基于多模态记忆知识的密集视频描述方法

方豪杰[1,2]，李永刚[2,3]，曹宗瑞[1,2]，叶利华[2,3]

（1. 浙江理工大学计算机科学与技术学院（人工智能学院），浙江 杭州 310018；
2. 嘉兴大学人工智能学院，浙江 嘉兴 314001；
3. 嘉兴大学全省多模态感知与智能系统重点实验室，浙江 嘉兴 314001）

摘　要：密集视频描述旨在从未修剪的视频中定位事件，并为每个有意义的事件生成相应的描述。现有方法主要利用源视频输入来生成描述，无法捕捉到视频中的隐含知识，即视频中隐含的视觉、音频、文本等多模态记忆知识，其中多模态记忆知识可以理解为视频内对象、动作和属性对应的有意义词集合。为解决该问题，提出了基于多模态记忆知识的密集视频描述方法，不仅利用了视频本身的多模态信息，还拓展了与视频相关的多模态记忆知识，极大地提高了密集视频描述生成的准确性。首先，该方法构建了多模态记忆知识库，设计了基于模态共享编码器的事件定位模块，实现源视频多模态特征之间的深层次融合并生成高质量事件提案。然后，模型从多模态记忆知识库中检索与候选事件提案密切相关的视觉、音频和文本记忆知识作为描述生成的先验信息。最后，该方法通过记忆增强解码器，有效地整合了多模态记忆知识和视频多模态信息，生成详细的密集视频描述。在 ActivityNet Captions 和 YouCook2 数据集上进行了对比实验和消融实验，结果验证了该方法的有效性。

关键词：密集视频描述；多模态记忆知识；记忆增强解码器；交叉注意力

中图分类号：TP391

文献标志码：A

doi: 10.11959/j.issn.1000−0801.2025154

Approach of dense video captioning based on multimodal memory knowledge

FANG Haojie[1,2], LI Yonggang[2,3], CAO Zongrui[1,2], YE Lihua[2,3]

1. School of Computer Science and Technology (School of Artificial Interlligence), Zhejiang Sci-Tech University, Hangzhou 310018, China
2. College of Artificial Intelligence, Jiaxing University, Jiaxing 314001, China
3. Provincial Key Laboratory of Multimodal Perceiving and Intelligent Systems, Jiaxing University, Jiaxing 314001, China

收稿日期：2024−12−30；**修回日期：**2025−04−15

通信作者：李永刚，liyonggang@zjxu.edu.cn

基金项目：国家重点研发计划项目（No.2023YFC3305900）；浙江省自然科学基金资助项目（No.LTGG24F020001）；嘉兴市科技计划项目（No.2023AY11047，No.2023AY11030）

Foundation Items: The National Key Research and Development Program of China (No.2023YFC3305900), Zhejiang Provincial Natural Science Foundation of China (No.LTGG24F020001), Jiaxing Science and Technology Project (No.2023AY11047, No.2023AY11030)

Abstract: Dense video captioning aims to localize events in an untrimmed video and generate a corresponding captions for each meaningful event. Existing methods mainly utilize the source video input to generate captions, and these methods are unable to capture the implicit knowledge in the video, i.e., the multimodal memory knowledge such as visual, audio, text, etc., implicit in the video, where the multimodal memory knowledge can be understood as a collection of meaningful words corresponding to the objects, actions, and attributes within the video. In order to solve the problem, an approach of dense video captioning based on the multimodal memory knowledge was proposed. Not only the multimodal information of the video itself was utilized, but also the multimodal memory knowledge related to the video was expanded, by which the accuracy of dense video captioning generation was greatly improved. Firstly, a multimodal memory knowledge base was constructed, a modal sharing encoder-based event localization module was designed to achieve deep fusion between multimodal features of the source video and generate high-quality event proposals. Then, visual, audio and textual memory knowledge closely related to the candidate event proposals was retrieved from the multimodal external memory knowledge base as a priori information for caption generation. Finally, with the designed memory-enhanced decoder, the multimodal memory knowledge and video multimodal information were effectively combined to generate detailed and dense video captioning. The results of extensive comparison experiments with current mainstream algorithms on ActivityNet Captions and YouCook2 datasets as well as ablation experiments demonstrate the effectiveness of the method.

Key words: dense video captioning, multimodal memory knowledge, memory-augmented decoder, cross-attention

0 引言

随着视频理解需求的增长，视频描述领域正快速发展。视频描述任务旨在为经过人工剪辑、仅包含单一事件的视频片段生成精确的描述。然而，现实世界中的视频监控、在线视频和短视频等场景通常涉及时长较长且包含多个事件的视频片段，单一句子的描述无法提供理解视频所需的全部信息。为解决这一问题，Krishna等[1]提出了密集视频描述任务，其目标是在未剪辑的视频中定位重要事件片段的边界，并用自然语言对每个事件片段进行描述。为了实现高效的密集视频描述，正确地建模事件定位与描述生成之间的交互显得至关重要。近期，众多密集视频描述方法[2-4]被提出，旨在获得更详尽的视频描述。这些方法中的大多数[5-6]遵循“先定位后描述”的框架，其中事件定位模块通常采用锚点机制、提案头等技术从源视频中定位重要事件片段，而描述生成模块利用事件片段的视觉上下文特征依次生成描述。此外，已有研究表明，音频线索[7-8]能够显著地提升密集视频描述模型性能。

