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　　**文献标识码**:A　　**文章编号**:1006-8961(2026)02-0556-17

论文引用格式:Xiao J F, Jiang W H, Fang Y M, Fang C Y and Zhao X W. 2026. Video captioning with trajectory-based spatiotemporal perceiving and adaptive semantic focusing. Journal of Image and Graphics, 31(2):0556-0572(肖景富, 姜文晖, 方玉明, 方承炀, 赵小伟. 2026. 融合轨迹时空感知与自适应语义聚焦的视频描述方法. 中国图象图形学报, 31(2):0556-0572)[DOI:10.11834/jig.250266]

融合轨迹时空感知与自适应语义聚焦的视频描述方法

肖景富[1],姜文晖[1*],方玉明[1],方承炀[1],赵小伟[2]

1. 江西财经大学计算机与人工智能学院,南昌　330032; 2. 盛景智能科技(嘉兴)有限公司,嘉兴　314502

摘　要:**目的**　视频内容描述任务旨在自动生成自然语言句子,精准表达视频视觉语义信息。尽管编码器—解码器方法在视觉表达与语言生成上已有进展,但视频编码器难以建模目标级运动与事件,解码器也难以实现跨模态语义对齐,限制了生成文本质量。为此,提出融合轨迹时空感知与自适应语义聚焦的方法,以增强目标运动建模能力并改善多模态语义对齐。**方法**　首先,提出基于点轨迹的视觉特征聚合方法,通过时空建模生成兼具空间外观与时间连续性的轨迹特征,并与局部运动特征融合,以增强模型在运动和形变场景下的目标追踪能力和语义连贯性;同时,设计无监督自适应关键轨迹聚焦学习方法,利用密集点轨迹动态信息,通过注意力权重自适应筛选关键轨迹并引入聚焦损失,引导模型优先关注关键语义区域、抑制背景干扰,从而提升跨模态语义关联能力。**结果**　在MSR-VTT(Microsoft research video to text)和MSVD(Microsoft research video description corpus)两个公开数据集上进行实验,所提方法在CIDEr(consensus-based image description evaluation)指标上分别取得61.2和130.1的得分,显著优于现有主流方法,验证了所提方法在描述准确性与语义丰富性方面的有效性。定性分析表明,该方法在提升描述的时序连贯性和语义表达能力方面表现优异。**结论**　本文方法有效提升了视频描述模型在复杂动态环境下的目标语义连续性建模能力,并通过无监督的自适应关键轨迹聚焦学习方法改善了注意力机制对视频与文本语义关联的能力。

关键词:视频内容描述;多模态语义关联;时空点轨迹;特征聚合;自适应语义聚焦

Video captioning with trajectory-based spatiotemporal perceiving and adaptive semantic focusing

Xiao Jingfu[1], Jiang Wenhui[1*], Fang Yuming[1], Fang Chengyang[1], Zhao Xiaowei[2]

1. School of Computer Science and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang　330032, China;
2. Shengjing Intelligent Technology (Jiaxing) Co., Ltd., Jiaxing　314502, China

Abstract: **Objective** The video captioning task aims to automatically generate natural language sentences that accurately convey the semantic content of videos, serving as a crucial intersection between visual understanding and natural language processing. It has shown broad potential in practical applications such as automated surveillance, assistive technologies for the visually impaired, and video summarization. While most existing approaches adopt an encoder-decoder architecture

收稿日期:2025-06-18;**修回日期**:2025-09-05;**预印本日期**:2025-09-12

*通信作者:姜文晖　wenhui@jxufe.edu.cn

基金项目:国家自然科学基金项目(62562035);江西省重点研发计划资助(20252BCE310034);江西省双千计划项目(jxsq2023101092);江西省自然科学基金项目(20252BAC200182,20242BAB23012);江西省职业早期青年科技人才项目(20252BEJ730121)

Supported by: National Natural Science Foundation of China(62562035); Key R&D Program of Jiangxi Province, China(20252BCE310034); Double Thousand Talents Program of Jiangxi Province (jxsq2023101092); Natural Science Foundation of Jiangxi Province, China (20252BAC200182, 20242BAB23012); Early-Career Program for Young Science and Technology Talents of Jiangxi Province (20252BEJ730121)

and have achieved notable progress in visual representation and language generation, they still face two key challenges. First, the video encoder often struggles to effectively model object-level motion and event information, which are essential for expressing action-related semantics. Real-world videos frequently involve complex dynamic processes, such as rapid motion and object deformation, making it difficult for models to capture spatiotemporal continuity. Second, the decoder encounters difficulties in cross-modal semantic association during training, as the model often fails to consistently focus on visual regions that are highly relevant to the current word being generated. The mapping between visual and textual modalities is inherently complex and unstable, particularly when multimodal alignment is weak. These limitations severely undermine the quality and accuracy of generated captions in dynamic scenes. To address these issues, this study proposes a novel video captioning method that integrates trajectory-aware spatiotemporal modeling with adaptive semantic focusing, aiming to enhance the model's ability to capture object-level dynamics and improve multimodal semantic alignment. **Method** To overcome the identified challenges, this study introduces a video description approach that leverages video point trajectories as a core constraint to enhance semantic continuity and multimodal alignment. The methodology comprises two innovative components designed to address the limitations of existing approaches. First, a visual feature trajectory aggregation method is developed to improve the modeling of target semantic continuity. This method explicitly captures the spatiotemporal semantics of target objects by aggregating visual features along point trajectories, which represent the temporal positions and movements of objects within a video. Through an average pooling operation, features along these trajectories are integrated to generate smooth and stable trajectory representations that preserve spatial appearance and temporal continuity. This approach surpasses traditional methods relying on frame-level feature extraction or independent sequence modeling, which often fail to maintain semantic consistency in scenarios involving occlusions or multiobject interactions. By focusing on continuous representations of individual targets, the trajectory aggregation method enables the model to track objects across frames, capture dynamic associations, and produce descriptions with enhanced temporal coherence and robustness, particularly in complex scenes with frequent motion or visual obstructions. Second, an unsupervised adaptive key trajectory focusing learning strategy is proposed to mitigate semantic misalignment between visual and textual modalities. This method exploits the dynamic spatiotemporal information embedded in dense video point trajectories to uncover latent semantic associations between visual and textual elements. By analyzing the distribution of attention weights during decoding, it adaptively identifies key trajectories——those most relevant to the generated description——and distinguishes them from nonkey trajectories that contribute less to semantic content. A novel unsupervised key trajectory focusing loss guides the model to prioritize semantically significant regions while suppressing interference from irrelevant background information. Unlike conventional attention mechanisms that either struggle to align words with appropriate visual regions or rely on labor-intensive manual annotations, this approach is annotation-free and avoids dependence on fixed geometric regions, such as rectangles. Instead, it allows semantic alignment regions to flexibly adapt to the dynamic boundaries and complex shapes of target objects, ensuring precise and contextually relevant attention. The adaptive key trajectory selection process further adjusts dynamically to changes in attention weights as each word is generated, thereby enhancing the flexibility and accuracy of semantic alignment. Together, these methods provide a comprehensive solution that strengthens the model's ability to generate coherent and accurate descriptions by addressing challenges in temporal semantic continuity and multimodal alignment. **Result** The proposed approach was evaluated on two widely used benchmark datasets, Microsoft research video to text (MSR-VTT) and Microsoft research video description corpus (MSVD), to demonstrate its effectiveness in generating accurate and semantically rich video descriptions. On the MSR-VTT dataset, the model achieved a consensus-based image description evaluation (CIDEr) score of 61.2, surpassing several leading methods such as IcoCap (60.2) and OmniViD (56.6), and attained a metric for evaluation of translation with explicit ordering (METEOR) score of 32.0. These results indicate that our method——particularly the introduced unsupervised focusing loss——effectively enhances the semantic association between fine-grained spatiotemporal video features and words. Moreover, the model outperformed approaches that rely on large-scale vision-language pretraining, including MELTR, VL-Prompt, and OmniViD, with CIDEr improvements ranging from +4.6 to +11.2, highlighting the practical advantages of the proposed trajectory-guided attention mechanism. On the MSVD dataset, which features more fine-grained and diverse video content, our model consistently outperformed all baselines across every evaluation metric. In particular,

it achieved a BLEU@1 score of 88.6, a BLEU@4 score of 66.0, a METEOR score of 42.5, a ROUGE-L score of 80.1, and an outstanding CIDEr score of 130.1. These scores represent a 7.6-point gain over OmniViD and a 1.6-point gain over VL-Prompt on CIDEr, underscoring the robustness and effectiveness of our model in handling dynamic video scenarios. Beyond quantitative metrics, qualitative analyses reveal that the generated descriptions exhibit strong temporal coherence, reduced redundancy, and improved alignment with video semantics. Attention visualizations confirm that the model successfully attends to semantically relevant moving objects while filtering out background clutter. This characteristic mitigates issues of attention dispersion——such as incorrectly identifying background elements as primary subjects——demonstrating the interpretability and precision of the proposed attention mechanism. **Conclusion** This paper presents a robust and interpretable video description framework that addresses two long-standing challenges: insufficient modeling of semantic continuity and inadequate cross-modal semantic alignment. By introducing a trajectory-based visual representation approach together with an unsupervised adaptive key trajectory focusing strategy, the proposed method strengthens the temporal and semantic integrity of video representations while guiding the model toward precise alignment with textual outputs. Requiring no additional annotations and being fully end-to-end trainable, the approach demonstrates strong generalization across multiple datasets and proves particularly effective in complex scenarios involving occlusions, background clutter, and multiobject dynamics. The substantial gains in quantitative performance——achieving CIDEr scores of 61.2 and 130.1 on MSR-VTT and MSVD, respectively——along with qualitative improvements in description quality, underscore the potential of integrating motion-aware semantics and attention guidance into video-language models. Overall, the proposed method provides a scalable and efficient solution for video content description, with broad implications for applications such as automated surveillance, assistive technologies, video summarization, and content retrieval.

