面向多模态关联不确定性的视频描述方法

姜文晖 官文彬 黎海军 方承炀 方玉明 左一帆
（江西财经大学计算机与人工智能学院 南昌 330013）

摘 要 视频描述技术的核心在于构建输入视频到输出文本的映射关系。然而，受训练数据的标注噪声影响，视频与对应的文本描述之间存在关联的不确定性，具体体现在：1）对于视觉模态，视频中存在文本描述未涉及的视觉片段，且这些视觉片段在视频的时空位置不确定；2）对于文本模态，局部文本错误地描述了视频中未出现的目标和事件，且这些错误在文本中出现的位置不确定。视频与文本描述的关联不确定性阻碍模型学习正确的跨模态语义关系，影响模型生成准确的视频描述。为解决以上挑战问题，本文提出一种多模态关联不确定性感知的视频描述方法。针对视觉模态的不确定性，提出一种文本内容感知的视频特征表达模型。通过帧聚类构建视频的场景原型，并基于场景原型与文本语义的相关性采样视频关键帧，形成与文本语义一致性高、冗余性低的关键帧序列，最终实现语义丰富、内容相关的视频表征。针对文本模态的不确定性，提出短语级质量感知的抗噪训练。联合建模文本标注的模态内语义关联性以及文本与视频内容的跨模态相关性，以无监督评估方式量化文本短语级标注的多维度质量，克服真值质量标签缺失条件下文本短语级质量评价的挑战。通过短语级质量指导视频描述模型的训练，抑制局部标注噪声的影响，充分利用了局部噪声数据的价值。所提方法可应用于大部分视频描述模型，在改善模型性能的同时不会增加测试时的计算量。本文以具有代表性的SwinBERT和VideoLLaMA2为基础模型，在MSR-VTT与MSVD数据集上开展实验。分析结果表明，该方法在多项评价指标上显著高于其他代表性方法。消融实验与定性分析也进一步佐证了本文方法可以有效克服多模态关联的不确定性。

关键词 视频描述；多模态关联；场景原型；短语级质量评估；抗噪训练
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2026.00481

Video Captioning with Multi-Modal Correlation Uncertainty

JIANG Wen-Hui GUAN Wen-Bin LI Hai-Jun FANG Cheng-Yang
FANG Yu-Ming ZUO Yi-Fan
（School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang 330013）

Abstract The core of video captioning lies in establishing a precise and semantically rich mapping from input video sequences to natural language descriptions. This task requires deep understanding of both visual dynamics and linguistic structure, as well as the ability to align them coherently across modalities. However, a major obstacle in training effective video captioning models is the widespread presence of annotation noise in commonly used datasets. Such noise introduces significant uncertainty in the alignment between video content and corresponding textual descriptions, which severely hinders the model′s capacity to learn accurate cross-modal

收稿日期：2025-06-03；在线发布日期：2025-11-03。本课题得到国家重点研发计划（No. 2023YFE0210700）、国家自然科学基金重点项目（No. 62132006）、国家自然科学基金国际（地区）合作与交流项目（No. 62132006）、国家自然科学基金专项项目（No. 62441203）、国家自然科学基金面上项目（No. 62271237）、国家自然科学基金地区项目（No. 62562035）、江西省双千计划（No. jxsq2023101092）、江西省重点研发计划（No. 20252BCE310034）资助。**姜文晖**，博士，副教授，中国计算机学会（CCF）会员，主要研究方向为图像内容理解、跨媒体分析。E-mail：wenhui@jxufe.edu.cn。**官文彬**，硕士研究生，主要研究方向为图像视频内容理解。**黎海军**，硕士研究生，主要研究方向为多模态数据分析。**方承炀**，博士，讲师，主要研究领域为场景文字视觉问答、多模态内容理解。**方玉明**（通信作者），博士，教授，中国计算机学会（CCF）会员，主要研究领域为计算机视觉、多媒体信号处理和视觉质量评估。E-mail：leo.fangyuming@foxmail.com。**左一帆**，博士，副教授，中国计算机学会（CCF）会员，主要研究领域为图像处理和多媒体信号处理。

representations and ultimately limits its performance. This alignment uncertainty manifests in two distinct yet interrelated ways. First, in the visual modality, videos often contain redundant, irrelevant, or background segments, which are not mentioned in the associated text. The spatial and temporal locations of these unannotated visual elements are unpredictable, making it challenging for models to identify which parts of the video are truly salient for description generation. As a result, models may focus on misleading cues, leading to inaccurate or incomplete captions. Second, in the textual modality, human-generated annotations can contain erroneous or hallucinated phrases that refer to objects, actions, or events not actually present in the video. These inaccuracies occur at arbitrary positions within the sentence, disrupting the semantic coherence and weakening the reliability of supervision signals during training. This bidirectional misalignment distorts the learning process and degrades the quality of generated captions. To address these challenges, we propose a novel multimodal alignment uncertainty-aware framework for video captioning, designed to explicitly mitigate both visual and textual uncertainties. Our approach consists of two complementary components. First, to handle visual uncertainty, we introduce a Text-aware Video Representation module. This method starts by clustering semantically similar video frames into scene prototypes, effectively summarizing the video′s dynamic content into a set of representative states. Then, guided by the semantics of the textual description, the model computes cross-modal relevance scores between each prototype and the caption, enabling selective sampling of key frames that are most aligned with the described narrative. The resulting key frame sequence is temporally coherent, semantically focused, and minimally redundant, thereby enhancing the fidelity of the visual representation used for caption generation. Second, for the uncertainty in the textual modality, we introduce a Phrase-Level Quality-aware Noise-Tolerant Training strategy. By jointly modeling the intra-modal semantic consistency within textual annotations and the cross-modal relevance between text and video content, we develop an unsupervised approach to estimate the multidimensional quality of text phrases, overcoming the challenge of quality evaluation without ground-truth labels. This phrase-level quality estimation guides the training of the video captioning model, suppressing the impact of local annotation noise while effectively leveraging the informative value of noisy data. Our framework is model-agnostic and can be seamlessly integrated into various encoder-decoder-based video captioning architectures. Importantly, it introduces no additional computational overhead during inference, ensuring practical efficiency. We conduct extensive experiments on two widely used benchmarks, MSR-VTT and MSVD, using SwinBERT and VideoLLaMA2 as base models. Compared with state-of-the-art methods, our approach achieves consistent and significant improvements across BLEU, METEOR, ROUGE-L, and CIDEr metrics. Ablation studies confirm the effectiveness of each component, and qualitative analysis demonstrates enhanced accuracy, fluency, and factual consistency in generated captions.

Keywords video captioning; multi-modal correlation; scene prototypes; phrase-level quality assessment; noise-tolerant training

1 引 言

视频描述技术是指用完整、流畅的语句描述视频中对应的人物、场景和事件。这一技术是计算机视觉与自然语言处理交叉学科的研究热点，可以应用于多媒体信息检索、无人机巡检报告生成等实际场景，因此具有重要的研究价值和应用需求。

视频描述技术的核心在于构建输入视频到输出文本的关联与映射关系。深度学习技术的迅猛发展

和大规模视频-文本数据集的构建极大推动了视频描述技术的发展。近年来，视频描述技术普遍采用编码器-解码器框架[1-10]实现从视频到文本的映射。其中卷积神经网络（CNN）[1-4]或视觉Transformer[5-6]模型通常作为视频编码器提取视频内容的语义特征，长短时记忆网络（LSTM）[8-10]或其他类似的序列生成模块则作为解码器生成文本描述。大量研究人员通过改进编码器和解码器的结构不断提升视频描述技术的性能，并取得了显著进步。

上述模型结构通常假设训练数据中的视频与文本标注之间存在完美的对应关系。然而，受标注噪声的影响，视频与对应的文本标注之间存在关联的不确定性。一方面，对于视觉模态而言，视频的信息量大，持续时间长，但并非所有内容都与视频的主题密切相关。例如，如图1（a）所示，视频包含多个不同的场景片段，但其核心主题仅与视频的一个片段最相关，其余场景的信息未在文本中描述，即视频的局部内容与文本描述不关联。对于不同视频而言，不相关的片段在视频的时空位置不确定，降低了视频特征表达的区分性。近期，部分研究度量视频帧与文本描述的语义相似度，并采样与文本描述最相关的关键帧构建视频特征[11-12]，但该方案容易产生冗余的关键帧，降低了视频特征表达的语义完备性。另一方面，对于文本模态而言，由于文本标注的随意性，一些文本描述存在局部噪声，无法准确反映视频内容。如图1（b）所示，文本标注的人物描述正确，但事件描述错误，即文本的局部描述与视频无关。文本局部噪声的出现也具有不确定性，会误导模型学习错误的“视频-文本”映射关系，从而影响生成描述的准确性和连贯性。部分研究成果评估文本的全局标注质量，并滤除低质量数据以消除其对模型训练的影响[13]。但局部错误的文本标注依然包含部分正确的文字描述，整体滤除难以充分利用标注数据的全部价值。因此，有效地建模视频与文本之间关联的不确定性，实现噪声抑制与语义保真的协同优化，是视频内容描述技术的关键问题之一。