尽管目前主流密集视频描述方法取得了显著的进展，但它们主要依赖于单一视频输入的视觉上下文特征来生成密集视频描述。这种方式限制了现有密集视频描述方法仅从单一源视频中提取信息，未能充分利用同类视频资源中蕴含的丰富线索来辅助描述生成。近期研究表明，多模态记忆知识对视频理解任务具有显著益处。视频中隐含的视觉、音频和文本主题线索被视为多模态任务的必要补充。这些线索为密集视频描述生成提供额外的上下文信息，从而显著提高描述的准确性。在视频描述领域，Jing等[9]在模型训练过程中引入了可学习向量以存储视觉上下文特征用于后续的推理过程。通过设计的上下文感知交叉注意力机制，建立了源视频与其在数据集中相似视频之间的联系，以辅助生成视频描述，从而显著提升了模型性能。尽管引入可学习向量是一种简单而高效的方法，避免了设计复杂结构的需求，但它也意味着知识记忆完全依赖于训练数据集。因此，模型的泛化能力可能受到限制。在视频–

文本检索领域，Cao 等[10]不同于以往依赖成对视频–文本数据学习视频和文本表示及相互关系的方法，提出了从大量视频数据中提取视觉概念并构建可学习图结构的方式，挖掘视觉模态和文本模态之间的知识。这种方法不仅提高了视频–文本检索的效率，还增强了检索结果的准确性。Kim 等[2]的工作首次将记忆知识库的概念引入密集视频描述领域，通过设计记忆读取模块，将丰富的人工描述资源作为文本记忆知识，与密集视频描述模型相结合。这种方法通过跨模态检索相关文本线索，并结合视觉和文本交叉注意力机制，有效提升了视频中重要事件定位和描述的准确性。尽管 Kim 等[2]构建了文本领域的记忆知识并取得了一定的效果，但是他们未考虑视觉、音频的记忆知识，这些领域的记忆知识对于提升密集视频描述的准确性和丰富性同样至关重要。

因此，本文提出整合跨视频中共有的视觉、音频和文本线索作为多模态记忆知识库，以增强模型在密集视频描述任务中的描述能力。这种多模态记忆知识的补充对于提升描述的质量和准确性至关重要，基于多模态记忆知识的密集视频描述方法总体架构如图 1 所示。首先，本文从视频集合中提取潜在的视觉、音频和文本线索，并基于这些线索构建了一个多模态记忆知识库。其次，本文通过设计的事件定位模块从未经修剪的视频中定位候选事件片段，该模块利用共享模态编码器实现深层次的多模态特征融合，从而生成高质量候选事件提案。这些候选事件提案随后从多模态记忆知识库中检索匹配的记忆信息，为模型提供额外的多模态记忆知识。为了进一步优化记忆知识与源视频多模态特征的交互，本文引入了一个高效的门控网络和一个改进的记忆增强解码器结构。其中，门控网络通过动态调整多模态记忆知识的输入，有效减少冗余知识的影响；记忆增强解码器通过集成一系列注意力机制，实现了多模态记忆知识和源视频特征的有效融合，并生成准确的事件描述。通过从大量外部视频学习视觉、音频和文本多模态记忆知识，本文方法显

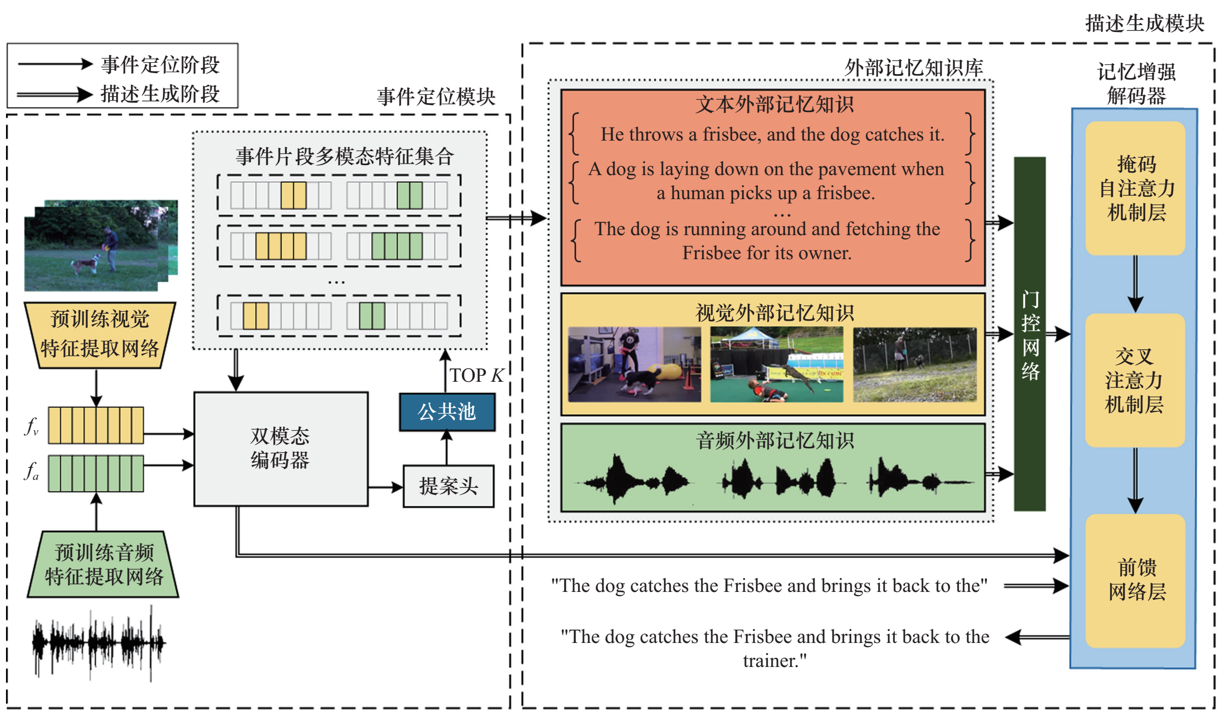


图 1　基于多模态记忆知识的密集视频描述方法总体架构

著增强了生成描述的多样性，在密集视频描述领域实现了性能的显著提升。本文主要贡献可以概括为以下几点。

（1）本文提出了一种基于多模态记忆知识的密集视频描述方法。首先，该方法从大规模视频数据中提取视觉、音频和文本记忆知识，构建多模态记忆知识库。其次，该方法通过从记忆知识库中检索与输入视频相关的多模态记忆知识，显著提升生成描述的准确性和多样性。此外，本文还引入了一个高效的门控网络动态平衡记忆知识抑制冗余，进一步优化生成描述的质量。