Key words: video captioning; multimodal semantic alignment; spatiotemporal point trajectory; feature aggregation; adaptive semantic focusing

0 引　言

视频内容描述任务旨在生成自然语言句子，以准确表达视频中所呈现的视觉内容。作为视觉理解与自然语言处理交叉的重要研究方向，该任务在自动安防监控、视障辅助以及视频摘要生成等实际应用场景中展现出广泛的潜力。

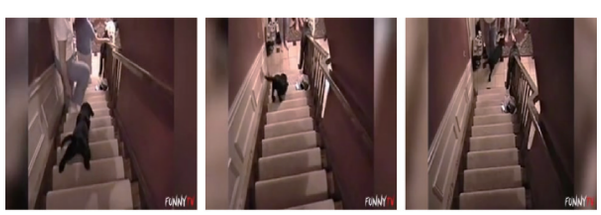
现有的视频内容描述方法通常采用编码器—解码器架构，先通过编码器提取视频的视觉语义特征，再由解码器逐步生成自然语言描述。该架构在提升模型的表达能力和语言生成质量方面取得了一定成效，但视频内容描述任务依然面临两个关键挑战。

首先，视频编码器需要准确理解目标级的运动和事件信息。物体的运动状态及其所经历的事件，往往构成视频语义的核心，是生成自然语言描述中动作相关词汇的关键依据。然而，视频中的关键物体通常跨越多个时间片段，并伴随剧烈运动和形变等复杂动态过程。这些因素使得模型在建模物体行为演化时面临较大挑战，从而难以准确识别和表达与动作相关的语义内容，最终导致生成描述中动作相关词汇的准确性较低。

其次，解码器在训练过程中面临跨模态语义关联困难的问题。由于视频中不同时空区域对最终描述的重要性各不相同，模型需准确聚焦与当前文本单词最相关的视觉区域，才能生成高质量的描述。然而，视频视觉特征与文本之间存在复杂且多变的对应关系，且这种对应关系是隐含的，难以直接建模其关联。引入语义关联监督，能够为模型提供显式的视觉—文本关联信号，指导解码器在生成每个单词时精准定位到与之相关的视频片段或空间区域。但是语义关联监督存在人工标注成本高的局限。

针对目标级的运动和事件信息建模难的挑战，当前主流方法通过建模视频的时空关系提取物体的运动信息。例如，Feichtenhofer等人(2016)提出卷积双流网络，通过双流卷积神经网络(two-stream convolutional neural network, Two-Stream CNN)结构和结合全局池化机制，提取视频序列的局部运动细节和全局空间外观特征。但是该方法仅对空间维度进行全局建模，忽略了时间维度上物体动态演化的连续性，难以有效捕捉跨帧的行为变化。为增强时间建模能力，部分研究(Ul Amin等，2024；Zhou等，2019)采用基于长短期记忆网络(long short-term memory, LSTM)，结合全局池化生成视频的整体时序

特征，同时引入微自编码器或卷积模块提取局部区域的短期动态变化与空间细节。尽管此类方法具备一定的时序建模能力，但 LSTM 对长时序建模能力有限，难以准确捕捉同一物体在长时间跨度内的动态演化。为建模更复杂的时空语义依赖一些研究（Xie 等，2024；Song 等，2018；赵永强 等，2023）通过引入自注意力机制以提取视频的全局时空特征，在一定程度上提升了对跨帧长距离动态关系的建模能力。空间位移难以准确聚焦于同一物体的时空轨迹上，导致模型对目标级的运动和事件的感知能力不足。如图 1 所示，先进模型 CVG（comprehensive visual grounding for video description）（Jiang 等，2024）在复杂动态场景下出现动作相关词汇预测错误的情况。其中，GT（ground truth）表示真实标签，红色表示生成错误的单词。



GT1: a black dog slides down the stairs
GT2: a dog is sliding down a flight of stairs
CVG: a dog is playing with a man

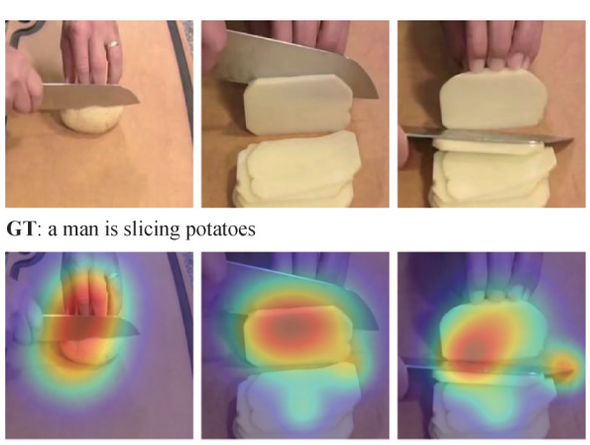
图 1　参考描述和 CVG 方法生成视频描述示例

Fig. 1　Example video captions generated by the reference description and the CVG method

针对跨模态语义关联困难的问题，部分研究尝试隐式监督的形式指导跨模态语义关联。交叉注意力机制因其能够有效引导模型聚焦于描述相关的视觉区域，近年来广泛引入到视频描述任务中（Yao 等，2015；Yan 等，2020；姜文晖 等，2022）。具体而言，交叉注意力机制可在编码阶段为不同的视觉特征分配权重，使模型在解码过程中能够重点关注关键帧与显著区域，从而提高生成句子的相关性和准确性。例如，Tang 等人（2023）提出一种顺序注意力框架，先通过空间注意力提取帧内对象和场景细节，再利用时间注意力建模时序变化，生成逻辑连贯的描述；Zhong 等人（2023）提出在编码器中引入循环区域注意力，挖掘多样化的区域特征，并结合时间注意力实现帧序列建模。尽管这些基于注意力的模型表现良好，研究（Liu 等，2017；Jiang 等，2022）表明，注意力机制仍未能有效地将单词与语义相关的视觉区域对齐，导致语义关联质量较差。为提升模型的语义关联能力，部分研究尝试引入显式监督机制，Liu 等人（2017）和 Zhou 等人（2019）的研究引入人工标注且形状固定的注意力标签，通过监督信号引导模型聚焦关键区域，标注成本高昂。Jiang 等人（2024）提出一种低成本生成注意力伪标注的方法，旨在减轻标注负担。然而，现有方法仍存在明显局限性。如图 2 所示，先进的视频描述生成模型（Jiang 等，2024）虽然能够聚焦语义相关区域，但注意力区域仍然存在偏差，注意力同时分散在土豆和小刀上，导致模型错误地生成了“小刀”，影响最终描述的准确性。值得注意的是，包括 Liu 等人（2017）和 Zhou 等人（2019）的研究在内，现有方法普遍采用从外部模型或人工标注获取注意力标签的技术路线。然而，依赖外部模型会引入额外的计算开销，而人工标注则显著增加成本。此外，语义关联区域通常采用矩形框标注，无法适应不规则或形变的物体。

为应对上述编码器—解码器架构中的两个挑战问题，本文提出一种融合轨迹时空感知与自适应语义聚焦的视频描述方法。该方法引入视频点轨迹作为辅助监督信号，通过挖掘视频中具有语义一致性的轨迹特征，提升模型物体时空上下文建模能力。同时，设计聚焦学习方法引导模型更有效地捕捉关键语义区域。

为了提升模型对物体运动和事件的理解能力，本文提出一种基于视频点轨迹的视觉特征轨迹聚合



GT: a man is slicing potatoes
CVG: a man is cutting a knife

图 2　CVG 的注意力图可视化示例
（红色表示高注意力区域，蓝色表示低注意力区域）

Fig. 2　Attention map visualization example of CVG (red indicates high-attention regions, blue indicates low-attention regions)

方法。视频点轨迹作为物体在时间维度上的时空位置表示，能够精准反映目标在视频不同时间维度上的语义一致性。另外，轨迹主要位于运动显著性较高的区域，其上对应的视觉特征包含丰富的物体时空上下文信息。基于这些特性，该方法将点轨迹上的视觉特征通过平均池化操作进行时空聚合，得到平滑且稳定的轨迹特征。轨迹特征作为物体的全局语义演变，视觉特征作为物体的局部运动细节，二者的结合既保证了语义信息的连续性，又充分捕捉了物体的细节变化，使模型能够更精准、更稳定地聚焦于目标实体。

同时，针对语义关联监督难以获得的问题，本文设计了无监督的自适应关键轨迹聚焦学习方法，用于优化模型的文本—视频语义关联能力。该方法利用视频密集点轨迹提供的动态时空信息，挖掘视觉与文本模态之间的潜在语义关联。具体而言，依据注意力权重的分布和视频密集点轨迹信息，通过自适应关键轨迹挖掘方法，将视频中的密集点轨迹划分为关键轨迹和非关键轨迹。与此同时，对应地设计了一种无监督的关键轨迹聚焦损失，引导模型聚焦关键区域。该损失通过增强关键轨迹在生成描述中的主导作用，引导注意力机制在文本—视频语义关联过程中尽可能多地关注关键轨迹区域，同时抑制非关键轨迹区域的干扰，无须依赖昂贵的注意力监督标注。此外，与现有方法受限于静态区域划分且语义关联区域通常采用固定的矩形形状不同，此方法允许语义关联区域的形状随目标轮廓灵活变化，这一特性使得注意力机制能够更精确地适应目标物体的动态边界和复杂形状。

本文的主要贡献总结如下：1）提出基于视频点轨迹的视觉特征轨迹聚合方法，通过对轨迹上的目标视觉特征进行聚合建模，保留了物体的空间外观信息，还维持了其在时间维度上的上下文连续性，有效增强了模型对物体时空连贯性的建模。2）设计一种无监督的自适应关键轨迹聚焦学习方法，利用点轨迹所携带的时空结构信息，动态挖掘视觉与文本之间的潜在语义关联，引导注意力机制聚焦语义相关区域，同时抑制背景干扰，提升了模型的语义关联能力。3）提出的视频描述方法在MSR-VTT（Microsoft research video to text）（Xu 等，2016）和MSVD（Microsoft research video description corpus）（Chen 和 Dolan，2011）标准数据集上进行了广泛实验，定量和定性比较均验证了该方法的有效性。其在两个数据集上的CIDEr（consensus-based image description evaluation）得分分别达到61.2和130.1，凸显了模型生成描述在准确性和语义丰富性上的显著优势。