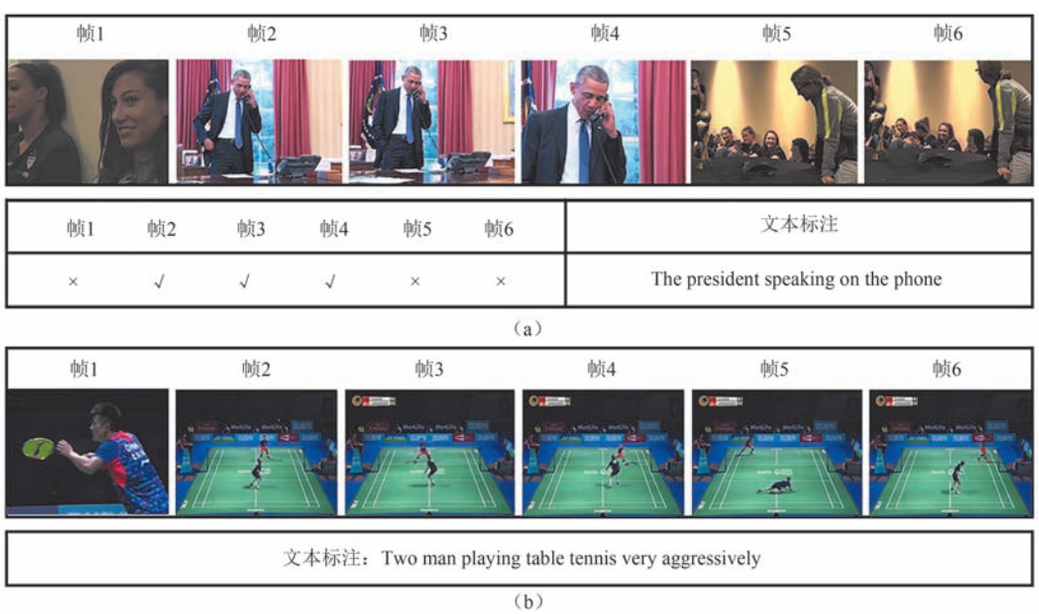

图1　MSR-VTT数据集中不同描述的示例（其中（a）表示文本只描述部分视频帧，（b）表示文本描述的部分错误，符号“√”表示文本描述与视频帧对齐）

针对以上挑战，本文提出一种多模态关联不确定性感知的视频描述方法。该方法针对视频模态的不确定性提出文本内容感知的视频特征表达；针对文本模态的不确定性提出短语级质量感知的抗噪训练。两者共同作用，降低“视频-文本”关联的不确定性对视频内容描述模型的影响。

对于文本内容感知的视频特征表达模块，其核心思想在于采样视频中语义信息重要且冗余性低的关键帧，进而构建与文本描述内容相关且语义完备的视频特征表达模型。为此，首先借助视觉大模型对静态图像特征表达的区分性，结合特征聚类构建视频的场景原型（prototypes），其中不同的场景原型代表视频中发生的不同场景或事件。然后，基于场景原型与文本描述之间的相关性，从每个场景原型中选择与文本语义相关的关键帧，以增强视频与文本之间的语义一致性。由于关键帧来自大量不同的

场景原型，因此冗余性低、语义丰富性高。最后，对采样的关键帧构建时空特征模型，最终生成语义丰富的视频表示。该模块不仅提高了视频内容与文本之间的语义对齐，还降低了关键帧的冗余，提高了视频特征的语义完备性。

对于短语级质量感知的抗噪训练，本文量化文本的短语级标注质量，并以此指导视频描述模型的抗噪训练。由于文本描述的质量评价缺少黄金标准和监督标签，本文提出一种面向无监督的文本标注质量评价方法，其核心思想在于大多数与同一视频对应的文本标注覆盖了视频的显著时空区域，因此文本模态语义一致性高的标注为高质量描述的概率越大。同时，结合提示调优评估文本标注与视频的跨模态语义一致性，形成单词级文本标注的多维度质量评价。最后设计渐进式融合策略形成单词级文本标注的综合质量，以此指导视频描述模型的抗噪训练，提高模型面向噪声标注训练的鲁棒性。

本文所提方法可应用于大部分视频描述模型的结构，如基于Transformer的模型结构和基于大语言模型的结构。本文以具有代表性的SwinBERT和VideoLLaMA2为基础模型，利用文本内容感知的视频特征表达和短语级质量感知的抗噪训练对其改进，并在MSR-VTT[14]和MSVD[15]数据集上进行了全面分析，验证了本文方法的显著优势。特别地，以VideoLLaMA2为基础模型，本文方法在MSR-VTT数据集上取得了CIDEr分数76.8，提高了4.2；在MSVD数据集上取得了CIDEr分数179.1，提高了4.8。

本文的主要贡献总结如下。

1）提出文本内容感知的视频特征表达。该方法通过构建场景原型，以采样与文本标注语义一致性高、冗余性低的关键帧，形成语义完备的视频特征表示，提高视频与文本标注关联的一致性。

2）短语级质量感知的抗噪训练。联合建模文本标注的模态内语义关联性以及文本与视频内容的跨模态相关性，通过无监督评估方式量化单词级文本标注的综合质量，有效解决了在缺少真值质量标签下文本局部噪声的评估问题。以文本标注的短语级质量指导视频描述模型的学习，充分挖掘了噪声标注的价值，提高了模型训练的有效性。

3）对包括多模态大模型在内的多种视频描述模型的结构进行实验，在MSR-VTT[14]和MSVD[15]数据集上展示了优异的性能。

2 相关工作

2.1 视频内容描述方法

视频描述任务旨在通过自然语言对视频内容进行总结。现有的绝大多数方法采用编码-解码架构[1-10]。其中编码器提取视频的语义特征，解码器生成文本描述。常用的视频编码策略包括ResNet[3]、C3D[2]、视觉变换器（Vision Transformer）[6]和对象特征[16]等。长短时记忆网络（LSTM）[8-9]、Transformer[17-18]和大语言模型[19]则广泛应用于解码器。

尽管现有方法已经取得了显著进步，现有特征提取策略的局限性仍是制约描述质量的关键瓶颈。传统的视频编码器采用固定间隔采样或均匀分段策略提取视频关键帧及相应特征[20]。例如，SwinBERT[6]和VALOR[21]模型将均匀采样的关键帧输入Video Swin Transformer提取视频特征。Vid2Seq模型[22]在关键帧编码的基础上额外进行时序编码以提高时空建模的有效性。CVG模型[9]进一步引入经动作识别任务预训练的ViT模型提取视频的运动特征。LSTR[23]、HSRA[24]、HMN[8]等模型基于均匀采样的关键帧提取视频切片，并融合2D、3D和物体级特征进行多维度视频编码。然而，均匀采样策略既容易产生连续的冗余信息，也可能采样到与主题不相关的背景信息，从而影响视频描述的准确性。

为了提高视频关键帧与文本描述的语义一致性，部分研究人员通过大规模多模态预训练提高视频与文本特征的全局表达能力[25-26]。Ryu等学者[12]基于已解码的部分文本描述筛选视频的关键帧，以实现对视频特征的动态更新。Lin等学者[11]则基于视频帧与文本描述的语义一致性过滤相关性低的关键帧，提高视频与文本的语义一致性，但保留的关键帧存在一定的冗余性。Shi等学者[17]则通过额外的检索模块从外部数据中挖掘与视频内容相关的文本数据，并利用这些文本修正视频的特征表达，但该方法受限于外部文本数据的质量。在本研究中，本文的特征表达方法既考虑视频特征与文本标签的语义一致性，又考虑了采样关键帧的冗余性，因此能够有效提升视频描述的准确性与丰富性。

2.2 噪声感知的文本描述方法

从含噪标签中学习是多模态关联学习的重要研究任务。一种直接应对噪声数据的方法是基于视觉-文本数据的CLIP[27]相似性过滤相关性低的训练数据。然而，排除过滤后的数据会减少可学习的信息

量，从而限制了文本描述生成过程中的表达能力。另一种常用策略是调整训练的损失函数。例如，Reed等人[28]结合原始标签和预测标签来计算延迟损失；Wang等人[29]利用模型预测标签的熵，渐进地修正潜在错误训练标签的语义类别。然而这些方法主要是为图像分类任务设计。为了更好地解决多模态关联学习中的噪声数据，Li等人[30]提出动量蒸馏，通过动量模型生成伪目标，以提升图像-文本对比学习中的跨模态对齐效果；Kang等人[13]用预训练的CLIP计算图像-文本的相似性，并以此指导模型学习不同对齐级别的图像-文本关联关系。然而，这些方法都只建模了文本与视觉数据的全局相关性，忽略了文本不同单词与视频内容的相关度存在差异的普遍现象。相比之下，本文综合考虑文本不同单词的质量，为视频-文本的关联学习提供细粒度的指导。