（2）本文设计了一个基于跨模态共享表征的事件定位器。该事件定位器构建跨模态共享表征来捕捉多模态特征的共性，并利用共享模态编码器促进跨模态共享表征和各模态特征之间的交互，从而增强多模态融合效果，生成更高质量的候选事件提案。

（3）本文在ActivityNet Captions和YouCook2数据集上进行了大量实验以验证密集视频描述任务中多模态记忆知识的有效性。与现有方法相比，本文提出的方法在上述数据集多个指标上获得了最好的性能。

1 相关工作

1.1 密集视频描述

密集视频描述任务作为视频描述领域[11-12]的一个分支，其核心需求在于检测长视频中的多个事件片段，并为每个事件生成相应的描述性句子。目前主流研究方法首先利用预训练的特征提取网络视觉特征，然后通过各自设计的方法来生成密集视频描述。这些方法可以分为两大类。第一类方法是将密集视频描述任务划分为两阶段模型，分别训练事件定位模块和描述生成模块[13-15]。第二类方法则是将事件定位模块和描述生成模块联合训练，端到端地实现视频中每个事件的定位和描述生成[16-19]。Krishna等[1]首次提出密集视频描述方法，该方法包含多尺度事件定位模块和基于注意力机制的长短期记忆（long short-term memory，LSTM）网络描述生成模块。Zhou等[20]进一步提出了首个端到端的密集视频描述方法。该方法包含3个主要模块：视频编码器、提案解码器和描述解码器。视频编码器和提案解码器的输出被用于指导描述解码器，并辅助整个模型的训练。虽然这一模型为后续的端到端研究开辟了新的方向，但是该模型的视频编码器和提案解码器需协同指导描述解码器，测试阶段需大量时间生成事件提案，导致运行效率低下，严重限制了该方法在现实应用场景中的可行性。Wang等[16]提出了PDVC框架，该框架通过在Transformer解码器中集成多个任务头来实现事件定位和描述生成并行处理，PDVC还引入了事件计数器，使得模型能够产生适当大小的事件集，从而在性能上有了显著提高。由于密集视频描述数据集的局限性，两阶段模型因其较高的归纳能力而成为实现高性能密集视频描述生成的更优选择。目前大多数模型采用两阶段训练方法。同时，端到端训练的模型在自然可解释性方面存在不足[5]。因此，本文选择两阶段模型BMT[21]作为本文研究的基线模型。

1.2 多模态密集视频描述

目前主流的密集视频描述研究方法主要依赖于单模态特征的提取和应用，少部分研究开始探索多模态特征融合（如结合音频、文本特征），以提升密集视频描述的全面性和准确性。Rahman等[8]提出了一个多模态特征交互模块，该模块旨在整合音频、视频特征，为视频事件提案生成更为丰富的描述。由于缺乏视频事件片段和描述之间准确的时序对齐信息，模型难以准确学习描述与视频内容之间匹配关系，极大地限制模型描述性能。鉴于Transformer架构在视觉领域的卓越表现，Iashin等[21]创新性地将其应用于密集视频描述生成任务，通过结合视觉和音频特征来生

成文本描述，这标志着Transformer架构在视频理解和多模态分析领域的新应用。尽管多模态密集视频描述取得了显著进展，现有方法在多模态特征的处理仍存在局限性，大多数方法依赖于简单的特征拼接操作，未能构建有效的模态信息融合机制。近期，在其他多模态领域的研究已经证实，构建跨模态共享表征对提升模型性能具有显著贡献[22-24]。基于这一发现，本文构建了跨模态共享表征捕捉多模态特征共享，并利用设计的共享模态编码器促进跨模态共享表征和各模态特征之间的交互。这种设计有效地提升了多模态特征融合效果，从而生成更准确、更丰富的描述。

1.3　基于记忆知识的学习

为了提升描述生成准确性，图像、视频和人工描述中隐含线索的理解与利用已引起学术界的高度关注。在视频描述任务中，You等[25]从训练数据中选择一组固定的概念作为记忆知识，并设计了专用于概念检测的神经网络，通过使用检测到的概念信息结合语义注意力机制指导最终描述生成。在视觉推理领域任务中，Zhen等[26]采用视觉感知模型，融合候选视觉概念与实际场景，并利用预训练的大型语言模型自适应识别关键视觉概念完成后续工作。在密集视频描述任务中，Kim等[2]首次将记忆知识的概念引入密集视频描述领域，通过设计的记忆读取模块，将丰富的人工描述资源作为语言记忆知识。这种方法通过跨模态检索相关文本线索，并结合视觉、文本交叉注意力机制，有效提升了事件定位和事件描述的准确性。然而，该方法未能充分考虑视觉、音频领域的记忆知识对密集视频描述方法的帮助。这些领域的记忆知识对于增强视频描述的准确性和丰富性具有不可忽视的作用。因此，本文整合了大规模视频数据中共有的视觉、音频和文本线索为多模态记忆知识库。通过检索与输入视频相关的记忆知识辅助，本文方法使模型能够生成更准确、更丰富的密集视频描述。

2　基于多模态记忆知识的密集视频描述方法

本文提出的方法包含两个关键组成部分：事件定位模块和描述生成模块。该方法旨在依托视觉特征、音频特征以及多模态记忆知识来指导密集视频描述生成。接下来，本文将详细介绍事件定位模块的设计、多模态记忆知识库的构建、多模态记忆知识的检索以及事件描述模块的设计。此外，本文采用预训练的CLIP Vit-L/14网络[27]提取视频特征f_v和文本特征c，并采用VGGish[28]来提取音频特征f_a。其中，CLIP指采用对比语言–图像预训练（contrastive language-image pretraining）方法，Vit指图像编码器基于视觉变换器架构（vision transformer），14表示输入图像分割为14×14像素的patch大小。