1 算法原理和实现

本文提出的融合轨迹时空感知与自适应语义聚焦的视频描述方法总体框架如图3所示。该方法主要包括3个模块，分别是特征编码模块、文本—视频语义关联模块和视频描述生成模块。其中，特征编码模块针对模型输入的文本—视频对数据进行编码表示，为后续处理提供基础特征输入；文本—视频语义关联模块是该模型的核心，通过对齐文本与视频的语义关联，确保模型在生成每个单词时精准聚焦最相关的视频编码特征；最后，视频描述生成模块将视频多源特征进行逐词预测，形成最终的视频描述。

1.1 特征编码模块

1.1.1 视频与文本特征提取

特征编码模块中的视频编码器从输入视频中提取出信息丰富的视觉信息。具体而言，使用强大的视觉编码器作为视频编码器提取视频中包含物体局部运动细节的视觉特征$\boldsymbol{V}$。首先，从每个输入视频中均匀采样F帧，这些视频帧被输入到Video Swin Transformer模型中以计算空间感知的视觉特征。对于每一帧f，Video Swin Transformer产生N个网格特征向量$\boldsymbol{v}_f=[\boldsymbol{v}_f^1,\boldsymbol{v}_f^2,\cdots,\boldsymbol{v}_f^N]$。为了获得视频的视觉特征，将所有帧的特征向量进行拼接，视觉特征向量表示为$\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_G]$，其中$G$是从整个视频中提取的视觉网量的总数，$G=F\times N$，视觉特征$\boldsymbol{V}\in\mathbf{R}^{G\times C}$。

此外，文本编码器将标注文本输入预训练的BERT（bidirectional encoder representations from Transformers）模型，以生成编码后的文本特征$\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_{t-1}]$，$\boldsymbol{Y}\in\mathbf{R}^{(t-1)\times C}$。借助BERT的语义建模能力，可有效捕获上下文信息，为后续任务提供丰富的语义支持。

1.1.2 轨迹特征提取

为了进一步增强目标语义连续性与动态变化过程的显式建模，在特征编码模块中还提取了基于视

频时空轨迹片段聚合的视频中的轨迹特征，为视频的时序建模提供更具物理意义的动态信息。

具体而言，首先采用先进的点跟踪模型估计视频帧间的时空轨迹片段集合，以获取图像中各像素点在视频帧中的位置变化 $\boldsymbol{I}$ 和运动轨迹。过程具体表示为

$$\{\boldsymbol{T}_1,\boldsymbol{T}_2,\boldsymbol{T}_3,\cdots,\boldsymbol{T}_M\} = Q(\boldsymbol{I}) \tag{1}$$

式中，$\boldsymbol{I}$ 表示从视频中均匀采样的 F 帧的集合，$Q(\cdot)$ 表示视频点跟踪模型，用于估算视频中像素点在每帧之间的运动轨迹，本文使用 PIPs++（persistent independent particles）（Zheng 等，2023）估计 F 帧视频帧的时空轨迹片段。其中，每条轨迹片段 $\boldsymbol{T}_i \in \mathbf{R}^{1\times F}$ 是由视觉特征序号构成的集合，用于记录对应像素点或关键区域在 F 帧图像中的位置变化与运动路径，每条时空轨迹片段的长度均满足 $len(\boldsymbol{T}_i) \leqslant F$。

根据时空轨迹片段的时序一致性，通过将视觉特征聚合得到轨迹特征。轨迹特征 $\boldsymbol{R}$ 由同一条时空轨迹片段上的所有视觉特征 $\boldsymbol{V}$ 经过轨迹特征聚合得到，模型中选择平均池化的方式进行轨迹特征聚合，具体过程表示为

$$\boldsymbol{R} = \left[\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_M\right] \tag{2}$$

$$\boldsymbol{r}_i = \sum_{j=1}^{n}\frac{1}{n}\cdot \boldsymbol{v}_i^j \boldsymbol{v}_{\boldsymbol{T}_i}^j \in \mathbf{R}^{1\times C} \tag{3}$$

式中，$\boldsymbol{R} \in \mathbf{R}^{M\times C}$ 表示视频所有时空轨迹片段对应的轨迹特征，表示轨迹集中的时空轨迹片段 $\boldsymbol{T}_i$ 上的第 j 个视觉特征，n 为时空轨迹片段长度。

1.2 文本—视频语义关联模块

本文提出的文本—视频语义关联模块采用基于 Transformer 的浅层架构，该浅层架构主要由多头自注意力（multi-head self-attention，MHA）和位置前馈层组成，其核心创新在于将传统的 MHA 机制改进为轨迹约束注意力模块。具体地，轨迹约束注意力模块主要包含自适应关键轨迹挖掘和基于关键轨迹的

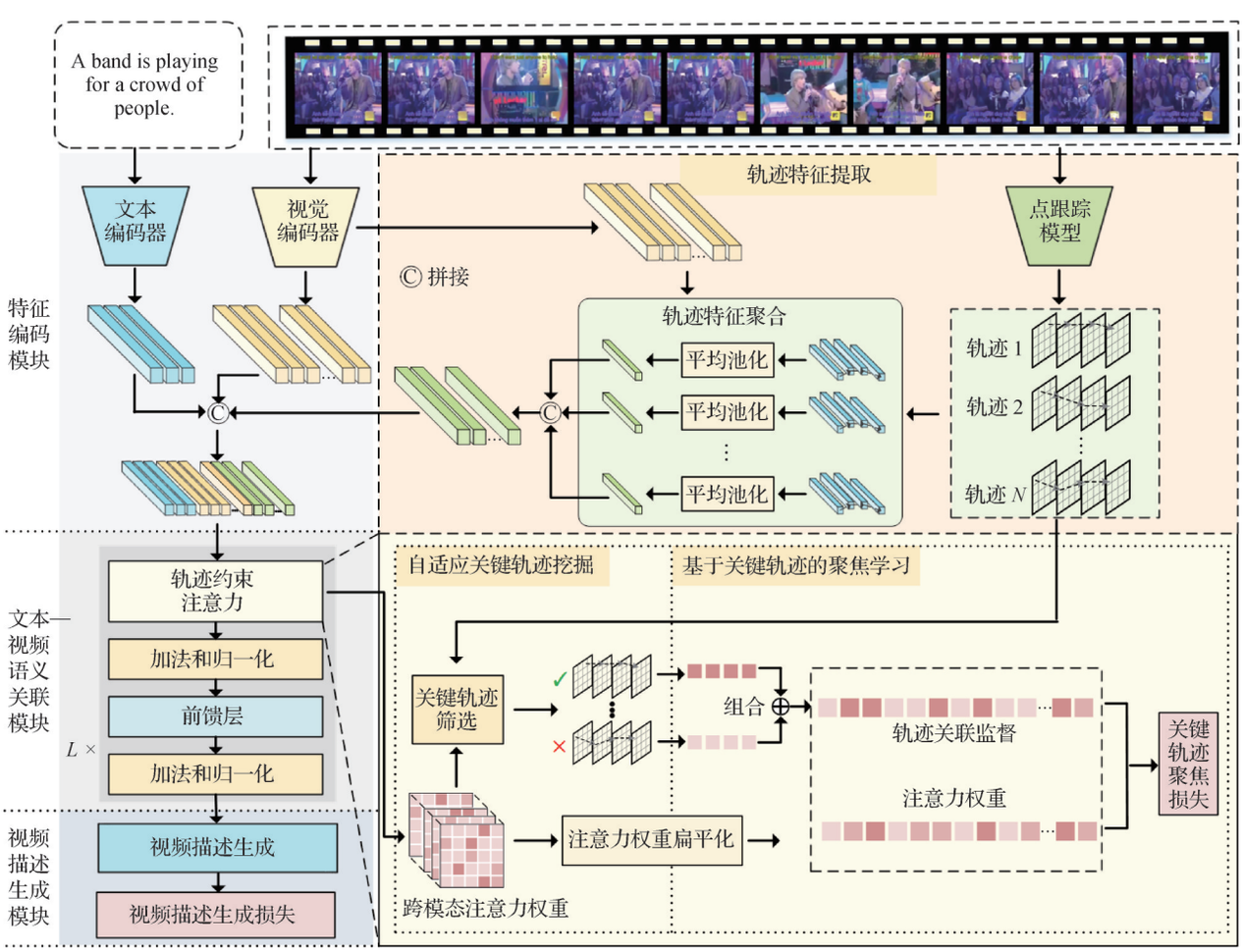


图 3　融合轨迹时空感知与自适应语义聚焦的视频描述方法总体框架

Fig. 3　Overall framework of video captioning with trajectory-based spatiotemporal perceiving and adaptive semantic focusing

聚焦学习两个模块组成，接下来对这两个模块进行全面的描述。

1.2.1 自适应关键轨迹挖掘

本文设计的自适应关键轨迹挖掘方法的过程如图4所示。

1)跨模态注意力权重。本文使用多模态融合模块将来自不同模态的信息进行整合，包括真实标注文本特征、视觉特征和轨迹特征。形式上，在时间步t，将先前预测的文本特征$\boldsymbol{Y}$、视觉特征$\boldsymbol{V}$和轨迹特征$\boldsymbol{R}$进行拼接作为多模态输入的多源特征，记为$\boldsymbol{U}=[\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_S]$，其中$S$为多模态特征总数。随后，将$\boldsymbol{U}\in\mathbf{R}^{S\times C}$输入到文本—视频语义关联模块。本文将预测词向量初始编码为$\boldsymbol{y}_t\in\mathbf{R}^{1\times C}$，MHA通过以下方式融合来自不同输入模态的信息，具体为

$$\boldsymbol{\alpha}_t=\mathrm{softmax}((\boldsymbol{y}_t\boldsymbol{W}_q)(\boldsymbol{U}\boldsymbol{W}_k)^{\mathrm{T}}/\sqrt{d_k}) \tag{4}$$

$$\boldsymbol{\alpha}_t=\left[a_t^1,a_t^2,\cdots,a_t^S\right] \tag{5}$$

$$\boldsymbol{q}_t=\sum_{i=1}^{S}a_t^i\cdot\boldsymbol{u}_i \tag{6}$$

式中，$\boldsymbol{W}_q\in\mathbf{R}^{C\times 2C}$和$\boldsymbol{W}_k\in\mathbf{R}^{C\times 2C}$为嵌入矩阵，$d_k$为特征的维数，$\mathrm{softmax}(\cdot)$表示归一化操作，$\boldsymbol{\alpha}_t\in\mathbf{R}^{1\times S}$表示$\boldsymbol{U}$在融合过程中注意力权重向量。式(4)得到的$\boldsymbol{\alpha}_t$中各标量具体在式(5)中表示，$a_t^i$表示$\boldsymbol{\alpha}_t$中的第$i$个注意力权重值。$\boldsymbol{q}_t$为预测词向量$\boldsymbol{y}_t$经过MHA后对应的输出特征，并通过$\boldsymbol{\alpha}_t$对输入特征进行重新加权后动态聚合得到。在这个过程中，模型学习到的注意力分配权重$\boldsymbol{\alpha}_t$直接影响$\boldsymbol{q}_t$。