2.3　文本质量评价方法

当前文本数据的质量评估广泛采用基于n-gram的匹配策略，例如BLEU[31]、ROUGE-L[32]、METEOR[33]和CIDEr[34]。另一类研究方法则利用预训练语言模型提取文本的语义特征，并在特征空间中进行文本的语义匹配。例如，BERTScore[35]借助BERT模型强大的文本语义编码能力，量化两段文本的语义相似程度。该方法结合了粗粒度（视频与文本描述）和细粒度（帧与词汇）层次的匹配得分，从而既考虑了视频文本描述的整体理解，又考虑了其具体属性。ViLBERTScore[36]使用预训练的ViLBERT[37]模型比较候选文本描述与参考之间的视觉对齐文本表示。TIGEr[38]评估文本描述质量时，不仅考虑文本描述如何呈现图像内容，还考虑机器生成的文本描述与人工生成的文本描述之间的匹配度。然而，这些指标旨在评估预测结果的全局准确性，并依赖于一组已标注的参考文本作为真值标签。此外，Zheng等人[39]针对视频描述生成式任务的特性，提出层次化多源不确定性聚合框架，识别语法、概念和对齐等层面的不确定性来源，并利用“条件空值度”（Conditional Vacuity）评估预测结果的全局不确定性，以应用于主动学习过程中标注样本的选择。与以上方法不同，本文提出的短语级质量评估模型针对文本标注自身的质量进行评估，实现短语级标注质量量化。

3　模型设计

3.1　总体概述

由于文本标注的主观性，视频描述任务通常汇集不同标注者对同一视频的多组描述以提高标注的完备性。因此，每个特定的视频X都有一组文本标注，记作$\Gamma=\{Y_1, Y_2, \cdots, Y_M\}$，其中$M$表示与该视频对应的文本描述数量。将输入视频$X$与每条文本标注$Y_m$分别配对，构造用于训练的样本对。

图2展示了本文模型的整体训练架构，主要包括文本内容感知的视频特征表达和短语级质量感知的抗噪训练。其中，文本内容感知的视频特征表达提取与文本描述一致性高、采样冗余性低的关键帧，进而形成语义完备的视频特征表达，以克服视觉模态的不确定性。短语级质量感知的抗噪训练首先通过文本标注的短语级质量评估从不同维度预测标注的短语级质量，再通过渐进式融合策略形成短语的综合质量，最终通过标注质量引导的自适应训练指导模型在学习过程中动态关注高质量的局部标注，同时降低标注噪声的影响，以克服文本模态的不确定性。

3.2　文本内容感知的视频特征表达

文本内容感知的视频特征表达首先将视频切分为多样且差异显著的场景原型；随后基于场景原型与文本描述之间的相关性，从每个场景原型中采样与文本语义一致的关键帧，以此提升视频与文本的对齐精度、减少冗余帧采样；最终送入视频编码器以形成文本内容感知的视频特征表达。整体结构如图2所示。

3.2.1　场景原型的建立

给定输入视频X，均匀采样N帧以构建候选帧集合，记作$F=[F_1, F_2, \cdots, F_N]$。利用预训练的多模态大模型CLIP，分别对每一候选帧和文本描述Y_i编码，得到候选帧的全局视觉特征$V^g=\{v_1^g, v_2^g, \cdots, v_N^g\}$和文本的全局特征$a^g$，其中$v_i^g$表示第$i$帧的全局特征向量。

为构建场景原型，采用基于欧氏距离的K-means算法对所有候选帧的全局特征聚类。聚类算法采用k-means++初始化策略，通过概率化中心点选择机制克服传统K-means对初始值的敏感性。聚类分析将候选帧的语义特征空间划分为K个语义一致的簇，每个簇对应一种视频场景原型。最终生成的K个场景原型以各聚类中心为语义特征表示，记为$[c_1, c_1, \cdots, c_K]$。

3.2.2　基于场景原型的帧采样与编码

为了筛选出最相关的帧，本文基于多模态语义的相关性设计关键帧选择策略。具体地，对于每一

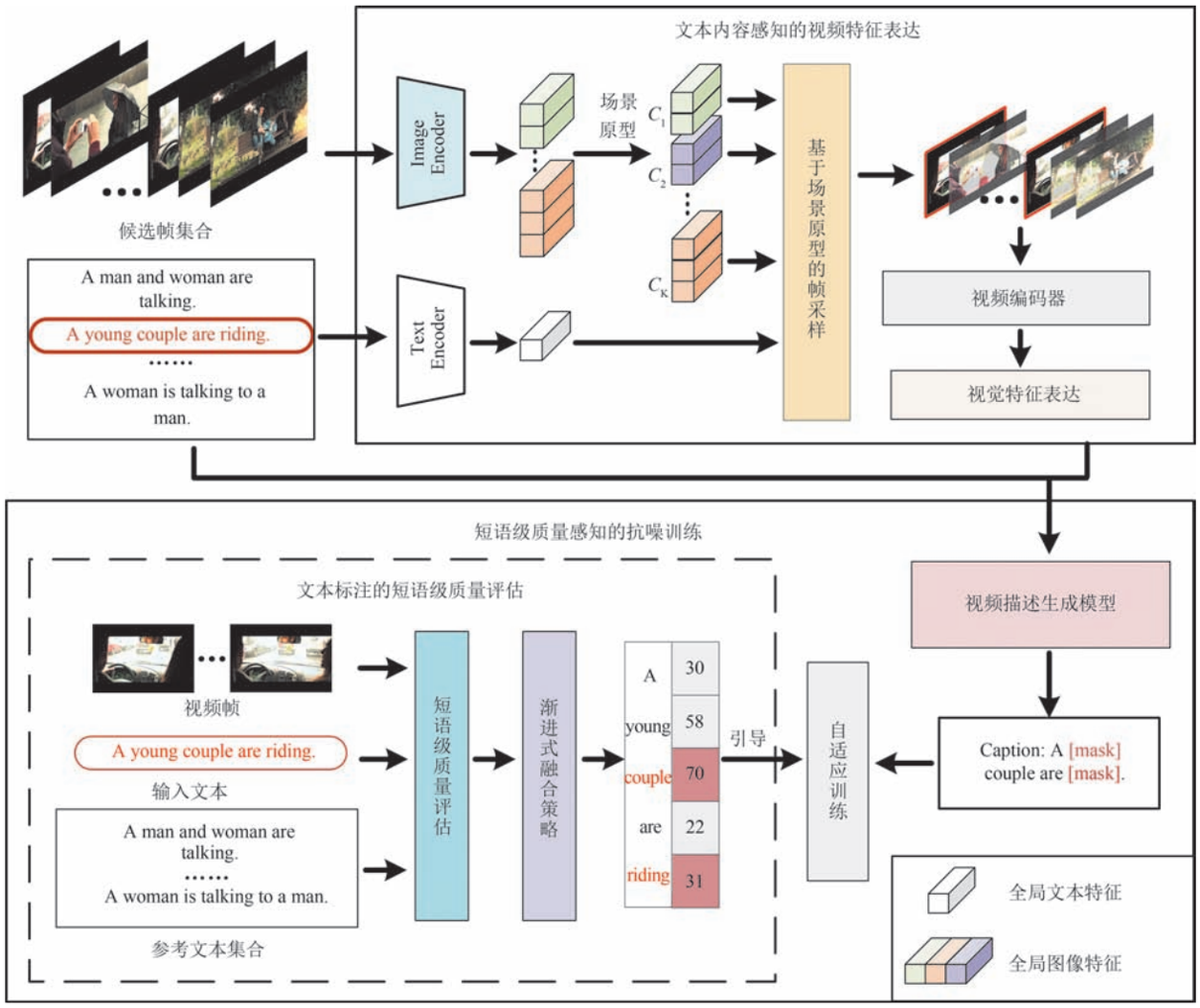

图 2 模型总体框架图

个场景原型，计算其与文本全局特征 a^g 之间的余弦相似度作为该场景原型的得分。第 k 个场景原型的得分 s_k 定义如下：

$$s_k=\frac{\langle a^g, c_k\rangle}{\|a^g\|_2\|c_k\|_2} \tag{1}$$

由于较高的得分表示该场景与文本之间的匹配程度更强，因此应从该场景采样更大比例的关键帧，反之亦然。为了由语义相似性确定采样比例，本文对场景原型分数 s_k 进行归一化：

$$r_k=\frac{\exp(s_k/\tau)}{\sum_{k=1}^{K}\exp(s_k/\tau)} \tag{2}$$

其中，τ 是温度参数，r_k 是采样比例。基于采样比例，在第 k 个场景原型中采样 f_k 帧：

$$f_k=\lfloor \bar{N}*r_k \rfloor \tag{3}$$

其中，$\bar{N}$ 代表模型中最终采样的帧数。接着从每个场景原型中提取每一帧的视觉特征，计算这些视觉特征与文本的余弦距离作为该帧与文本的相似度；随后从每个场景原型中选择相似度最高的前 f_k 帧。为了严格保持视频原有的时序关系并避免跨场景抽样导致的时序断裂，我们将所有筛选出的帧按其原始视频时间戳进行升序排列，形成最终的、全局时序一致的帧序列 $F=[F_{\hat{1}}, F_{\hat{2}}, \cdots, F_{\hat{N}}]$。最后，将筛选后的帧序列 F 输入视频编码器[40]，形成文本语义感知的视频特征表达 $\bar{V}$。