2.1　事件定位模块

事件定位模块旨在定位视频中所有重要事件，生成一系列事件提案，并从中筛选出置信度最高的100个提案作为候选事件提案。具体而言，首先，事件定位模块对先前提取的音频和视觉特征进行编码，然后，通过多模态编码器实现深层次多模态特征融合。事件定位器架构如图2所示，该编码器由n个编码层组成，每个编码层包含2个双模态编码器和1个共享模态编码器。双模态编码器由2个自注意力层和2个交叉注意力层组成。其中，自注意力机制引导模型聚焦于当前模态特征中关键语义信息，交叉注意力机制促进不同模态特征之间的初步交互和对齐。具体工作流程定义如下：

$$A^{\mathrm{self}}=\mathrm{MultiHeadAttention}(Q,K,V) \tag{1}$$

$$V^{\mathrm{self}}=\mathrm{MultiHeadAttention}(Q,K,V) \tag{2}$$

$$V=\mathrm{MultiHeadAttention}(V^{\mathrm{self}},A^{\mathrm{self}},A^{\mathrm{self}}) \tag{3}$$

$$A=\mathrm{MultiHeadAttention}(A^{\mathrm{self}},V^{\mathrm{self}},V^{\mathrm{self}}) \tag{4}$$

其中，Q、K和V分别代表查询（queries）、键

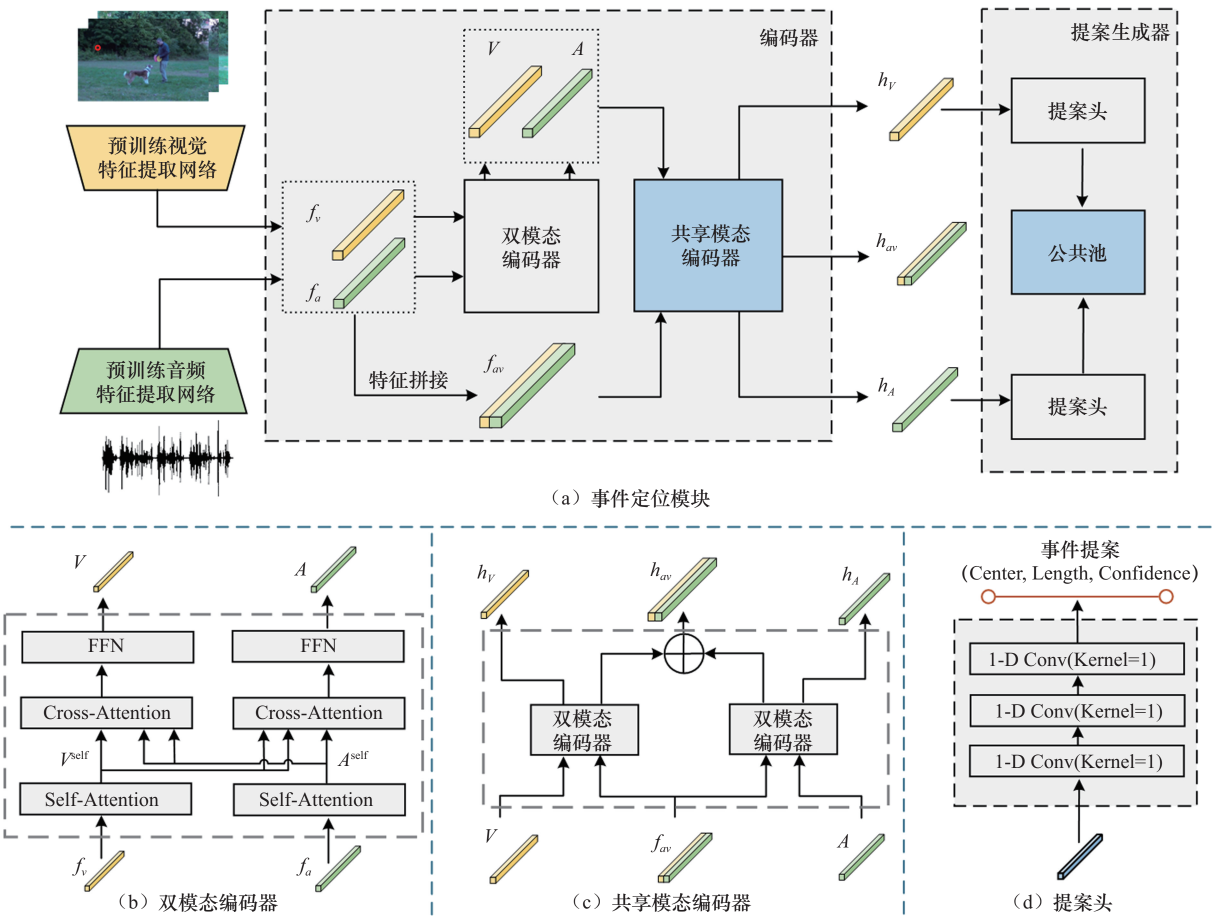


图2　事件定位器架构

(keys) 和值 (values) 的序列，MultiHead Attention() 代表多头注意力机制，A^{self}、V^{self}分别为音频特征和视觉特征自注意力机制交互结果。

此外，为了充分利用跨模态共享表征在多模态任务中的优势，首先进行跨模态共享表征构建。然后，本文将跨模态共享表征与上文得到的特征A、V分别输入共享模态编码器，以进一步增强多模态特征融合。共享模态编码器输出结果经过n层编码器处理并保留中间结果h_v和h_a。其中，跨模态共享表征学习视觉、音频模态之间的共性，捕获它们之间共同存在的语义信息。而单模态特征保留了各模态特性。通过交叉注意力机制，本文方法使跨模态共享表征与各模态特征进行充分交互，实现了对视频多模态特征共性与特性的有效整合，显著提升了多模态特征的融合效果。共享模态编码器的过程定义如下。

$$f_{av}=\mathrm{Concat}(f_a,f_v) \tag{5}$$

$$h_m=\mathrm{MultiHeadAttention}(m,f_{av},f_{av}) \tag{6}$$

$$h_{av}=\mathrm{LN}\left(\sum\nolimits_{m\in\{A,V\}}\mathrm{MultiHeadAttention}(f_{av},m,m)\right) \tag{7}$$