由式(4)可知，注意力权重$\boldsymbol{\alpha}_t$在训练中作为隐变量学习。然而，注意力机制常存在“关注偏移”问题，难以合理分配权重，影响视频内容的准确描述。为增强注意力权重的可靠性，本文引入文本—视频语义关联监督，利用轨迹片段的时空一致性，从中挖掘对文本语义贡献较大的关键轨迹，引导注意力权重聚焦于这些关键区域。具体地，视每条轨迹为语义关联单元，结合跨模态注意力权重划分，在多阶段训练中自适应挖掘关键轨迹，并指导模型在生成文本时对其分配合适的特征权重。

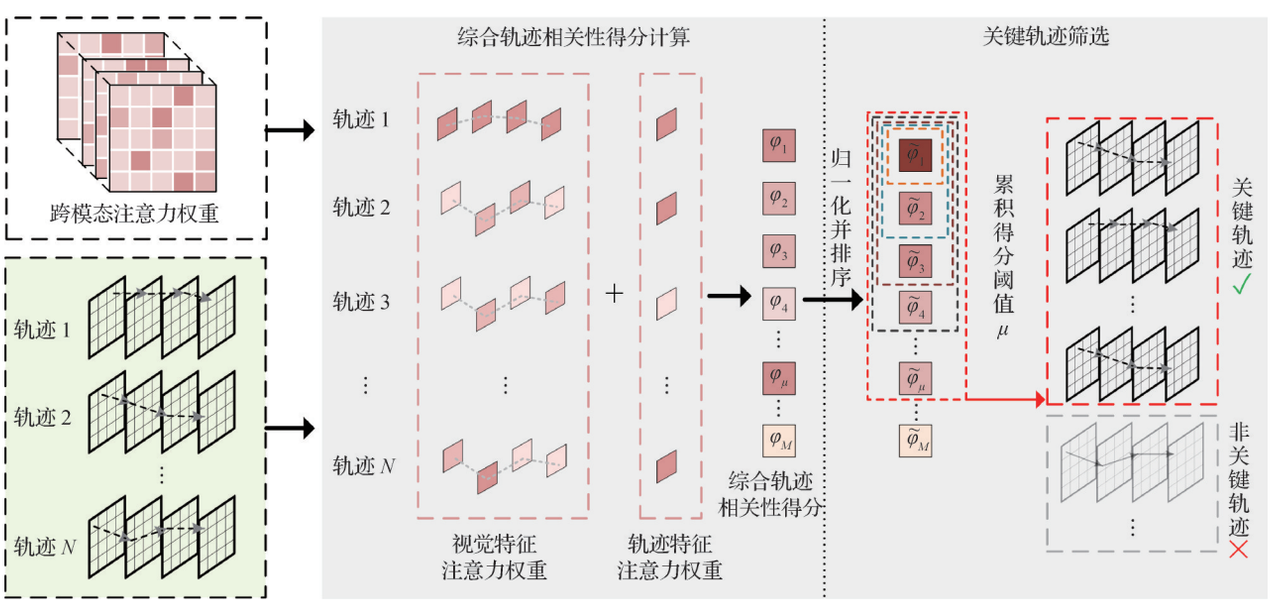


图4 自适应关键轨迹挖掘模块

Fig. 4 Adaptive key trajectory mining module

2)综合轨迹相关性得分计算。为了衡量每条时空轨迹片段在预测当前单词时的重要性，本文设计了计算轨迹片段与文本描述之间的相关性的得分函数。利用文本特征与多源视频特征间的跨模态注意力权重来度量它们之间的语义匹配度。具体地，多源视频特征与文本特征$\boldsymbol{y}_i\in\mathbf{R}^{1\times C}$之间的相关性得分过程可表示为

$$\boldsymbol{\beta}_t=\left[b_1,\cdots,b_G,c_1,\cdots,c_M\right]=softmax(\boldsymbol{\omega}_i) \tag{7}$$

式中，$\boldsymbol{\omega}_i=\left[a_i^t,a_i^{t+1},\cdots,a_i^S\right]$，$\boldsymbol{\beta}_i\in\mathbf{R}^{1\times(G+M)}$表示文本特征$\boldsymbol{y}_i$与视频所有视觉特征$\boldsymbol{V}$和所有轨迹特征$\boldsymbol{R}$的注意力权重归一化后的得分。具体地，$\boldsymbol{\beta}_i$可分为两部分，$[b_1,\cdots,b_G]$表示文本特征$\boldsymbol{y}_i$与所有视觉特征$\boldsymbol{V}$

的相关性，反映其与每条轨迹片段的局部关联；$[c_1, \cdots, c_M]$表示$\boldsymbol{y}_i$与轨迹特征$\boldsymbol{R}$的相关性，轨迹特征$\boldsymbol{R}$与每条轨迹一一对应，故该部分反映的是$\boldsymbol{y}_i$与每条时空轨迹的全局关联。

进一步将计算时空轨迹片段集合中每条轨迹片段与文本特征$\boldsymbol{y}_i$的轨迹相关性得分。$\boldsymbol{\beta}_i$中包含了两者之间的局部相关性和全局相关性，为更好地度量轨迹相关性得分，本文联合局部相关性和全局相关性得到综合轨迹相关性得分。为此，设计了综合轨迹相关性得分函数，具体表示为

$$\varphi_j = F_{\text{tr}}(\boldsymbol{\beta}_i, \boldsymbol{T}_j) = \sum_{k \in \boldsymbol{T}_j} b_k + c_j \tag{8}$$

3）关键轨迹筛选。对所有轨迹的综合轨迹相关性得分进行总体归一化并降序排列，得到排序后的列表中第j条轨迹对应的相关性得分，表示为$\tilde{\varphi}_j$，对应的时空轨迹片段为$\tilde{\boldsymbol{T}}_j$。本文定义轨迹累计集合$\boldsymbol{L}_j$为根据轨迹片段相关性得分排序后前j条轨迹的组合，即

$$\boldsymbol{L}_j = \left\{ \tilde{\boldsymbol{T}}_m, m \in [1, j], m \in \mathbf{N}^+ \right\} \tag{9}$$

进一步计算排序后轨迹片段相关性累积得分，过程表示为

$$\boldsymbol{X} = \left[x_1, x_2, \cdots, x_M \right], x_M = \sum_{\tilde{T}_n \in \boldsymbol{L}_M} \tilde{\varphi}_n \tag{10}$$

式中，x_M表示轨迹累计集合$\boldsymbol{L}_M$的累计得分。显然，x_j会随着j的增大从0到1，形成累计得分序列$\boldsymbol{X}$。

接着，本文提出基于轨迹片段相关性得分对时空轨迹进行策略筛选，以保留对文本语义贡献较大的关键轨迹。同时，筛选过程中自适应调整保留轨迹数量，使得在不同场景下都能灵活筛选关键信息，确保模型能够关注最具语义代表性的轨迹。通过设定累积得分阈值μ，以筛选保留语义贡献较大的关键轨迹，具体过程为

$$\sum_{\tilde{T}_n \in \boldsymbol{L}_\mu} \tilde{\varphi}_n = \text{argmax}\, \boldsymbol{X} < \mu \tag{11}$$

式中，μ为设置的超参数累积得分阈值。当满足式（11）时，轨迹累计集合$\boldsymbol{L}_\mu$为关键轨迹集合，在$\boldsymbol{L}_\mu$外的其余轨迹被视为包含语义信息弱的非关键轨迹。而后根据关键轨迹集合中包含的特征区域，动态调整注意力机制关注的重点，从而实现视频内容与文本描述的精确语义关联。

1.2.2 基于关键轨迹的聚焦学习

跨模态注意力机制缺乏显式监督，易导致注意力偏离语义相关区域，影响文本—视频语义对齐。为此，本文提出基于关键轨迹的聚焦学习方法，引导注意力权重$\boldsymbol{\alpha}_t$聚焦于语义关键区域，抑制非关键轨迹的干扰，提升跨模态语义关联的准确性。

通过构造关键与非关键轨迹的注意力分布约束，鼓励注意力在关键轨迹上的权重总和更大，非关键轨迹上尽可能接近0。同时，引入聚焦损失机制增强对“难样本”的学习，进一步优化语义关联效果。该损失的具体形式为

$$L_{\text{vg}} = -\log\left(\sum_{j \in \boldsymbol{L}_\mu} \varphi_j \right) - \sum_{i \in \boldsymbol{L}_\mu} \varphi_i^\gamma \cdot \log(1 - \varphi_i) \tag{12}$$

式中，$\sum_{j \in \boldsymbol{L}_\mu} \varphi_j$代表关键轨迹区域的注意力权重之和。由式（12）可知，关键轨迹区域权重越接近1，损失越小，引导注意力聚焦于重要的时空区域；非关键轨迹权重为0时损失最小，从而抑制无关注意力分布。φ_i^γ控制不同“难样本”的学习强度，γ设置为0.5。