3.3 文本标注的短语级质量评估

文本描述可能包含与视频内容无关的错误短语。为了缓解文本标注的局部噪声对视频描述模型的影响，提出文本标注的短语级质量评估模块，克服标注质量无参考的局限性，联合建模文本标注的模态内语义关联性以及文本与视频内容的模态间相关性，从多维度评估文本的短语级质量，整体结构如图3所示。

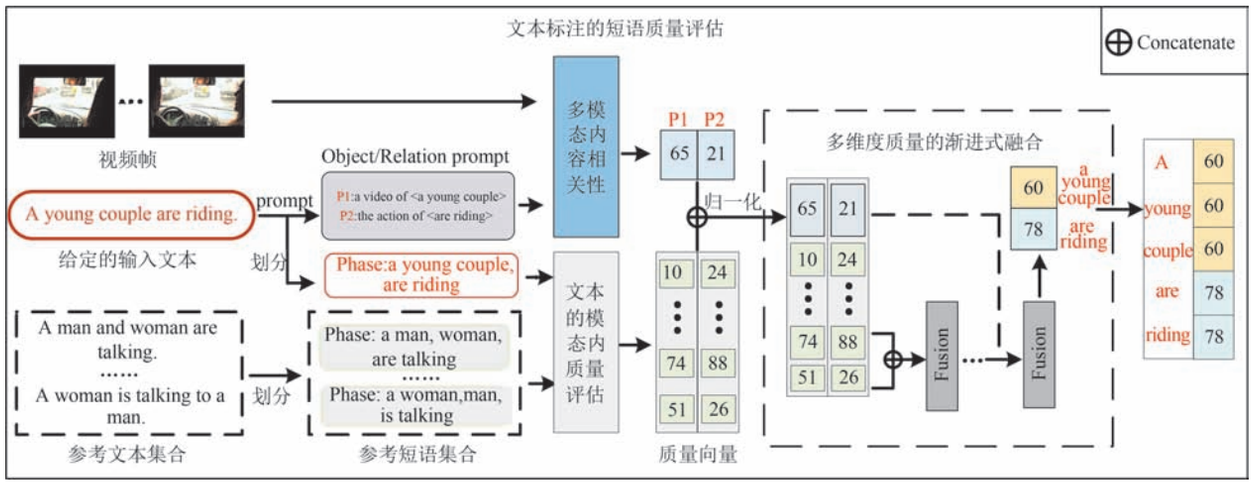

图 3　文本的短语级质量评估流程图

3.3.1　文本的模态内质量评估

由于文本标注的质量缺少真值质量标签，本文从统计角度提出无监督的文本全局质量评估方法。其核心思想在于，标注集合Γ中，若某个标注文本与集合内越多数量的其他标注呈现高语义相关性，则表明它是高质量标注文本。通过这种方式，将标注质量评估转化为可量化的语义相关性统计分析，解决真值质量标签缺失的挑战。

具体地，本文采用留一法，对于文本描述Y_m，将同一视频中的其余文本描述$\tilde{\Gamma}_m=\{Y_1,\cdots,Y_{m-1},Y_{m+1},\cdots,Y_M\}$视为伪真值标注。为实现短语级文本标注质量的评估，本文将Y_m拆解为若干短语，记为$h_m=\{h_1,h_2,\cdots,h_Z\}$($Z$为短语总数)。随后，采用BLEU等指标对$h_z$的质量进行量化。

$$\tilde{\boldsymbol{q}}_{z,j}=\mathrm{Indicator}_j\left(h_z,\tilde{\Gamma}_m\right) \tag{4}$$

其中，$\mathrm{Indicator}_j$表示第j个评估指标，$\tilde{\boldsymbol{q}}_{z,j}$表示$Y_m$中第$z$个短语与$\tilde{\Gamma}_m$计算的语义一致性。

由于现有的文本评估指标并不完美，且不同指标具有一定互补特性(如BLEU强调词序列匹配度，CIDEr侧重文本的共识性)。因此，本文应用多个指标评估h_z的质量，并将所有质量分数构成向量$\tilde{\boldsymbol{q}}_z=[\tilde{\boldsymbol{q}}_{z,1},\tilde{\boldsymbol{q}}_{z,2},\cdots,\tilde{\boldsymbol{q}}_{z,J}]$，其中$J$表示评估指标的数量。

3.3.2　基于多模态内容相关性的质量评价

此外，进一步通过文本标注与视频的语义一致性来评估文本标注的短语级质量。然而，由于短语本身缺乏上下文语境，直接计算其与视频的跨模态一致性难以捕捉准确的语义对应关系。为此，本文借鉴预训练多模态模型中的提示调优(Prompt Tuning)策略[41]，将短语嵌入自然语言提示模板中，使得短语在被输入模型前具备更明确的语义结构和语境信息，从而提升其在多模态空间中的可识别性与可对齐性，最终增强与视频内容的语义一致性。

具体来说，本文设计不同的提示词模板分别生成对象短语和动词短语[41]，例如，对象短语的提示词模板为*“a video of*[名词]*”*，动词短语的提示词模板为*“the action of*[动词]*”*。在实际使用过程中，[名词]为*“persons”“an airplane”*等具体表述，[动词]为*“are walking”“is landing”*等对应内容。随后，采样的视频帧F与提示词同时输入预训练的视觉-语言模型计算文本与视频的跨模态相似性，得到其跨模态评估质量$\tilde{q}_{z,J+1}$。将$\tilde{\boldsymbol{q}}_z$与$\tilde{\boldsymbol{q}}_{z,J+1}$进行拼接，形成短语级质量向量$\boldsymbol{q}_z=[\tilde{\boldsymbol{q}}_{z,1},\tilde{\boldsymbol{q}}_{z,2},\cdots,\tilde{\boldsymbol{q}}_{z,J},\tilde{\boldsymbol{q}}_{z,J+1}]$。最终将所有文本描述的短语级质量向量合并形成$\boldsymbol{Q}=[\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_U]^{\mathrm{T}}$。其中$U$表示所有文本描述的短语数量，$\boldsymbol{Q}$的每一行表示同一个短语在不同指标上的质量，每一列表示不同的句子在同一个指标上的质量。

3.3.3　多维度质量的渐进式融合

不同评估指标的评分分布存在显著差异，例如BLEU@4指标的评分均值比BLEU@1的均值低。此外，由于评估指标自身的局限性，导致不同文本集合Y的评分分布受噪声干扰，形成不同程度的长尾效应，难以直接比较。为克服以上问题，本文对质量矩阵$\boldsymbol{Q}$的每一列进行最小-最大归一化，得到归一化的质量向量$\bar{\boldsymbol{q}}_z=[\bar{\boldsymbol{q}}_{z,1},\bar{\boldsymbol{q}}_{z,2},\cdots,\bar{\boldsymbol{q}}_{z,J+1}]$。

最后，本文提出渐进式融合策略对不同的文本描述质量进行动态融合，通过非线性、逐步的信息聚合实现更全面且精确的文本全局质量评估。该策略

能够避免低质量评分对整体表示造成突发性干扰，有效提升融合结果的稳定性与鲁棒性。具体地，对于两个质量分数$\bar{\boldsymbol{q}}_{z,1}$，$\bar{\boldsymbol{q}}_{z,2}$，采用门机制进行动态融合，如式5所示：

$$\alpha=\mathrm{Sigmoid}(\mathrm{Linear}(\mathrm{Concat}(\bar{\boldsymbol{q}}_{z,1},\bar{\boldsymbol{q}}_{z,2}))) \quad (5)$$

$$\bar{\bar{q}}_z=\alpha*\bar{\boldsymbol{q}}_{z,1}+(1-\alpha)*\bar{\boldsymbol{q}}_{z,2}$$

其中，Linear(•)表示将输入向量映射为一个标量值。$\bar{\bar{q}}_z$依次与下一个质量分数按照相同方式融合，直到所有元素合并为一个值，得到h_z的短语级质量评分q_z，将Y_m的所有短语质量拼接得到最终的质量评分$q_m=[q_1,q_2,\cdots,q_Z]$。