其中，Concat() 表示特征拼接，LN() 表示归一化层，$m\in\{A,V\}$，h_A和h_V分别表示音频特征、视觉特征与跨模态共享表征的交互结果。

最后，本文方法将中间结果输入前馈网络 (feed-forward network，FFN) 获得最终编码输出。具体过程定义如下：

$$f_V=\mathrm{FullyConnected}(h_v) \tag{8}$$

$$f_A=\mathrm{FullyConnected}(h_a) \tag{9}$$

其中，FullyConnected()代表全连接层。

随后，将经过充分多模态特征融合得到的特征f_V和f_A分别送入由一系列提案头组成的提案生成器中。每一个提案头均由3层全卷积网络构成。其中第一层卷积层的核大小为k，而第二层和第三层则采用卷积核大小为1的卷积。通过一系列提案头，事件定位模块能够在区间[1, T]内对每个时间戳进行预测，识别出不同尺度的事件片段。最终，计算得出事件提案的时间边界和置信度，包括片段中心（Center）、片段长度（Length）和置信度得分（Confidence）。具体过程定义如下：

$$\text{Center}=p+\sigma(c) \tag{10}$$

$$\text{Length}=\text{anchor}+\exp(l) \tag{11}$$

$$\text{Confidence}=\sigma(o) \tag{12}$$

其中，exp()是缩放因子，p、anchor由每一个提案头初始化确定，$\sigma()$是一个sigmoid函数。c、l、o是提案头预测的3个值。

然后，生成的提案集合输入公共池，并依据置信度得分从公共池中保留得分最高的100个事件提案。相较于传统的视频描述任务，密集视频描述需要为每个事件提案生成描述。对于最终选定的提案集合E中的每个事件提案e，描述生成器的任务是学习如何将每个提案e转化为详细的视频描述。对于每一个事件提案e，首先获取该事件提案的起始与结束时间，然后从完整的视频特征中提取相应事件片段的视觉特征f_V^{evt}和音频特征f_A^{evt}。具体过程定义如下：

$$\text{Start}=\text{Center}-\frac{\text{length}}{2} \tag{13}$$

$$\text{End}=\text{Center}+\frac{\text{length}}{2} \tag{14}$$

$$f_m^{\text{evt}}=(f_m^{\text{Start}}:f_m^{\text{End}}) \tag{15}$$

其中，Start和End为对应多模态特征开始位置和结束位置的索引，$m\epsilon\{A,V\}$,f_V^{evt}、f_A^{evt}分别表示当前事件片段的视觉特征和音频特征。

最后，记忆检索网络和编码器分别接收当前事件片段的视觉特征f_V^{evt}、音频特征f_A^{evt}并进行记忆知识匹配和多模态特征融合。

2.2　多模态记忆知识库构建

2.2.1　视觉、音频记忆知识库构建

目前主流的密集视频描述方法主要依赖于输入视频的多模态特征来生成描述，这导致生成的描述缺乏丰富的语义信息。一个直接的解决方案是引入相似视频作为记忆知识，以补充源视频信息的不足。然而，将源视频和大量候选视频进行关系建模将显著增加模型的计算成本和时间开销。借鉴其他多模态语言生成任务的经验，图像和视频中隐含的线索能够辅助模型生成高质量描述。因此，本文通过无监督的学习方式从大量视频中提取隐含的视觉、音频线索，构建视觉、音频记忆知识库，并利用记忆知识作为描述生成的先验知识，极大增强模型生成描述的准确性。

具体而言，首先，本文收集了ActivityNet Captions和YouCook2数据集中的所有视频，并采用CLIP Vit-L/14和VGGish分别提取这些视频的视觉特征$\{V_i\}_{i=1}^{O}$和音频特征$\{A_i\}_{i=1}^{P}$，其中O和P分别表示视频集合总视觉特征长度和总音频特征长度，V_i和A_i分别表示第i帧的视觉特征和第i帧的音频特征。基于此过程，获得了2个数据集合中所有视频的多模态特征。视觉音频记忆知识库构建流程如图3所示，利用K-means算法对这些视觉特征和音频特征分别进行聚类处理，最终得到M个视觉聚类中心和N个音频聚类中心，记作$f_v^m=\{f_1^v,f_2^v,\cdots,f_M^v\}$，$f_a^m=\{f_1^a,f_2^a,\cdots,f_N^a\}$，其中$f_L^v$、$f_L^a$分别表示第$L$个视觉聚类中心和第$L$个音频聚类中心。其中，每一个聚类中心可以理解为一个记忆知识。本文将f_a^m和f_v^m定义为音频和视觉记忆知识库，它们将在记忆增强解码器中辅助源视频生成更为详尽的描述。