与传统Focal Loss（Lin等，2017）相比，本文提出的L_{vg}损失在设计目标和数学形式上均有显著创新。Focal Loss通过$(1 - p_t)^\gamma$调制单模态分类任务中的样本权重，主要解决类别不平衡问题；而L_{vg}则通过语义引导的集合约束机制$\sum_{j \in \boldsymbol{L}_\mu} \varphi_j$和非关键轨迹抑制项$\varphi_i^\gamma$，实现跨模态注意力的定向优化。这种设计使模型能够更精准地关注视频中与文本描述相关的关键轨迹，提升跨模态语义对齐能力。

1.3 视频描述生成模块

多源特征经过文本—视频语义关联模块的引导后，最终得到综合视频特征$\boldsymbol{q}_t$作为文本—视频语义的高层表示。随后，通过全连接层和softmax(·)操作计算出词汇表上的概率分布$\boldsymbol{p}_t$，具体表示为

$$y_t^* \sim \boldsymbol{p}_t = softmax(FC(\boldsymbol{q}_t)) \tag{13}$$

式中，$FC(\cdot)$为全连接层，$\boldsymbol{p}_t \in \mathbf{R}^{1 \times M}$表示输出词语在词汇表上的概率分布，$M$表示词汇表长度，生成的词$y_t^*$是$\boldsymbol{p}_t$在词汇表上的预测概率最高词。

1.4 模型训练目标

在文本生成过程中，本文使用了一个因果自注意力掩码，使得每个单词在预测时只能关注已经预测的前置单词。对于视频描述生成任务，本文将输出句子与真实标注文本进行比较，得到对应的模型训练损失。具体而言，本文采用交叉熵损失，定

义为

$$\mathcal{L}_{\text{cap}} = -\sum_{t=1}^{T}\log\left(\boldsymbol{p}_t\left(y_t \middle| y^{1:t-1}\right)\right) \tag{14}$$

另外，本文致力于在模型中联合优化视频描述生成任务和文本—视频语义关联任务，以提升视频描述的生成质量。模型联合目标损失定义为

$$\mathcal{L} = \mathcal{L}_{\text{cap}} + \lambda \cdot \mathcal{L}_{\text{vg}} \tag{15}$$

式中，$\mathcal{L}_{\text{cap}}$表示视频描述生成损失；$\mathcal{L}_{\text{vg}}$对应于式(12)中定义的语义关联损失，作为模型语义关联的正则项，有效地指导注意力机制的权重分布；λ作为超参数，平衡两种类型的损失。

2 实验结果与分析

2.1 数据集与评估指标

2.1.1 数据集

本文方法在MSR-VTT(Xu等，2016)和MSVD(Chen和Dolan，2011)数据集上进行了充分的实验。MSR-VTT是一个大规模数据集，由来自YouTube等在线平台的10 000多个视频片段组成。数据集包括6 573个训练样本、497个验证样本和2 990个测试样本。MSVD是一个短视频片段数据集，每个片段均配有单条人工生成的文本描述。尽管MSVD规模小于MSR-VTT，但内容丰富，具有较高的多样性。数据集划分为1 200个训练样本、100个验证样本和670个测试样本。

2.1.2 评估指标

按照标准的视频描述评估协议，本文使用4种常用的评估指标衡量描述质量，包括基于精确度的BLEU(bilingual evaluation understudy)(Papineni等，2002)、计算句子级相似度得分的METEOR(metric for evaluation of translation with explicit ordering)(Denkowski和Lavie，2014)、基于共识的CIDEr(Vedantam等，2015)，以及通过最长公共子序列估计句子相似度的ROUGE-L(recall-oriented understudy for gisting evaluation — longest common subsequence)(Lin，2004)。上述指标分别记为$B@N$、M、C和R，其中N的取值为1和4。

2.1.3 实施细节

参考先进模型CVG(Jiang等，2024)的方法，本文使用在ImageNet上预训练的Video Swin Transformer提取视觉特征，此外，使用在PointOdyssey数据集上预训练的PIPs++模型(Zheng等，2023)实现视频的点跟踪。对真实标注文本仅进行必要的预处理，包括将文本转换为小写、去除标点符号，在每个描述的开头和结尾分别添加[BOS]和[EOS]标记，对于词汇表中未收录的文本，以[UNK]标记替换。所有描述统一设定句子长度为50，对于超出长度的描述直接截断，不足的则在末尾填充至指定长度。本文采用与CVG(Jiang等，2024)模型相同的超参数，将视觉特征和文本特征的嵌入维度分别设置为512、和768，多模态Transformer的层数L设置为12。在训练过程中，模型使用Adam优化器进行15个epoch的优化，初始学习率设置为3×10^{-5}，每3个epoch衰减0.8倍。

2.2 与最新技术的比较

2.2.1 MSR-VTT上的实验结果

表1总结了本文方法和15种先进方法的性能。其中，RSFD(refined semantic enhancement towards frequency diffusion for video captioning)(Zhong等，2023)、Cocap(accurate and fast compressed video captioning)(Shen等，2023)和IcoCap(improving video captioning by compounding images)(Liang等，2024)使用2D卷积神经网络提取视觉特征；LSRT(long short-term relation Transformer for video captioning)(Li等，2022)、MELTR(meta loss transformer for learning to fine-tune video foundation models)(Ko等，2023)、HMN(learning hierarchical modular networks for video captioning)(Li等，2024)和VCRN(visual commonsense-aware representation network for video captioning)(Zeng等，2025)使用3D卷积神经网络提取视觉特征；SwinBERT(Lin等，2022)、VL-Prompt(Yan等，2023)、OmniViD(generative framework for universal video understanding)(Wang等，2024)、MAN(memory-based augmentation network for video captioning)(Jing等，2024)和HSRA(hierarchical semantic representation and aggregation)(Han等，2025)使用视觉Transformer提取视觉特征；CMG(cross-modal graph with meta concepts for video captioning)(Wang等，2022)、CVG(Jiang等，2024)和Track4Cap(tracking-guided information augmentation for captioning)(Luo等，2025)使用多种视觉编码器提取混合视觉特征。

所提方法首先在MSR-VTT数据集上将其与当

前最新的方法进行比较。从表1的左侧可以看出，该方法在主要评价指标上均优于其他最新方法。具体地，该方法在 CIDEr 指标上达到61.2，相比 IcoCap（60.2）和 OmniVID（56.6）等最新方法有明显提升，同时，在 METEOR 指标上达到32.0。所引入的聚焦损失有效增强了视频中细粒度时空特征与词汇之间的语义关联。

此外，本文方法也显著超越了利用大规模视觉和语言预训练的 MELTR、VL-Prompt 和 OmniViD（CIDEr分数提升+4.6~+8.4）。这表明所提出的聚焦损失在视频描述生成任务中发挥了积极作用。这些结果充分验证了本文方法的优越性。

2.2.2　MSVD上的实验结果

进一步在具有挑战性的MSVD数据集上进行了性能评估。如表1右侧所示，该方法在所有评估指标上均取得了最优结果。其中，BLEU@1达到88.6，BLEU@4达到66.0，METEOR达到42.5，ROUGE-L达到80.1，CIDEr更是达到130.1的高分，明显优于对比方法。尤其在CIDEr指标上，该方法相比OmniViD提升了7.6，相比VL-Prompt提升幅度达到2.0，充分展现了模型的有效性。这种性能的提升，得益于提出的聚焦损失机制，在编码过程中充分挖掘了视频中的时序结构和运动线索，强化了视频局部动态特征与全局语义之间的语义关联。

表1　MSR-VTT和MSVD测试集上与最新方法的结果对比

Table 1　Comparison with state-of-the-art methods on the MSR-VTT and MSVD test sets

方法	年份	MSR-VTT					MSVD				
		B@1 ↑	*B*@4 ↑	*M* ↑	*R* ↑	*C* ↑	*B*@1 ↑	*B*@4 ↑	*M* ↑	*R* ↑	*C* ↑
SwinBERT	2022	83.1	41.9	29.9	62.1	53.3	–	58.2	41.3	77.5	120.6
LSRT	2022	–	42.6	28.3	61.0	49.5	–	55.6	37.1	73.5	98.5
CMG	2023	83.5	44.9	29.6	62.9	53.0	–	59.5	38.8	76.2	107.3
Cocap	2023	–	44.4	30.3	63.4	57.2	–	60.1	41.4	78.2	121.5
MELTR	2023	–	44.2	29.3	62.4	52.8	–	–	–	–	–
VL-Prompt	2023	–	43.2	30.1	62.7	55.3	–	63.5	41.6	78.9	128.1
RSFD	2023	–	43.4	29.3	62.3	53.1	–	51.2	35.7	72.9	96.7
HMN	2024	–	45.1	28.8	63.6	54.2	–	59.9	36.2	74.2	104.7
IcoCap	2024	–	47.0	31.1	64.9	60.2	–	59.1	39.5	76.5	110.3
MAN	2024	–	41.3	28.0	61.4	49.8	–	59.7	37.3	74.3	101.5
OmniVID	2024	–	44.3	29.9	62.7	56.6	–	59.7	42.2	78.1	122.5
CVG	2024	84.8	46.5	31.2	64.6	60.2		–	–	–	–
HSRA	2025	–	46.9	30.9	64.8	55.3	–	62.2	39.2	78.4	110.1
Track4Cap	2025	–	44.6	30.5	63.6	57.7	–	62.1	**42.5**	79.8	127.2
VCRN	2025	–	41.5	28.1	61.2	50.2	–	59.1	37.4	74.6	100.8
本文	2025	**86.0**	**47.5**	**32.0**	**65.1**	**61.2**	**88.6**	**66.0**	**42.5**	**80.1**	**130.1**

注：加粗字体表示各列最优结果。“↑”表示数值越高越好，“–”表示没有可用的结果。

2.3　消融实验

为了验证在该模型中添加轨迹特征和引入聚焦损失对视频描述任务的有效性，同时验证部分模型参数对性能的影响，在MSR-VTT验证集上进行了消融研究。具体地，从MSR-VTT数据集中的视频中均匀采样8帧进行消融实验。

2.3.1　轨迹特征的有效性

为了验证提出的轨迹特征的有效性，以每个视频均匀采样8帧的基线模型作为起点，并在此基础上加入轨迹特征模块。具体而言，本文通过视频点轨迹估计模型获取时间轨迹片段，并进一步生成时空轨迹片段，随后将其轨迹视觉向量与其他特征拼