3.4 标注质量引导的自适应训练

本文可以采用大部分视频描述生成模型作为基础模型，包括基于Transformer的模型结构（如SwinBERT）和基于大语言模型的结构（如VideoLLaMA2）。以SwinBERT为例，其采用跨模态联合建模方法实现文本与输入视觉特征的深度语义融合。模型主体由级联的浅层Transformer模块构成，每个Transformer模块均包含多头自注意力机制层和动态前馈网络层。通过层级式特征抽象机制，模型逐步聚合多尺度时空特征，最终生成文本描述。模型采用自回归式的生成方案，将$\bar{\mathrm{V}}$和生成的词序列$y_{1:t-1}$进行拼接后作为输入，视频生成模型预测下一个词y_t。

本文将短语级质量评估模块（PCE）计算出的短语级别的文本描述质量作为引导信号融入损失函数中进行引导训练。该训练策略在训练过程中引导模型优先学习高质量的文本描述，同时有效地忽略噪声文本描述的干扰，从而显著减轻注释噪声对模型训练的负面影响。

遵循常规做法，本文方法以视频-文本描述对为输入，通过短语级质量q_t修正交叉熵损失函数，指导模型的训练：

$$\mathcal{L}_{cap}=-\sum_{t=1}^{T}q_t*\log(p_t(y_t|y_{1:t-1},\bar{\mathrm{V}})) \quad (6)$$

式(6)中，低质量的文本标注提供的监督信息被弱化，从而有效降低了模型受标注噪声的影响。

4 实验结果与分析

4.1 实验准备

（1）数据集：本文利用MSR-VTT[14]和MSVD[15]两个广泛使用的数据集对视频描述的性能进行评价。其中MSR-VTT数据集包含10 000个视频片段，每个视频配有20个多样的句子描述，提供了全面且多维的视角以反映视频内容。按照标准数据集划分，本文使用6513个视频用于训练，497个用于验证，2990个用于测试。MSVD数据集包含1970个视频片段，每个片段配有约40个自然语言句子，捕捉视频内容的各个方面。与标准划分一致，文本使用1200个样本用于训练，100个用于验证，670个用于测试。

（2）评估指标：为了全面评估模型生成的视频描述的质量，本文采用视频描述任务的标准评估指标，即BLEU[31]、METEOR[33]、ROUGE-L[32]和CIDEr[34]定量评估模型的效果。

（3）实现细节：本文采用两种具有代表性的基础模型，分别是SwinBERT和VideoLLaMA2。对于SwinBERT，文本内容感知的视频特征表达模块中将候选帧数N设置为64，最终采样帧数$\bar{N}$设置为32。场景原型的取值K设置为3，由消融实验确定。视频编码器使用在ImageNet[42]上预训练的Video Swin Transformer[40]模型。对于短语级质量感知的抗噪训练模块，本文与CVG[9]一致，对标注的文本描述应用了最小的预处理步骤，包括转换为小写、去除标点符号，并在每个文本描述的开头和结尾添加[*BOS*]和[*EOS*]标记。词汇表中不存在的单词将被替换为[*UNK*]标记。句子被限制为固定长度50，对于超过该长度的句子进行截断，对于不足的句子进行填充。词嵌入的大小设定为512。在训练阶段，本文使用Adam优化器进行25个epoch的模型优化。初始学习率设为3e-5，并在每经过3个epoch后将其降低为原来的0.8倍。对于VideoLLaMA2，将$\bar{N}$设置为16，模型其余设置与VideoLLaMA 2-7B保持一致。在训练阶段，本文使用AdamW优化器进行1个epoch的模型优化。初始学习率设为2e-5，并以余弦衰减。

4.2 消融实验与分析

为验证本文模型中各个模块的有效性，以SwinBERT为基础模型，在MSR-VTT验证集上进行了消融实验。不失一般性，所有实验中$\bar{N}$设置为8。

4.2.1 文本内容感知的视频特征表达的有效性

为了验证所提的文本内容感知的视频特征表达模块的有效性，本文对比了基线模型，即视频中均匀采样8帧，并禁用短语级质量感知的抗噪训练模块，

结果如表1所示。无论场景原型K的取值如何变化，模型的效果均始终优于基线模型。值得注意的是，$K=1$时，模型退化为基于CLIP相似度进行关键帧筛选；随着K的不断增大，模型性能持续提升，表明原型数量的增长有助于增强帧采样的多样性，从而构建更丰富的视频语义表示；当$K=3$时，模型整体性能提升最为显著，特别是在CIDEr分数上达到了57.7，比基线提高了2.4，显示出该配置下帧丰富性与语义对齐达到了较佳平衡。随着K的进一步扩大，模型性能逐渐降低，这是因为原型数量过多可能引入与描述文本相关性较低的帧，进而削弱语义一致性，从而降低模型的性能。总体而言，文本内容感知的视频特征表达模型通过构建语义原型并筛选关键帧的方式，不仅提升了帧与文本之间的相关性，也有效缓解了单一选帧策略的局限，进一步增强了模型的鲁棒性与生成描述的丰富性。

表1　MSR-VTT验证集上场景原型的数量对视频描述生成任务性能的影响

	B@1	B@4	M	R	C
基线方法	84.5	41.6	30.0	61.8	55.3
TAVR(K=1)	85.6	43.3	30.6	62.7	56.3
TAVR(K=2)	85.8	43.2	30.6	62.9	56.9
TAVR(K=3)	86.0	43.8	30.8	63.4	57.7
TAVR(K=4)	85.7	43.7	30.2	63.0	57.6

注：其中B@1、B@4、M、R、C分别表示BLEU@1、BLEU@4、METEOR、ROUGE-L、CIDEr。

4.2.2　文本标注的短语级质量评估的有效性

本文研究了在短语级质量评估中不同质量指标对模型性能影响。实验中，选择了BLEU@N、ROUGE-L、METEOR、CIDEr、BERTScore和多模态语义一致性(V-L)作为质量指标来评估文本描述质量。所有实验均未使用文本内容感知的视频特征表达以进行公平比较，具体结果见表2。

首先，从单一指标角度来看，基于文本模态内的质量评估指标(如BLEU和CIDEr)以及使用多模态一致性(V-L)获得的文本质量均表现出比基线模型更为稳定和优越的性能。这是因为将文本质量纳入损失函数中，可以指导模型从高质量的注释中学习，从而生成更优化的文本描述。具体来说，V-L相较基线模型将CIDEr分数提高了2.7；在基于文本模态内的评估指标中，METEOR表现最为优秀，比基线提升了4.0，这是由于METEOR通过同义词和短语匹配评估词汇丰富度和句子连贯性，因此能够准确反映相似场景下短语的质量。此外，表2结果显示，融合不同质量指标融合进一步提高了视频描述模型的性能——尤其是当V-L与METEOR融合时，模型的CIDEr分数达到61.0。相比基线模型在CIDEr指标上提升了5.7。融合METEOR、CIDEr和V-L三个维度质量时，模型性能CIDEr分数显著提升至61.3分，相比基线模型在CIDEr指标上提升了6.0。这归因于本文在评估文本质量时不仅考虑了候选句子，还特别关注了文本描述与视频内容语义一致性。以上结果验证了多维度质量的渐进式融合策略的有效性以及基于文本的短语级质量引导的抗噪训练有效提高了视频描述方法的准确性。

表2　MSR-VTT验证集上文本的短语级质量评估的结果，针对不同度量标准进行的消融研究

质量指标						评估指标
B@1	R	M	C	BERTs	V-L	C
-	-	-	-	-	-	55.3
√	-	-	-	-	-	58.7
-	-	√	-	-	-	59.3
-	√	-	-	-	-	57.9
-	-	-	√	-	-	58.3
-	-	-	-	√	-	57.8
-	-	-	-	-	√	58.0
√	-	-	-	-	√	59.0
-	-	√	-	-	√	61.0
-	√	-	-	-	√	59.3
-	-	-	√	-	√	59.3
-	-	√	√	-	√	61.3

注：其中符号“√”表示包含以下度量标准，B@1、M、R、C、BERTs、V-L分别表示BLEU@1、METEOR、ROUGE-L、CIDEr、BERTScore、多模态语义一致性。

4.3　对比实验结果

为验证本文所提框架的有效性，本文将其与24种前沿方法展开全面对比，涵盖SwinBERT[6]、MAN[7]、HMN[8]、CVG[9]、SAAT[43]、CMG[44]、CoCap[45]、RSFD[46]、TextKG[47]、IcoCap[48]、Track4Cap[49]、VCRN[50]等经典方法，以及VideoLLaMA2[19]、VALOR[21]、Vid2Seq[22]、MELTR[51]、VL-Prompt[52]、OmniViD[53]、等基于大模型的方法。比较结果如表3和表4所示。

1)在MSR-VTT测试集的结果：在MSR-VTT数据集的实验中，本文方法通过构建文本语义感知的视频特征表达，并聚焦标注的短语级质量感知的学习模型，展现出全面的性能优势。如表3所示，基于SwinBERT基线模型，本文方法在所有评估指标

