network for unbiased scene graph generation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(3): 1743-1756.

[12] Dong Xingning, Gan Tian, Song Xuemeng, et al. Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 19405-19414.

[13] Zheng Chaofan, Xinyu Lyu, Gao Lianli, et al. Prototype-based embedding network for scene graph generation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 22783-22792.

[14] Zhang Haoji, Wang Yiqin, Tang Yansong, et al. Flash-VStream: Memory-based real-time understanding for long video streams[PP/OL]. V2. arXiv (2024-06-30)[2025-07-02]. https://doi.org/10.48550/arXiv.2406.08085.

[15] Cheng Dingxin, Li Mingda, Liu Jingyu, et al. Enhancing long video understanding via hierarchical event-based memory[C]//2025 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE, 2025: 1-6.

[16] Tu Yunbin, Li Liang, Su Li, et al. Query-centric audio-visual cognition network for moment retrieval, segmentation and step-captioning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, 2025: 7464-7472.

[17] Hu Jingjing, Guo Dan, Li Kun, et al. Unified static and dynamic network: Efficient temporal filtering for video grounding[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025. https://doi.org/10.48550/arXiv.2403.14174.

[18] Lu Jiale, Chen Lianggangxu, Guan Haoyue, et al. Improving rare relation inferring for scene graph generation using bipartite graph network[J]. Computer Vision and Image Understanding, 2024, 239: 103901.

[19] Kim J, Park J, Park J, et al. Groupwise query specialization and quality-aware multi-assignment for transformer-based visual relationship detection[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 28160-28169.

[20] Wang Guan, Li Zhimin, Chen Qingchao, et al. OED: Towards one-stage end-to-end dynamic scene graph generation[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 27938-27947.

[21] Han Xianjing, Song Xuemeng, Dong Xingning, et al. DBiased-P: Dual-biased predicate predictor for unbiased scene graph generation[J]. IEEE Transactions on Multimedia, 2023, 25: 5319-5329.

作者简介

李文生　男,1994年2月生,山东淄博人。现为北京工业大学博士研究生。主要研究方向为视频理解。

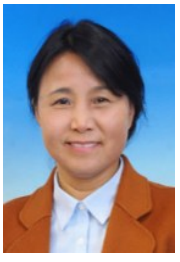
E-mail: liwensheng@emails.bjut.edu.cn

张　菁　女,1975年2月生,广东梅县人。博士,现为北京工业大学教授、博士生导师。主要研究方向为人工智能与计算机视觉等。

E-mail: zhj@bjut.edu.cn

王艺晓　女,2002年6月生,河南安阳人。现为北京工业大学硕士研究生。主要研究方向为视频理解。

E-mail: wyx0504@emails.bjut.edu.cn



卓　力　女,1971年10月生,江苏徐州人。博士,现为北京工业大学教授、博士生导师。主要研究方向人工智能与计算视觉等。

E-mail: zhuoli@bjut.edu.cn