接后输入多模态融合模块。

实验结果如表2所示，与基线模型相比，添加轨迹特征后，模型在所有评估指标上均有所提升，表明该模块对生成文本的质量具有显著的正向作用。其中，ROUGE-L(+1.2)和CIDEr(+0.8)的提升尤为明显，说明模型生成的文本在整体结构和内容匹配度方面得到了优化。同时，CIDEr指标的显著提升进一步表明，该模块能够帮助模型更准确地提取和表达视频中的关键信息，从而增强描述文本的连贯性和内容相关性。

表2　在MSR-VTT验证集上验证不同模块的有效性

Table 2　Evaluate the effectiveness of different modules on the MSR-VTT validation set

模块		评估指标				
轨迹特征	聚焦损失	*B*@1	*B*@4	*M*	*R*	*C*
–	–	84.5	41.6	30.0	61.8	55.3
√	–	85.0	42.1	**30.3**	63.0	56.1
–	√	85.0	42.8	30.2	**63.2**	56.7
√	√	**85.5**	**43.0**	**30.3**	**63.2**	**57.2**

注：加粗字体表示各列最优结果。“–”表示不包含对应模块，“√”表示包含对应模块。

2.3.2　聚焦损失的有效性

为评估聚焦损失模块的效果，本文在基线模型中单独引入该模块，通过关键轨迹构建跨模态注意力损失，引导模型自适应关注语义关键区域。结果如表2显示，各项指标均有提升(BLEU@4(+1.2)、ROUGE-L(+1.4)、CIDEr(+1.4))表明该模块有效提升文本的连贯性与内容匹配度。与主要提升视频表征能力的轨迹特征模块不同，聚焦损失通过优化跨帧注意力分配，提升语义一致性和关键内容表达质量，CIDEr的显著提升尤为突出。

此外，在已有轨迹特征模块基础上引入聚焦损失，BLEU@4和CIDEr进一步提升(+0.9和+1.1)，表明两者具备良好互补性，轨迹特征强化视频表征，聚焦损失优化注意力机制，结合使用可显著提升文本生成质量和视频语义对齐能力。

2.3.3　累积得分阈值μ的影响

为进一步验证累积得分阈值μ的不同取值对模型性能的影响，将累积得分阈值μ在不同设置下，模型的CIDEr评分以折线图的形式进行展示。实验结果如图5所示，从实验结果来看，轨迹累积得分阈值μ对模型的CIDEr评分具有显著影响，并且μ = 0.7是模型的CIDEr评分的最优值。

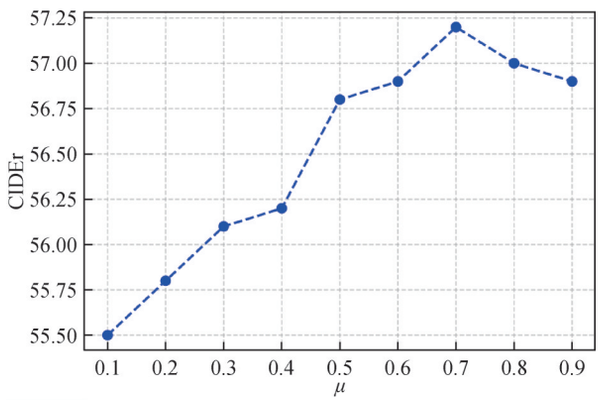


图5　累积得分阈值μ对模型性能的影响

Fig. 5　Effect of cumulative score threshold μ on model performance

具体而言，在μ由0.1增加到0.7的过程中，模型的CIDEr评分整体呈上升趋势，从55.5增长至57.2，表明随着轨迹过滤阈值的提高，模型能够有效筛选出更具代表性的轨迹信息，减少无关的轨迹干扰，从而提升模型对关键物体身上的运动和事件建模能力。然而，当μ继续增大至0.8和0.9时，CIDEr评分开始下降，分别回落至57.0和56.9，这说明过高的轨迹过滤阈值可能放大关键轨迹范围，引入冗余或噪声信息，分散注意力，削弱模型对关键信息的聚焦，从而影响描述准确性。

2.3.4　聚焦损失系数λ的影响

本文还研究了重要的超参数对视频描述的影响。这里λ平衡了视频语义对齐和视频描述的重要性。从表3中可以看出，随着λ的变化，与基线模型相比，视频内容描述模型取得了一致更好的性能。在λ = 0.1至λ = 0.2区间小幅提升，由57.0增加至57.2，随后随着λ增大略有下降，最低降至56.2。当λ = 0.2时，描述性能达到峰值，适度的语义关联监督有助于提升模型在细粒度描述匹配方面的表现，而较大的λ可能会过分强调视频接地的贡献，抑制了语言生成的表达能力。在接下来的实验中，如果没有特别说明，本文将λ设置为0.2。

2.3.5　视频采样帧数的影响

为了探究视频采样帧数对模型性能的具体影响，在MSR-VTT验证集上设置了8、16和32这3种

表 3　在 MSR-VTT 验证集上验证聚焦损失系数 λ 的影响

Table 3　Evaluation of the impact of the focus loss coefficient λ on the MSR-VTT validation set

损失系数 λ	评估指标				
	B@1	*B*@4	*M*	*R*	*C*
0.1	85.1	42.5	30.2	62.8	57.0
0.2	**85.5**	**43.0**	**30.3**	**63.2**	**57.2**
0.3	85.1	42.6	30.1	63.1	56.9
0.4	85.0	42.3	30.1	62.5	56.3
0.5	84.9	42.2	30.1	62.4	56.2

注：加粗字体表示各列最优结果。

表 4　在 MSR-VTT 验证集上验证视频采样帧数的影响

Table 4　Impact of the number of sampled video frames on the MSR-VTT validation set

帧数	评估指标				
	B@1	*B*@4	*M*	*R*	*C*
8	85.5	43.0	30.3	63.2	57.2
16	85.8	43.2	30.4	63.1	58.1
32	86.1	43.8	30.7	63.3	59.7
100	**86.4**	**45.1**	**31.0**	**63.4**	**60.9**

注：加粗字体表示各列最优结果。

不同采样帧数进行对比实验，实验结果如表 4 所示。从结果可以明显看出，随着采样帧数的增加，模型在各项评估指标上均呈现出持续上升的趋势。具体而言，CIDEr 指标从帧数为 8 时的 57.2 提升至帧数为 32 时的 59.7，增长幅度较为显著，表明更多帧信息有助于增强模型对视频时序结构和关键动作的理解能力。同时，BLEU@1、BLEU@4、METEOR 和 ROUGE-L 指标也随帧数增加而稳步提升，其中 BLEU@4 从 43.0 提升至 43.8，METEOR 从 30.3 增长至 30.7。在长视频场景下（100 帧），模型仍能保持较好的性能表现（BLEU@4 指标达到 45.1，CIDEr 指标为 60.9），这表明本文所提的轨迹特征提取与聚合方法在长视频上依然有效。同时，帧数为 100 时性能达到最佳，进一步反映出更高帧数能够为生成描述提供更丰富的细节支撑。适当提高视频采样帧数能够有效缓解关键信息缺失的问题，提升模型对视频多模态特征的建模能力，提升内容描述质量。

2.3.6　聚焦损失应用层数的影响

本文还研究了聚焦损失在跨模态注意力模块的不同位置引入聚焦损失对模型性能的影响。从表 5 可以看出，当仅在第 12 层引入损失时，各项指标整体表现最佳，CIDEr 指标达到 57.2，优于在更多层引入聚焦损失的情况。同时，BLEU@4 和 METEOR 等指标也有小幅提升。结果表明，聚焦损失在模型的高层应用效果更佳，主要原因在于高层语义信息更为丰富，此时施加轨迹引导可以更有效地对齐视觉与文本模态的关键区域。而过早在较低层引入该损失，可能会限制模型对底层视觉特征的自由表达，导致最终生成描述的整体质量略有下降。

2.3.7　轨迹聚合方法的影响

为验证不同聚合方式对模型性能的影响，在 MSR-VTT 验证集上设计了对比实验，选取最大池化、平均池化与自适应池化 3 种聚合方式，实验结果如表 6 所示。实验结果显示，本文采用的平均池化在各项指标上均取得了最优表现，这得益于其能对轨迹序列中的特征进行全面整合，平衡不同帧特征的贡献，避免单一特征过度主导聚合结果，从而更完整地保留轨迹的时空语义信息。相比之下，最大池化性能稍逊，主要因为其仅筛选轨迹中响应最强的特征，容易忽略那些强度较低但对语义表达至关重

表 5　聚焦损失作用于多模态 Transformer 不同层数在 MSR-VTT 验证集的影响

Table 5　Impact of applying the focusing loss at different layers of the multimodal Transformer on the MSR-VTT validation set

模块			评估指标				
第 10 层	第 11 层	第 12 层	*B*@1	*B*@4	*M*	*R*	*C*
√	√	√	85.2	42.5	30.2	62.8	56.8
–	√	√	85.3	42.3	30.1	63.1	57
–	–	√	**85.5**	**43.0**	**30.3**	**63.2**	**57.2**

注：加粗字体表示各列最优结果。“√”表示采用，“–”表示未采用。

要的细节信息。此外，自适应池化的权重更新依赖于特征自身的梯度反馈，在轨迹聚合场景中，这种机制容易受到噪声特征的干扰。综合来看，轨迹聚合方式的选择对模型性能有着直接影响，合理的聚合策略能够有效增强模型对视频轨迹特征的利用，进而提升视频与文本语义关联的准确性。

表6　在MSR-VTT验证集上验证轨迹聚合方法的影响
Table 6　Effects of verification trajectory aggregation methods on MSR-VTT validation set

聚合方式	评估指标				
	B@1	*B*@4	*M*	*R*	*C*
最大池化	84.1	42.3	30.0	62.4	56.1
平均池化	**85.5**	**43.0**	30.3	**63.2**	**57.2**
自适应池化	84.6	42.5	**30.4**	62.7	55.9