表3 与主流方法在MSR-VTT测试集性能比较

方法	出版时间	MSR-VTT				
		B@1	B@4	R	M	C
SAAT[43]	2020	79.6	39.9	61.2	27.7	51.0
APML[11]	2021	82.9	43.8	63.6	30.3	52.2
SGN[12]	2021	-	40.8	60.8	28.3	49.5
SwinBERT[6]	2022	83.1	41.9	62.1	29.9	53.8
LSRT[23]	2022	-	42.6	61.0	28.3	49.5
CMG[44]	2022	83.5	44.9	62.9	29.6	53.0
Cocap[45]	2023	-	44.4	63.4	30.3	57.2
RSFD[46]	2023	-	43.4	62.3	29.3	53.1
TextKG[47]	2023	-	46.6	64.8	30.5	60.8
HMN[8]	2024	81.3	43.5	62.7	29.0	51.5
IcoCap[48]	2024	-	47.0	64.9	31.1	60.2
MAN[7]	2024	-	43.5	62.2	28.9	50.9
CVG[9]	2024	84.8	-	64.6	31.2	60.2
HSRA[24]	2025	-	46.9	64.8	30.9	55.3
Track4Cap[49]	2025	-	44.6	63.6	30.5	57.7
VCRN[50]	2025	-	41.5	61.2	28.1	50.2
本文(+SwinBERT)	2025	86.8	48.3	65.4	31.7	63.6
MELTR[51]	2023	-	44.2	62.4	29.3	52.8
VL-Prompt[52]	2023	-	43.2	62.7	30.1	55.3
OmniViD[53]	2024	-	44.3	62.7	29.9	56.6
Vid2Seq[22]	2023	-	-	-	30.8	64.6
MA-LMM[26]	2024	-	-	-	33.4	74.6
COSA[25]	2024	-	53.7	-	-	74.7
VideoLLaMA2[19]	2024	88.1	52.5	67.2	33.1	72.6
VALOR-L[21]	2025	-	54.4	68.0	32.9	74.0
本文(+VideoLLaMA2)	2025	88.3	54.4	68.1	33.6	76.8

注：B@1、B@4、M、R、C分别表示BLEU@1、BLEU@4、METEOR、ROUGE-L、CIDEr。

表4 与主流方法在MSVD测试集性能比较

方法	出版时间	MSVD				
		B@1	B@4	R	M	C
APML[11]	2021	86.4	58.0	76.2	39.2	108.3
SwinBERT[6]	2022	-	58.2	77.5	41.3	120.6
Cocap[45]	2023	-	60.1	78.2	41.4	121.5
TextKG[47]	2023	-	60.8	75.1	38.5	105.2
VL-Prompt[52]	2023	-	63.5	78.9	41.6	128.1
MAN[7]	2024	-	60.1	74.2	37.3	101.5
IcoCap[48]	2024	-	59.1	76.5	39.5	110.3
HMN[8]	2024	-	59.7	74.3	37.3	101.5
OmniViD[53]	2024	-	59.7	78.1	42.2	122.5
HSRA[24]	2025	-	62.2	78.4	39.2	110.1
Track4Cap[49]	2025	-	62.1	79.8	42.5	127.2
VCRN[50]	2025	-	59.1	74.6	37.4	100.8
本文(+SwinBERT)	2025	89.6	65.2	81.2	43.8	138.5
VL-Prompt[52]	2023	-	63.5	78.9	41.6	128.1
OmniViD[53]	2024	-	59.7	78.1	42.2	122.5
Vid2Seq[22]	2023	-	-	-	45.3	146.2
MA-LMM[26]	2024	-	-	-	49.8	179.1
COSA[25]	2024	-	76.5	-	-	178.5
VideoLLaMA2[19]	2024	-	73.9	85.9	49.3	174.3
VALOR-L[21]	2025	-	80.7	87.9	51.0	178.5
本文(+VideoLLaMA2)	2025	-	77.4	87.9	51.8	179.1

注：B@1、B@4、M、R、C分别表示BLEU@1、BLEU@4、METEOR、ROUGE-L、CIDEr。

上均超越现有最优方案，尤其在CIDEr指标上取得63.6，较当前最佳模型TextKG提升2.8，其余指标也均实现0.5点以上的提升。与采用帧级掩码策略的APML方法相比，本文在CIDEr指标上实现11.1的绝对增益，这得益于对齐性强化与聚类驱动的帧选择策略共同作用。进一步对比近期主流模型，本文方法持续保持领先：相较于CVG提升3.4，IcoCap提升3.4，HSRA提升8.3，Track4Cap提升5.9，HMN提升12.1。值得一提的是，本文方法对比MAN和VCRN上提升显著，分别提升了12.7和13.4。基于VideoLLaMA2基础模型，相比基线模型在CIDEr指标上提高4.2。对比其他基于大规模视觉-语言预训练的Vid2Seq（提升12.2）、COSA（提升2.1）和VALOR（提升2.8），本方法仍展现出显著优势。

2）在MSVD测试集上的结果：本文进一步在MSVD数据集上进行了分析。如表4所示，基于SwinBERT基础模型，本文方法在所有指标上均达到最佳性能，具体表现为：BLEU@1分数89.6，BLEU@4分数65.2，METEOR分数43.8，ROUGE-L分数81.2，CIDEr分数138.5。进一步对比近期主流模型，本文方法持续保持领先：相较于HSRA，本文在CIDEr指标上提升28.4，Track4Cap提升11.3，VCRN提升37.7。值得一提的是，本模型相较当前主流的OmniViD框架实现16.0分的跨越式提升，较采用视觉-语言预训练范式的VL-Prompt模型也取得10.4分的显著优势；基于VideoLLaMA2基础模型，相比基线模型在CIDEr指标上提高4.8，达到179.1，与MA-LMM持平，但METEOR显著高于MA-LMM。与其他基于大模型的Vid2Seq、COSA、VALOR等方法相比，本方法展现出一致的性能优势。这充分证明模型在视频内容深度理解和语义层次化表达方面的有效性。

4.4 主观对比实验

本文在MSR-VTT测试集中随机抽取100个样本进行人工评估，从相关性和丰富性两个维度对基

线模型(SwinBERT)和本文模型的预测结果进行对比分析。其中相关性表示生成的标题是否反映了视频的内容，而丰富度则衡量句子中富含信息量的描述是否充分。具体地，将每个视频及其对应的两个模型生成的描述(随机标记为A和B)展示给3位不同的评估人员，要求他们分别从相关性和丰富性两个角度在“A更好”、“B更好”和“无法区分”三个选项中进行选择。统计规则为：若某个描述获得的“更好”评价更多，则判定该结果“胜出”；若两者获得的“更好”评价数量相同，则判定为“无法区分”。从表5中可以看出，我们的模型在相关性和丰富性两个指标上的“胜出”比例均显著高于基线模型。

表5　MSR-VTT测试集上用户评测的结果

模型	丰富性	相关性
SwinBERT更好	19%	24%
本文方法更好	33%	35%
无法区分	48%	41%

4.5　计算参数量比较

本文所提出的文本内容感知的视频关键帧采样和文本标注的短语级质量评估可以在训练开始阶段一次性计算完成，后存储在硬盘中供训练过程中读取。模型训练过程中，与基线模型相比仅渐进式融合策略中包含一个全连接层，因此对训练效率几乎无影响，测试阶段不涉及帧抽取和质量评估模块，因此不会对推理速度造成任何额外开销。

表6展示了不同模型的参数量与在MSR-VTT数据集上的性能比较。在训练阶段，本文方法在参数量上仅比基准模型增加了0.1 M，这是由于所引入的渐进式融合策略中仅包含一个轻量级的全连接层，因此几乎不引入额外的计算负担。在取得63.6的CIDEr分数(较CoCap提高6.4)的同时，参数量仅为CoCap的51.2%，远低于MAN等大型模型(参数量为1500 M)，体现了良好的成本与性能的平衡。

4.6　可视化分析

(1)基于场景原型的帧采样的可视化

图4展示了本文方法在帧采样方面的有效性分析。图4(a)中对比了传统的均匀采样策略(第一行)与本文所提出方法选取的帧序列(第二行)。可以观察到，虽然均匀采样能够覆盖视频的整个时间跨度，但其后四帧与文本描述“a man and child play games at a theme park”的语义相关性较弱。这些不相关帧的存在可能对模型训练产生干扰，降低描述生成的准确性与连贯性。相比之下，本文所提出的帧采样策略在结合文本语义信息的基础上，通过选择与描述高度相关的帧，同时保留场景多样性，有效提升了帧序列的语义代表性与覆盖广度。图4(b)进一步对比了在另一场景下均匀采样与本文方法的帧选择效果。均匀采样结果中，部分帧存在明显语义不一致现象，甚至出现了全黑帧的无效帧，不仅无法提供有用的语义信息，反而会引入干扰。相较之下，本文方法选出的帧在确保与描述“a woman with wings and horns flies in the sky and then lands”语义一致的同时，还富有多样性，有效覆盖视频的关键变化过程，从而生成更为完整、自然且富有表现力的视频语义表征。