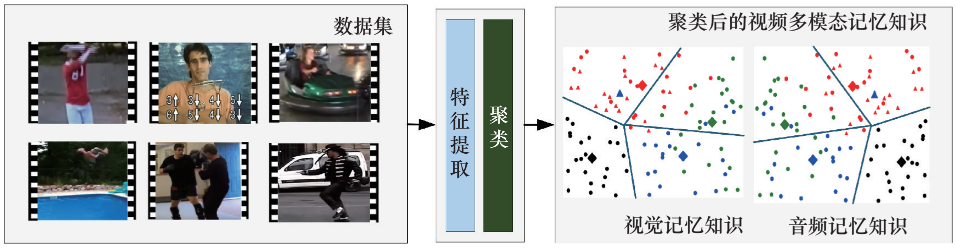


图3 视觉音频记忆知识库构建流程

2.2.2 文本记忆知识库构建

为了构建包含丰富语义信息的文本记忆知识库，本文同样收集了ActivityNet Captions和YouCook2数据集中所有的人工描述作为文本记忆知识库，并参考Kim等[2]的做法，选择以CLIP Vit-L/14作为文本编码器[27]。得益于其在预训练阶段所学习到的丰富视觉和语言特征表示，CLIP Vit-L/14展现了卓越的特征对齐能力，为后续记忆知识准确匹配奠定基础。最终，本文将获得的所有文本嵌入整合为一个文本记忆知识库f_t^m。

2.3 记忆知识匹配

记忆知识匹配的目的在于从音频、视觉和文本多模态记忆知识库中提取与视频相关的记忆知识$f_x^m, x\epsilon\{a,v,t\}$。记忆知识匹配流程如图4所示，该过程主要通过计算特征$\bar{f}_V^{\mathrm{evt}}$，$\bar{f}_A^{\mathrm{evt}}$与记忆知识库特征f_a^m、f_v^m、f_t^m之间的相似性得分来实现。具体过程定义如下：

$$\bar{f}_m^{\mathrm{evt}} = \mathrm{meanpool}(f_m^{\mathrm{evt}}) \tag{16}$$

$$d(\bar{f}_A^{\mathrm{evt}}, f_a^m) = \frac{\bar{f}_A^{\mathrm{evt}} \cdot f_a^m}{\|\bar{f}_A^{\mathrm{evt}}\| \|f_a^m\|} \tag{17}$$

$$d(\bar{f}_V^{\mathrm{evt}}, f_v^m) = \frac{\bar{f}_V^{\mathrm{evt}} \cdot f_v^m}{\|\bar{f}_V^{\mathrm{evt}}\| \|f_v^m\|} \tag{18}$$

$$d(\bar{f}_V^{\mathrm{evt}}, f_t^m) = \frac{\bar{f}_V^{\mathrm{evt}} \cdot f_t^m}{\|\bar{f}_V^{\mathrm{evt}}\| \|f_t^m\|} \tag{19}$$

其中，$\cdot$为点积运算符，meanpool()为平均池化操作，$d()$为相似性得分计算操作，$m\epsilon(A,V)$。

然后，本文方法根据所有匹配对的相似性得分进行降序排列并筛选出每个事件的多模态特征中匹配度最高的N个记忆知识，表示为$r_m^M = \{r_m^1, r_m^2, \cdots, r_m^N\}, m\epsilon\{v,a,t\}$，其中$N$代表匹配的第$N$个记忆知识。随后，将这些记忆知识$r_m^M, m\epsilon\{v,a,t\}$分别送入编码器层，并进行简单的自注意力交

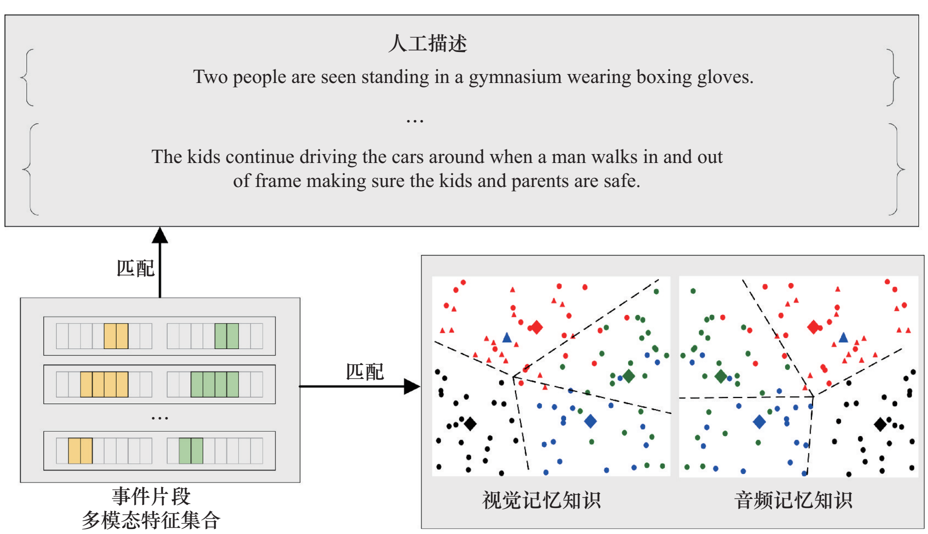


图4 记忆知识匹配流程

互，以得到进一步处理后的记忆知识表示 $\tilde{r}_m^M, m\epsilon\{v,a,t\}$。考虑后续工作的简化实现，本文方法将最终的多模态记忆知识定义为 $\tilde{r}^M=[\tilde{r}_v^M;\tilde{r}_a^M;\tilde{r}_t^M]$，其中[;]为拼接操作。

除了上述模块，本文还引入了一个高效的门控网络，用于动态平衡多模态记忆知识的信息流。该门控网络接收多模态记忆知识为输入，通过线性变化层和sigmoid激活函数为每个多模态记忆知识计算权重得分，从而实现对多模态记忆知识流的控制。这种设计使模型能够自适应整合与当前视频最相关的记忆知识，具体过程定义如下：