注：加粗字体表示各列最优结果。

2.3.8　不同点跟踪算法的影响

为验证不同点跟踪算法提取的视频点轨迹对提出视频描述方法性能的影响，在MSR-VTT验证集上设计了对比实验，选取PIPs++（Zheng等，2023）、CoTracker3（Ko等，2023）和TAP-Vid（Doersch等，2022）3种主流点跟踪算法，实验结果如表7所示。实验结果表明，三者的评估指标差异较小，整体性能接近：PIPs++在部分指标上略优，CoTracker3和TAP-Vid则在其他指标上各有侧重，但差距均未超过0.5%。这3个先进的点跟踪算法在复杂场景下的轨迹生成能力已较为成熟，生成的轨迹在完整性和准确性上达到近似水平。因此，在本模型中，不同先进点跟踪算法对后续注意力权重分配的干扰程度相近，最终体现为实验性能的小幅波动。

2.3.9　模型复杂度分析

在表8中，从模型大小（参数量）、推理计算成本（浮点运算量）和推理速度（单视频推理时间）3个维度，对模型复杂度进行了全面比较。本文提出的轨迹特征和聚焦损失模块均可作为即插即用模块嵌入任意Transformer架构的视频描述模型，且对基线模型的参数量和浮点运算量影响较小。这一特点的重要原因是，模型中使用的点轨迹均为离线提取，因此对模型复杂度的影响降至最低。特别地，在推理过程中聚焦损失模块不会为基线模型引入任何新参数和浮点运算量。

表7　在MSR-VTT验证集上验证不同点跟踪算法的影响
Table 7　Effects of verifying different point tracking algorithms on MSR-VTT validation set

点跟踪算法	评估指标				
	B@1	*B*@4	*M*	*R*	*C*
PIPs++	**85.5**	43.0	30.3	**63.2**	**57.2**
CoTracker3	85.0	**43.1**	**30.5**	63.0	56.6
TAP-Vid	85.3	42.5	30.4	62.7	56.8

注：加粗字体表示各列最优结果。

表8　各方法的参数量大小、浮点运算量和运算效率
Table 8　Parameters, floating point calculations and operation efficiency of each method

模型配置		指标		
轨迹特征	聚焦损失	参数量/M ↓	浮点运算量/G ↓	运算效率/s ↓
–	–	222.9	89.78	0.214 4
√	–	223.1	89.96	0.304 1
–	√	222.9	89.78	0.214 4
√	√	223.1	89.96	0.304 1

注：“↓”表示数值越低越好。“√”表示采用，“–”表示未采用。

2.4　可视化分析

2.4.1　轨迹特征与关键轨迹挖掘可视化分析

为了验证所提出模型在轨迹特征提取与关键轨迹挖掘方面的有效性，本文设计了可视化分析实验，对模型在实际视频场景中学习到的轨迹级视觉语义进行了直观展示，如图6所示。图中显示的轨迹为模型预测标红单词时挖掘的关键轨迹，不同颜色表示不同的轨迹。模型在处理动态场景时，能够通过点跟踪模块准确捕捉目标对象的运动轨迹，并在时序维度上形成连续的轨迹线段。模型进一步聚合轨迹上各点的视觉特征，生成轨迹级特征。从图中可见，这些轨迹特征有效聚焦于与视频语义相关的动态区域。

此外，图6还展示了模型在预测部分标红词语（目标词）时所选取的关键轨迹。这些轨迹由本文提出的无监督关键轨迹挖掘策略筛选而得，依据为与目标词在跨模态注意力中响应最高的轨迹集合。可见，选出的关键轨迹通常覆盖了与目标词高度相关的视觉区域，如动作主体、交互对象或关键运动部位，验证了模型在视频—文本对齐方面

的有效性与可解释性。整体而言，本文提出的轨迹建模与聚焦机制，显著提升了模型对视频语义结构的理解能力，为高质量的视频描述生成提供了有力支持。

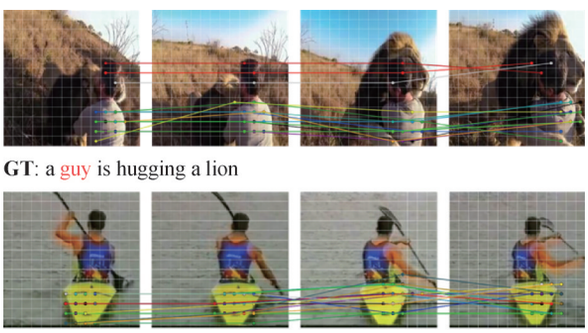


图6　关键轨迹挖掘可视化

Fig. 6　Visualization of key trajectory mining

2.4.2　注意力图可视化分析

为了直观展示本文模型的有效性，可视化对比了所提模型与CVG在生成句子关键词时的时空注意力区域，如图7所示。其中，红色表示高注意力区域，蓝色表示低注意力区域。从图7示列1可以看出，本文模型成功预测出“土豆”，而CVG错误地预测为“小刀”。通过注意力可视化发现，CVG模型关注的区域不够准确，注意力同时分散在土豆和小刀上。相比之下，得益于本文提出的文本—视频语义关联模块的精细对齐能力，本文模型显著改善了视觉注意力分布，能精确地关注视频中的“土豆”区域，不关注周围的其他物体。而CVG由于仅关注“土豆”时关注区域不够准确（过大），同时关注了无关的信息（小刀），进而产生了错误预测。从图7示列2可以看出，本文模型成功预测出“跑道”，这归功于模型在预测时能准确聚焦相关区域。反观CVG则错误预测为“草地”，究其原因在于预测单词时注意力区域出现偏差。从图7示列3可以看出，所提模型同样展现出更精准的注意力聚焦能力。实验结果表明，本文方法能够更精准地聚焦时空注意力区域，从而显著提升模型的预测准确性。

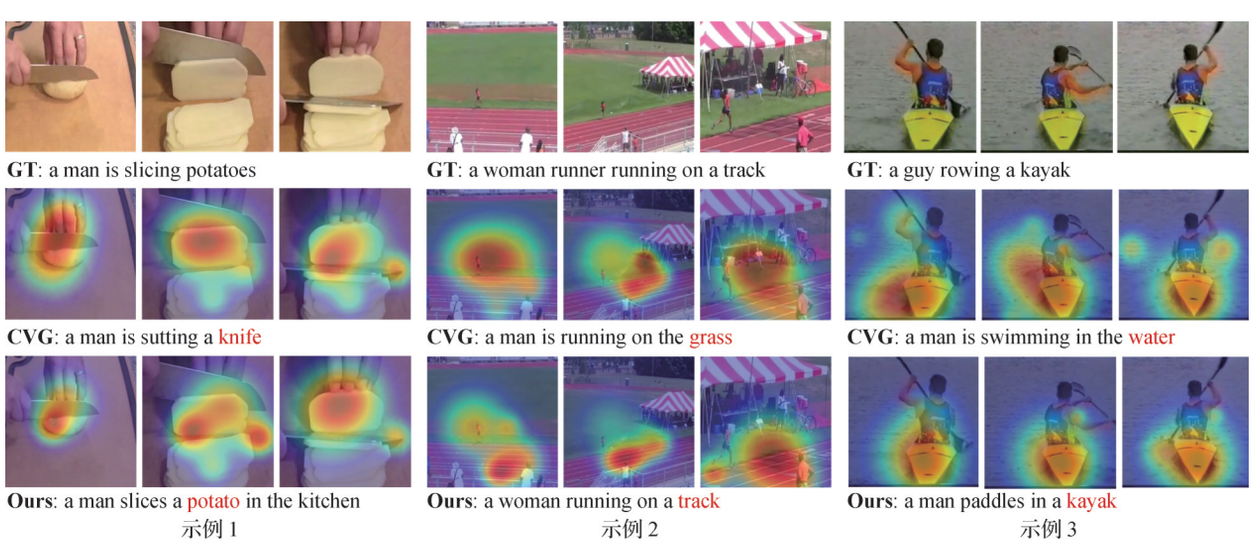


图7　本文方法和CVG注意力图可视化对比

Fig. 7　Visualization comparison of attention maps between proposed method and CVG

2.4.3　视频描述结果可视化分析

图8展示了本文方法与CVG方法在视频描述生成任务上的样例对比，直观呈现了两者在视觉—语言对齐精度与语义理解深度上的差异。其中，正确预测的单词标记为绿色，而错误预测的单词标记为红色。本文方法通过轨迹特征对关键目标的动态追踪与聚焦损失对语义权重的精准调控，能够更稳定地聚焦于视频中的核心区域，从而生成更贴合视频内容的准确描述。

具体来看，在示例1中，视频的核心主体为“男人”，但CVG方法因其注意力机制受到画面中“狮子”这一干扰元素的影响，未能准确锁定核心主体，最终生成了包含错误主体的描述语句；而本文方法通过轨迹特征对“男人”的运动路径进行全程锚定，有效抑制了背景干扰，确保描述中主体识别的准确性。第示例2更凸显了本文方法对动态动作的建模

优势：视频中关键人物的“划桨”动作是描述的核心语义，CVG方法虽然能识别出人物，但对动作的捕捉不够精准，生成的描述未能体现这一关键动态。本文方法借助轨迹特征对人物肢体运动轨迹的细粒度刻画，结合聚焦损失对动作相关语义的强化，不仅准确识别出“划桨”这一动作，还能在描述中清晰呈现动作与主体的关联，进一步提升了描述的语义完整性与细节丰富度。

图8　本文方法和CVG视频描述结果可视化对比

Fig. 8　Visualization comparison of video captioning results between proposed method and CVG

3　结　论

本文围绕视频描述任务中的两个挑战问题，即关键物体动态理解不充分和语义关联不准确，提出一种融合轨迹时空感知与自适应语义聚焦的视频描述方法。具体而言，本文设计了基于视频点轨迹的视觉特征聚合方法，通过轨迹聚合建模增强模型对目标语义的一致性表征，显式建模物体在时空维度上的连续特征，保留了物体的空间外观信息，维持目标特征的语义一致性，有效改善模型在物体运动和形变场景中物体上下文语义建模能力。此外，本文引入了无监督的自适应关键轨迹聚焦学习方法，利用视频密集点轨迹的动态时空信息，引导注意力机制精准聚焦于关键语义区域，同时抑制背景干扰，并且根据解码时注意力权重的变化自适应动态筛选，实现灵活的语义关联。本文方法在MSR-VTT和MSVD两个主流基准数据集上均优于现有对比的先进方法，显著提升了生成描述的准确性与连贯性。