表6　MSR-VTT数据集上参数量的比较

模型	参数量	CIDEr
SwinBERT	225.7 M	53.8
CoCap	441 M	57.2
MAN	1500 M	50.9
本文方法	225.8 M	63.6

(2)描述质量评估结果的可视化

接下来，本文从MSR-VTT与MSVD数据集挑选了四个视频片段，并在图5中展示了短语级质量评估的结果。在图5(a)的描述文本中，“a”和“girl”为噪声，其质量得分分别为0.05与0.08；而“driving”、“car”等为正确描述，其质量得分相对较高。图5(b)的描述中，“a pig”是错误的标注，并未在原视频中出现，因此预测的质量得分仅为0.11和0.15。类似地，图5(c)中的“the girl”也未在原视频中出现，这一动作主体可能是标注者的猜测而非真实观察的结果，此外“cut”这一动作也描述错误(视频中操作人在折纸而非剪纸)，因此该例中的名词和动词的得分都较低。图5(d)的描述“a group is dancing”中，该描述与视频中实际发生的内容并无关系，因此其短语级质量得分普遍都偏低。

(3)视频描述生成的结果可视化

图6展示了从MSR-VTT和MSVD数据集中选取的基线方法(SwinBERT)和本文的方法生成描述的例子。通过文本内容感知的视频特征表达以及短语级的文本描述噪声感知训练策略，本文的方法不仅能够使得生成的文本描述内容更加丰富，还能够更有效地识别标注文本的局部噪声并削弱噪声数据对模型的影响，从而生成更加准确的文本描述。图6(a-c)展示了本文的模型在文本描述生成的丰富

图 4 基于场景原型的帧采样结果可视化

图 5 描述质量评估结果的可视化

性。例如图6(a)所示，基线方法只是简单的生成描述“a man is playing ping pong ball.”，而本文方法的描述更加丰富，增加了“in a blue shirt”这一细节，这是因为本文的模型在实现视频与文本内容对齐的过程中，使用聚类的方法自动地从每一类视频帧中自动地分配，确保所抽取的视频帧在维持语义一致性的同时，蕴含更多元化的细节，从而生成更加丰富的描述。图6(b)和图6(c)同样如此，基线方法生成“a man is cooking”等描述，但是用来描述整个视频却略显单调，而本文的方法分别比基线方法增加了“in a black shirt”、“in a pot”和“down an alley”这些细节描述，使得描述更加丰富饱满。这种对场景细节的

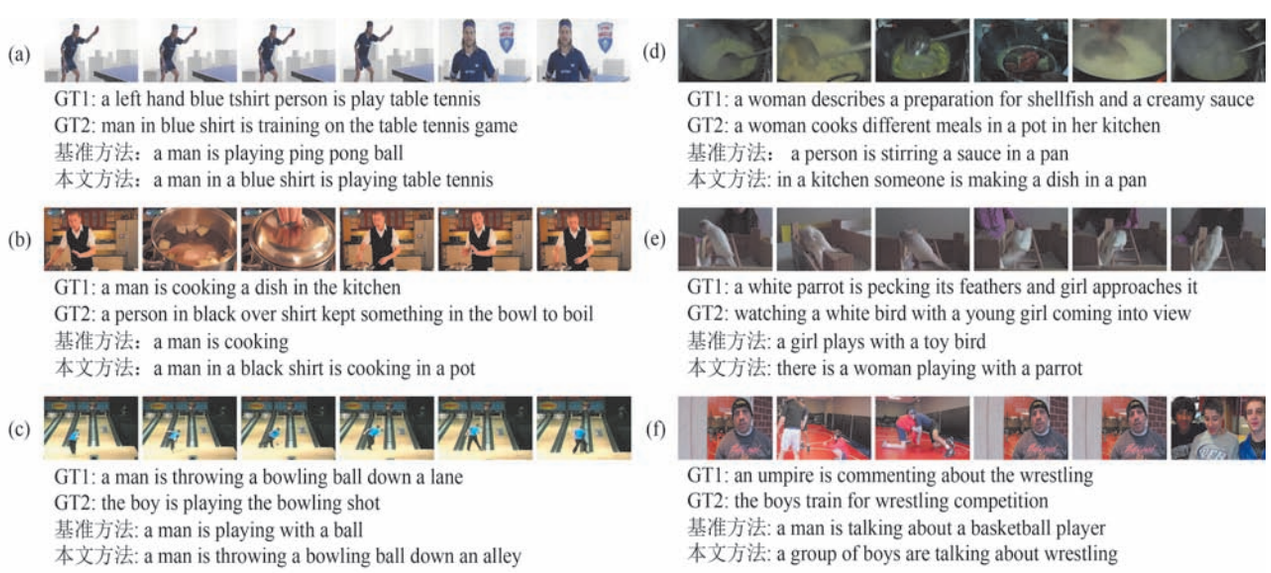

图 6 本文与其他代表性模型预测结果可视化对比

捕捉和表达，正是本文基于文本内容感知的视频特征表达方法所擅长的。

图6(d-f)展示了本文的模型在文本描述生成的准确性。例如图6(d)所示，视频标注文本GT1中存在噪声数据“a creamy sauce”，基线方法受其噪声误导错误地生成“a sauce”，而本文的方法着重从高质量标注文本进行学习，从而降低噪声影响以生成更为准确无误的描述；在图6(e)中，基线方法错误地将“parrot”预测为“a toy bird”，这是因为在噪声数据中学习会学习到错误的视频-文本之间的语义关系，使得模型预测到相关但不足够准确的描述，而本文的方法能降低噪声数据影响从而学习到准确的视频-文本之间的语义对齐关系，从而准确地预测出“a parrot”；在图6(f)中，基线方法同样受噪声的影响错误地将“wrestling”预测为“basketball”，而本文的方法基于噪声感知的训练策略，准确地预测出“wrestling”。

5 结 论

本文针对视频与对应的文本描述之间存在关联的不确定性提出了一个新的框架，通过引入文本内容感知的视频特征表达和短语级质量感知的抗噪训练模块，显著提升了文本描述生成的准确性和丰富性。其中，视频特征表达通过场景原型驱动的帧采样策略，结合文本语义筛选关键帧，在减少冗余的同时增强视频表征的语义丰富性与跨模态对齐性。其次，模型的抗噪声训练模块提出短语级质量评估，准确识别文本描述中的噪声，并优先从高质量文本片段中学习，避免了低质量文本描述对模型训练的负面影响。通过这两个模块的协同作用，模型能够更有效地处理标注噪声带来的问题，减少了文本描述生成中的错误和不一致性，从而提升了生成结果的准确性与连贯性。实验表明，本文设计的模型效果显著，在多个评价指标上得分均有显著提高。未来工作将进一步探索多模态预训练模型与采样的深度融合，以应对更复杂的视频描述生成任务。

参 考 文 献

[1] Tang Peng-jie, Wang Han-li. From video to language: a survey of research on video caption generation and description. Acta Automatica Sinica, 2022, 48(2): 375-397 (in chinese)
(汤鹏杰，王瀚漓. 从视频到语言：视频标题生成与描述研究综述，自动化学报，2022，48(2)：375-397)

[2] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks//Proceedings of the IEEE/CVF International Conference on Computer Vision. Santiago, Chile, 2015: 4489-4497

[3] He Kaiming, Zhang X., Ren Shaoqing, Sun Jian. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[4] Song Peipei, Dan Guo, Zhou Jinxing, Xu Mingliang, Meng Wang. Memorial GAN with joint semantic optimization for unpaired image captioning. IEEE Transactions on Cybernetics, 2023, 53(7):4388-4399

[5] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the AAAI Neural Information Processing Systems. Long Beach, USA, 2017: 5998-6008

[6] Lin K, Li Linjie, Lin Chung-Ching, alel. SwinBERT: End-to-end transformers with sparse attention for video captioning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 17928-17937

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

[8] Li GuoRong, Ye Hanhua, Qi Yuankai, et al. Learning hierarchical modular networks for video captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 46(2): 1049-1064

[9] Jiang Wenhui, Cheng Yibo, Liu Linxin, et al. Comprehensive visual grounding for video description//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 2552-2560

[10] Hou Jing-yi, Qi Ya-yun, Wu Xin-xiao, Jia Yun-de. Cross-lingual knowledge distillation for chinese subtitle generation in videos. Chinese Journal of Computers, 2021, 44(9): 1907-1921 (in chinese)
(侯静怡，齐雅昀，吴心筱，贾云得. 跨语言知识蒸馏的视频中文字幕生成. 计算机学报，2021，44(9)：1907-1921)

[11] Lin Ke, Gan Zhuoxin, Wang Liwei. Augmented partial mutual learning with frame masking for video captioning//Proceedings of the AAAI Conference on Artificial Intelligence. Online, 2021: 2047-2055