$$g(x)=\sigma(\mathrm{Linear}(x)) \tag{20}$$

$$\tilde{r}^{\mathrm{gated}}=g(\tilde{r}^M)\cdot\tilde{r}^M \tag{21}$$

其中，Linear()为线性层。

2.4　记忆增强解码器

记忆增强解码器旨在利用模型先前生成的单词嵌入表示 C、双模态编码器输出的特征向量 f_V^{evt}、f_A^{evt} 以及多模态记忆知识 $\tilde{r}^{\mathrm{gated}}$ 来生成事件描述。记忆增强解码器如图5所示，记忆增强解码器接收上述特征向量并进行后续处理。在每一个解码器层中，单词嵌入 C 首先通过掩码自注意力层计算当前句子单词嵌入间的相关性，然后通过交叉注意力层分别与多模态特征 f_V^{evt}、f_A^{evt} 和多模态记忆知识 $\tilde{r}^{\mathrm{gated}}$ 进行交互。随后，将处理结果送入描述生成器以生成最终的描述。字幕生成器负责为下一个单词的分布建模并将单词嵌入映射到与词汇表相对应的维度中。该模块由一个全连接层和激活函数组成。具体过程定义如下：

$$C^{\mathrm{self}}=\mathrm{MultiHeadAttention}(Q,K,V) \tag{22}$$

$$C_m=\mathrm{MultiHeadAttention}(C^{\mathrm{self}},m,m),\ m\epsilon\{f_V^{\mathrm{evt}},f_A^{\mathrm{evt}},\tilde{r}^{\mathrm{gated}}\} \tag{23}$$

$$C_M=[C_{f_V^{\mathrm{evt}}};C_{f_A^{\mathrm{evt}}};C_{\tilde{r}^{\mathrm{gated}}}] \tag{24}$$

$$C=\mathrm{softmax}(\mathrm{FFN}(C_M)) \tag{25}$$

其中，softmax()为激活函数，FFN()为全连接层。

3　实验及结果分析

本文在ActivityNet Captions[29]和YouCook2[30]数据集上进行了大量实验评估，验证了所提模型的性能表现。

3.1　数据集

ActivityNet Captions数据集[29]涵盖了20 000个YouTube视频，总计849个小时的视频内容，包含100 000个视频片段，平均每个视频片段提供3.65个带有时间戳的人工描述。该数据集被划分为训练集（50%）、验证集（25%）和测试集（25%）。鉴于测试集不提供人工描述，所有实验评估结果均在验证集上获得。

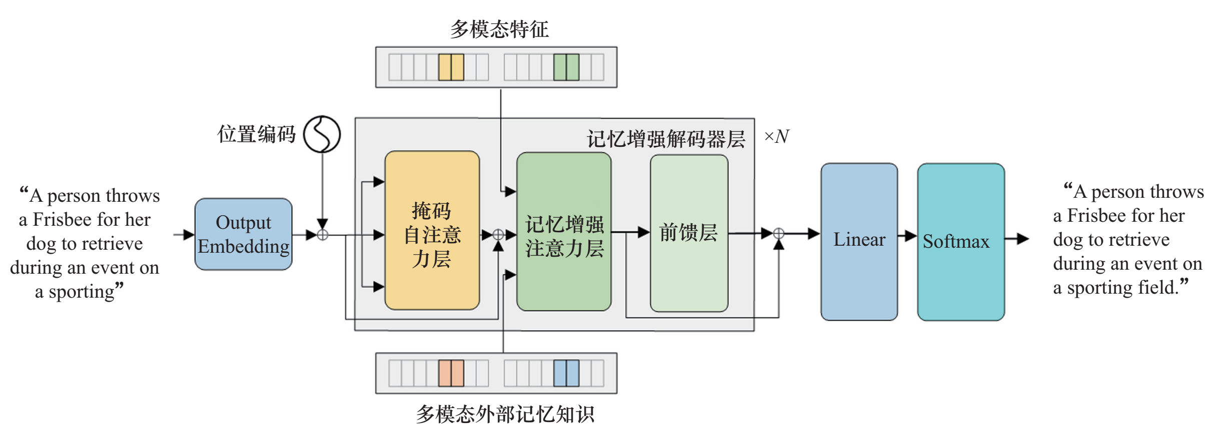


图5　记忆增强解码器

YouCook2数据集[30]包含约2 000个未经剪辑的烹饪教学视频。与ActivityNet Captions数据集相似，所有实验评估结果在验证集上获得。YouCook2数据集共包含1 333个训练视频和449个验证视频。这些视频时长平均为5.3 min，每个视频平均包含7.8个事件。

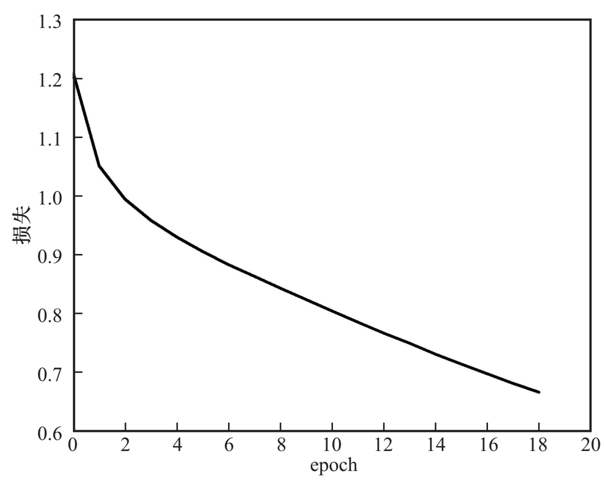
3.2 评估指标

本文采用官方提供的评估工具来衡量模型在事件定位和描述生成上的性能。具体而言，对于事件定位任务，本文在不同阈值{0.3，0.5，0.7，0.9}下评估预测事件提案的召回率（Rec）、准确率（Pre）和F1分数（F1）的平均值；对于描述生成任务，本文基于生成的描述计算METEOR[31]（M）、BLEU{3，4}[32]（B3,B4）和CIDEr（C）指标。与现有方式[16,21]一样，本文使用METEOR评分作为主要的性能评估指标。