然而，当前关键轨迹的选择依赖于注意力机制，训练前期容易受到早期特征噪声和注意力偏移的影响，限制了训练前期语义聚焦的准确性。未来将引入其他语义辅助信息，提高语义聚焦的准确性和稳定性。

参考文献（References）

Chen D L and Dolan W B. 2011. Collecting highly parallel data for paraphrase evaluation//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics: 190-200 [DOI: 10.5555/2002472.2002497]

Denkowski M and Lavie A. 2014. Meteor universal: language specific translation evaluation for any target language//Proceedings of the 9th Workshop on Statistical Machine Translation. Baltimore, USA: Association for Computational Linguistics: 376-380 [DOI: 10.3115/v1/W14-3348]

Doersch C, Gupta A, Markeeva L, Recasens A, Smaira L, Aytar Y, et al. 2022. TAP-vid: a benchmark for tracking any point in a video//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #989 [DOI: 10.5555/3600270.3601259]

Feichtenhofer C, Pinz A and Zisserman A. 2016. Convolutional two-stream network fusion for video action recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 1933-1941 [DOI: 10.1109/CVPR.2016.213]

Han T T, Xu Y C, Yu J, Yu Z and Zhao S C. 2025. Action-driven semantic representation and aggregation for video captioning. IEEE Transactions on Circuits and Systems for Video Technology, 35(4): 3383-3395 [DOI: 10.1109/TCSVT.2024.3502736]

Jiang W H, Cheng Y B, Liu L X, Fang Y M, Peng Y X and Liu Y. 2024. Comprehensive visual grounding for video description//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press: 2552-2560 [DOI: 10.1609/aaai.v38i3.28032]

Jiang W H, Zhan K, Cheng Y B, Xia X and Fang Y M. 2022. The integrated mechanism of hierarchical decoders and dynamic fusion for image captioning. Journal of Image and Graphics, 27(9): 2775-2787 (姜文晖, 占锟, 程一波, 夏雪, 方玉明. 2022. 结合多层级解码器和动态融合机制的图像描述. 中国图象图形学报, 27(9): 2775-2787) [DOI: 10.11834/jig.211252]

Jiang W H, Zhu M W, Fang Y M, Shi G M, Zhao X W and Liu Y. 2022. Visual cluster grounding for image captioning. IEEE Transactions on Image Processing, 31: 3920-3934 [DOI: 10.1109/TIP.2022.3177318]

Jing S Q, Zhang H N, Zeng P P, Gao L L, Song J K and Shen H T. 2024. Memory-based augmentation network for video captioning. IEEE Transactions on Multimedia, 26: 2367-2379 [DOI: 10.1109/

TMM.2023.3295098]

Ko D, Choi J, Choi H K, On K W, Roh B and Kim H J. 2023. MELTR: meta loss transformer for learning to fine-tune video foundation models//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 20105-20115 [DOI: 10.1109/CVPR52729.2023.01925]

Li G R, Ye H H, Qi Y K, Wang S H, Qing L Y, Huang Q M, et al. 2024. Learning hierarchical modular networks for video captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(2): 1049-1064 [DOI: 10.1109/TPAMI.2023.3327677]

Li L, Gao X Y, Deng J C, Tu Y B, Zha Z J and Huang Q M. 2022. Long short-term relation transformer with global gating for video captioning. IEEE Transactions on Image Processing, 31: 2726-2738 [DOI: 10.1109/TIP.2022.3158546]

Liang Y Z, Zhu L C, Wang X H and Yang Y. 2024. IcoCap: improving video captioning by compounding images. IEEE Transactions on Multimedia, 26: 4389-4400 [DOI: 10.1109/TMM.2023.3322329]

Lin C Y. 2004. ROUGE: a package for automatic evaluation of summaries//Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics: 74-81

Lin K, Li L J, Lin C C, Ahmed F, Gan Z, Liu Z C, et al. 2022. SwinBERT: end-to-end transformers with sparse attention for video captioning//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 17928-17937 [DOI: 10.1109/CVPR52688.2022.01742]

Lin T Y, Goyal P, Girshick R, He K M and Dollár P. 2017. Focal loss for dense object detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2999-3007 [DOI: 10.1109/ICCV.2017.324].

Liu C X, Mao J H, Sha F and Yuille A. 2017. Attention correctness in neural image captioning//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI Press: 4176-4182 [DOI: 10.1609/aaai.v31i1.11197]

Luo H L, Cai X and Shark L K. 2025. Frame-by-frame multi-object tracking-guided video captioning. IEEE Transactions on Circuits and Systems for Video Technology, 35(7): 6357-6370 [DOI: 10.1109/TCSVT.2025.3541965]

Papineni K, Roukos S, Ward T and Zhu W J. 2002. BLEU: a method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics: 311-318 [DOI: 10.3115/1073083.1073135]

Shen Y J, Gu X, Xu K, Fan H, Wen L Y and Zhang L B. 2023. Accurate and fast compressed video captioning//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 15512-15521 [DOI: 10.1109/ICCV51070.2023.01426]

Song H H, Liu Q S, Wang G J, Hang R L and Huang B. 2018. Spatio-temporal satellite image fusion using deep convolutional neural networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(3): 821-829 [DOI: 10.1109/JSTARS.2018.2797894]

Tang M K, Wang Z Y, Zeng Z Y, Li X and Zhou L P. 2023. Stay in grid: improving video captioning via fully grid-level representation. IEEE Transactions on Circuits and Systems for Video Technology, 33(7): 3319-3332 [DOI: 10.1109/TCSVT.2022.3232634]

Ul Amin S, Kim B, Jung Y, Seo S and Park S. 2024. Video anomaly detection utilizing efficient spatiotemporal feature fusion with 3D convolutions and long short-term memory modules. Advanced Intelligent Systems, 6(7): #2300706 [DOI: 10.1002/aisy.202300706]

Vedantam R, Zitnick C L and Parikh D. 2015. CIDEr: consensus-based image description evaluation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 4566-4575 [DOI: 10.1109/CVPR.2015.7299087]

Wang H, Lin G S, Hoi S C H and Miao C Y. 2022. Cross-modal graph with meta concepts for video captioning. IEEE Transactions on Image Processing, 31: 5150-5162 [DOI: 10.1109/TIP.2022.3192709]

Wang J K, Chen D D, Luo C, He B, Yuan L, Wu Z X, et al. 2024. OmniViD: a generative framework for universal video understanding//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 18209-18220 [DOI: 10.1109/CVPR52733.2024.01724]

Xie J Y, Meng Y D, Zhao Y T, Nguyen A, Yang X Y and Zheng Y L. 2024. Dynamic semantic-based spatial graph convolution network for skeleton-based human action recognition//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press: 16702-16710 [DOI: 10.1609/aaai.v38i6.28440]

Xu J, Mei T, Yao T and Rui Y. 2016. MSR-VTT: a large video description dataset for bridging video and language//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 5288-5296 [DOI: 10.1109/CVPR.2016.571]

Yan C G, Tu Y B, Wang X Z, Zhang Y B, Hao X H, Zhang Y D, et al. 2020. STAT: spatial-temporal attention mechanism for video captioning. IEEE Transactions on Multimedia, 22(1): 229-241 [DOI: 10.1109/TMM.2019.2924576]

Yan L Q, Han C, Xu Z L, Liu D F and Wang Q F. 2023. Prompt learns prompt: exploring knowledge-aware generative prompt collaboration for video captioning//Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: IJCAI: 1622-1630 [DOI: 10.24963/ijcai.2023/180]

Yao L, Torabi A, Cho K, Ballas N, Pal C, Larochelle H, et al. 2015. Describing videos by exploiting temporal structure//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 4507-4515 [DOI: 10.1109/ICCV.2015.512]

Zeng P P, Zhang H N, Gao L L, Li X P, Qian J and Shen H T. 2025. Visual commonsense-aware representation network for video cap-

tioning. IEEE Transactions on Neural Networks and Learning Systems, 36(1): 1092-1103 [DOI: 10.1109/TNNLS.2023.3323491]

Zhao Y Q, Jin Z, Zhang F, Zhao H Y, Tao Z W, Dou C F, et al. 2023. Deep-learning-based image captioning: analysis and prospects. Journal of Image and Graphics, 28(9): 2788-2816 (赵永强, 金芝, 张峰, 赵海燕, 陶政为, 豆乘风, 等. 2023. 深度学习图像描述方法分析与展望. 中国图象图形学, 28(9): 2788-2816) [DOI: 10.11834/jig.220660]

Zheng Y, Harley A W, Shen B K, Wetzstein G and Guibas L J. 2023. PointOdyssey: a large-scale synthetic dataset for long-term point tracking//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 19798-19808 [DOI: 10.1109/ICCV51070.2023.01818]

Zhong X, Li Z P, Chen S Q, Jiang K, Chen C and Ye M. 2023. Refined semantic enhancement towards frequency diffusion for video captioning//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press: 3724-3732 [DOI: 10.1609/aaai.v37i3.25484]

Zhou L W, Kalantidis Y, Chen X L, Corso J J and Rohrbach M. 2019. Grounded video description//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 6571-6580 [DOI: 10.1109/CVPR.2019.00674]

作者简介

肖景富，男，博士研究生，主要研究方向为计算机视觉。
E-mail：2202320632@stu.jxufe.edu.cn
姜文晖，通信作者，男，副教授，主要研究方向为计算机视觉、多模态信号处理。E-mail：wenhui@jxufe.edu.cn
方玉明，男，教授，主要研究方向为计算机视觉、多媒体信号处理、视觉质量评估。E-mail：fa0001ng@e.ntu.edu.sg
方承炀，男，讲师，主要研究方向为计算机视觉。
E-mail：fangchengyang@jxufe.edu.cn
赵小伟，男，博士，主要研究方向为计算机视觉。
E-mail：zhaoxw8@sany.com.cn