[12] Ryu Hobin, Kang Sunghun, Kang Haeyong, Yoo Chang Dong. Semantic grouping network for video captioning//Proceedings of the AAAI Conference on Artificial Intelligence. Online, 2021: 2514-2522

[13] Kang W, Mun J, Lee S, Roh B. Noise-aware learning from web-crawled image-text data for image captioning//Proceedings of the IEEE International Conference on Computer Vision. Paris, France, 2023: 2930-2940

[14] Xu Jun, Tao Mei, Ting Yao, Yong Rui. MSR-VTT: A large video description dataset for bridging video and language//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 5288-5296

[15] Chen D, Dolan W. Collecting highly parallel data for paraphrase

evaluation//Proceedings of the Association for Computational Linguistics. Portland, USA, 2011: 190-200

[16] Anderson P, He Xiaodong, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6077-6086

[17] Shi Yaya, Xu Haiyang, Yuan Chunfen, Li Bing, Hu Weiming, Zha Zhengjun. Learning video-text aligned representations for video captioning. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19: 1-21

[18] Wu Bofeng, Liu Buyu, Huang Peng, Bao Jun, Xi Peng, Yu Jun. Concept parser with multimodal graph learning for video captioning. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33: 4484-4495

[19] Cheng ZS, Leng SC, Zhang H, et al. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-LLMs. arXiv:2406.07476, 2024

[20] Li Junnan, Li Dongxu, Xiong Caiming, Hoi S. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 12888-12900

[21] Liu Jing, Chen, Sihan, He, Xingjian, et al. Valor: vision-audio-language omni-perception pretraining model and dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(2):708-724

[22] Yang A, Nagrani A, Seo P H, et al. Vid2seq: large-scale pretraining of a visual language model for dense video captioning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 10714-10726

[23] Li Liang, Gao Xingyu, Deng Jincan, et al. Long short-term relation transformer with global gating for video captioning. IEEE Transactions on Image Processing, 2022, 31: 2726-2738

[24] Han Tingting, Xu Yaochen, Yu Jun, et al. Action-driven semantic representation and aggregation for video captioning. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(4): 3383-3395

[25] Chen S, He X, Li H, et al. COSA: concatenated sample pretrained vision-language foundation model//Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2024: 49990--50008

[26] He Bo, Li Hengduo, YoungJang, et al. Ma-lmm: memory-augmented large multimodal model for long-term video understanding//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 13504-13514

[27] Radford A, Kim J, Hallacy C, et al. Learning transferable visual models from natural language supervision//Proceedings of the International Conference on Machine Learning. Online, 2021: 8748-8763

[28] Reed S, Lee K, et al. Training deep neural networks on noisy labels with bootstrapping//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015

[29] Wang Xinshao, Yang Hua, Kodirov E, Clifton D, Robertson N. Proselflc: progressive self label correction for training robust deep neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 752-761

[30] Li Junnan, Selvaraju R, Gotmare A, et al. Align before fuse: vision and language representation learning with momentum distillation//Proceedings of the Neural Information Processing Systems. Online, 2021: 9694-9705

[31] Papineni, Kishore, et al. Bleu: a method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA, 2002: 311-318

[32] Lin C. Rouge: ROUGE: a package for automatic evaluation of summaries//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain, 2004: 74:81

[33] Denkowski M, Lavie A. Meteor universal: language specific translation evaluation for any target language//Proceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore, USA, 2014: 376-380

[34] Vedantam R, Zitnick C, Parikh D. CIDEr: consensus-based image description evaluation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 4566-4575

[35] Zhang T, Kishore V, Wu F, Weinberger K, Artzi Y. Bertscore: evaluating text generation with bert//Proceedings of the International Conference on Learning Representations. Online, 2020

[36] Lee H, Yoon S, et al. ViLBERTScore: evaluating image caption using vision-and-language bert//Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems. Online, 2020: 34 - 39

[37] Lu Jiasen, Batra D, Parikh D, Lee Stefan. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks//Proceedings of the Neural Information Processing Systems. Vancouver, Canada, 2019: 13-23

[38] Jiang Ming, Huang Qiuyuan, Zhang Lei, et al. TIGEr: text-to-image grounding for image caption evaluation//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Hong Kong, China, 2019: 2141-2152

[39] Zheng Ervine, Yu Qi. Hierarchical multi-source uncertainty aggregation for interactive video captioning//Proceedings of the International Joint Conference on Artificial Intelligence. Vancouver, Canada, 2025: 14512-14519

[40] Liu Ze, Ning Jia., et al. Video swin transformer//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 3202-3211

[41] He Tao, Gao Lianli, Song Jingkuan, Li Yuan-Fang. Towards open-vocabulary scene graph generation with prompt-based finetuning//Proceedings of the European Conference on

Computer Vision. Tel Aviv, Israel, 2022: 56-73
[42] Deng Jia, Dong Wei, Socher R, et al. ImageNet: a large-scale hierarchical image database//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 248-255
[43] Zheng Qi, Wang Chaoyue, Tao Dacheng. Syntax-aware action targeting for video captioning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA, 2020: 13093-13102
[44] Wang Hao, Lin Guosheng, Hoi S, Miao Chunyan. Cross-modal graph with meta concepts for video captioning. IEEE Transactions on Image Processing, 2021, 31: 5150-5162
[45] Shen Yaojie, Gu Xin, Xu Kai, Fan Heng , Wen Longyin, Zhang Libo. Accurate and fast compressed video captioning// Proceedings of the IEEE International Conference on Computer Vision. Paris, France, 2023: 15512-15521
[46] Zhong Xian, Li Zipeng, Chen Shuqin, et al. Refined semantic enhancement towards frequency diffusion for video captioning// Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023: 3724-3732
[47] Gu Xin, Chen Guang, Wang Yufei, et al. Text with knowledge graph augmented transformer for video captioning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 18941-18951
[48] Liang Yuanzhi, Zhu Linchao, Wang Xiaohan, Yang Yi. Icocap: improving video captioning by compounding images. IEEE Transactions on Multimedia, 2024, 26: 4389-4400
[49] Luo HuiLan, Cai Xia, Shark L. Frame-by-frame multi-object tracking-guided video captioning. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(7), 6357-6370
[50] Zeng Pengpeng, Zhang Haonan, Gao Lianli, et al. Visual commonsense-aware representation network for video captioning. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(1): 1092-1103
[51] Ko D, Choi J, et al. MELTR: meta loss transformer for learning to fine-tune video foundation models//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 20105-20115
[52] Yan Liqi, Han Cheng, Xu Zenglin, et al. Prompt learns prompt: exploring knowledge-aware generative prompt collaboration For video captioning//Proceedings of the International Joint Conference on Artificial Intelligence. Macao, China, 2023: 1622-1630
[53] Wang Junke, Chen Dongdong, Luo Chong, et al. Omnivid: a generative framework for universal video understanding // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 18209-18220

JIANG Wen-Hui, Ph. D., associate professor. His research interests include image content understanding and cross media analysis.

GUAN Wen-Bin, master student. His research interest is image and video content understanding.

Li Hai-Jun, master student. His research interest is multimodal data analysis.

FANG Cheng-Yang, Ph. D., lecturer. His current research interests include visual question answering and multi-modal signal processing.

FANG Yu-Ming, Ph. D., professor. His current research interests include computer vision, multimedia signal processing and visual quality assessment.

ZUO Yi-Fan, Ph. D., associate professor. His current research interests include image processing and multimedia signal processing.

Background

Video captioning aims to describe the video content with accurate and fluent natural language sentences. It has a wide variety of applications such as assisting visually impaired persons, video retrieval and human-computer interaction.

The core challenge lies in effectively establishing the semantic alignment and mapping between the input video and the output textual representation. However, due to the presence of annotation noise, there often exists a degree of uncertainty in the correspondence between videos and their associated textual annotations. This misalignment hinders the model's ability to accurately learn cross-modal semantic relationships, thereby compromising the coherence and precision of the generated captions.

To address this issue, we propose a novel video captioning framework that integrates a Text-aware Video Representation (TAVR) module and a Phrase-Level Quality-aware Noise-Tolerant Training (PQNT) strategy. Specifically, the TAVR module constructs scene prototypes by clustering video frames and dynamically allocates sampling ratios based on the semantic content of the associated text. This enables the selection of key

frame sequences that are highly relevant to the text and low in redundancy, resulting in semantically rich video representations. This module effectively reduces the interference of redundant frames and enhances cross-modal alignment. Meanwhile, the PQNT module introduces a Phrase-Level Caption Evaluation (PCE) mechanism, which quantifies caption quality from both global (semantic clustering) and local (phrase-level alignment) perspectives. A quality-aware loss function is further applied to guide the model to prioritize high-quality textual annotations while mitigating the influence of noisy data.

The experimental results demonstrate that the proposed method achieves significant performance gains across multiple evaluation metrics compared with most other methods.