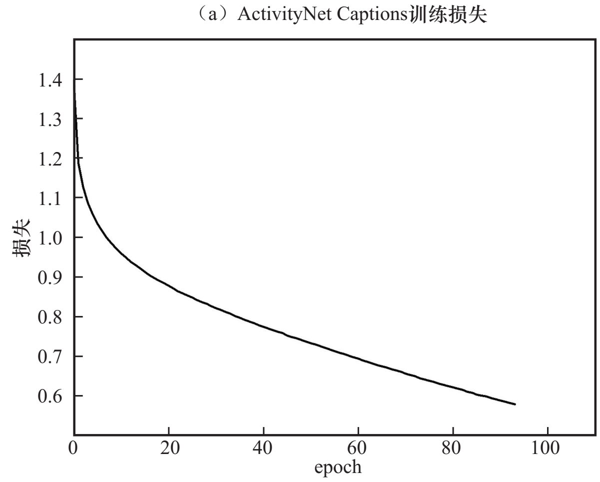
3.3 实验设置

在实验过程中，本文提取了维度分别为128的音频特征和768的视觉特征，并使用*K*-means算法为音频特征和视觉特征分别生成48个锚点，用于后续生成事件提案。训练事件提案模型时，批次大小设置为16。为了确保提案头的有效运行，本文将音频特征和视觉特征序列填充至800。对于获取的事件提案，仅保留置信度得分最高的100个事件提案。在构建多模态记忆知识库时，本文同样使用*K*-means算法生成500个视觉锚点和1 000个音频锚点。在知识检索过程中时，模型筛选出相似度得分最高的150个视觉、音频和文本记忆知识，为描述生成提供多模态信息。对于描述生成器，序列长度将填充至当前批次中最长序列的长度。多模态融合编码器和描述解码器块的数量设置为2，每层包含4个多头注意力机制。词汇表的构建基于训练集中出现过的所有单词，并将描述的最大长度限制为30，对超出此长度的人工描述进行截断处理。本文采用Adam优化器来优化网络参数，并将学习率设定为5×10^{-5}。

此外，本文还分析了描述生成器在两个不同数据集上的训练情况和验证阶段的性能表现。事件描述器在ActivityNet Captions和YouCook2数据集上的训练损失如图6所示，展示了描述生成器在训练阶段的损失情况，其中描述生成器在经过92次和19次迭代后，损失值持续稳定下降。事件描述器在ActivityNet Captions和YouCook2验证集上的性能曲线如图7所示，展示了描述生成器在验证阶段的性能曲线。所有实验均在配备4个NVIDIA GeForce RTX 4090的服务器上完成。

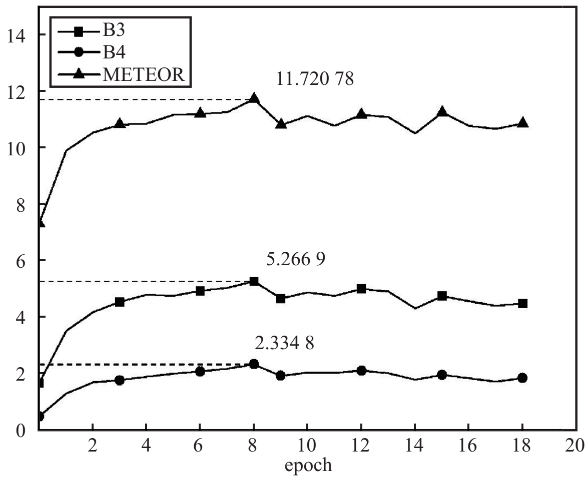


（a）ActivityNet Captions训练损失

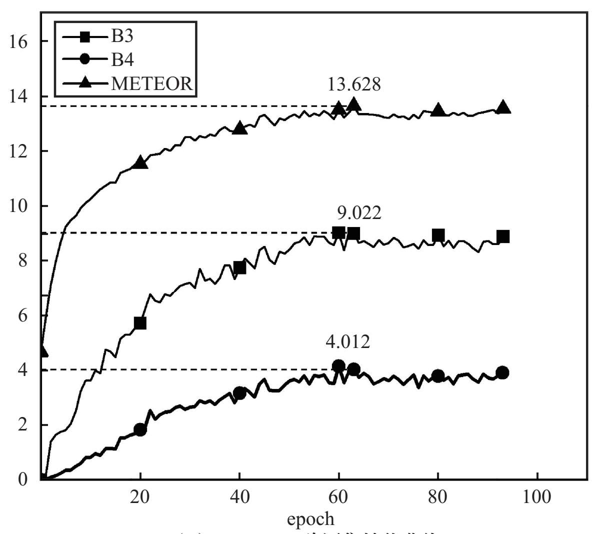


（b）YouCook2训练损失

图6 事件描述器在ActivityNet Captions和YouCook2数据集上的训练损失



（a）ActivityNet Captions验证集性能曲线



（b）YouCook2验证集性能曲线

图7　事件描述器在 ActivityNet Captions 和 YouCook2 验证集上的性能曲线

3.4　性能比较

基于 ActivityNet Captions 数据集的各方法定位性能比较见表 1，展示了各方法在 ActivityNet captions 验证集上的召回率、精确度和 F1 分数等事件检测性能指标，本文在 F1 分数上取得了最佳性能。PDVC[16]通过其设计的事件计数器，自适应保留了事件提案数量的同时实现了较好的事件定位性能。Vid2Seq[33]与 DIBS[3]通过引入额外设计的模块，并利用大量无人工描述的视频进行训练，实现了性能指标的均衡，但由于没有精确的时序先验知识，其定位性能相对较弱。与 BMT[21]相似，本文通过设计的提案生成器，保留了更多的事件提案，更加符合密集描述需求的同时，在事件定位能力表现更佳。此外，本文方法在 BMT[21]的基础上，通过构建跨模态共享表示，更有效地促进了多模态信息融合，从而进一步提升了事件定位的能力。

表 1　基于 ActivityNet Captions 数据集的各方法定位性能比较

对比项	ActivityNet Captions		
	Pre	Rec	F1
BMT[21]	48.23	**80.31**	60.27
PDVC[16]	58.07	55.42	59.64
Vid2Seq[33]	52.70	53.90	52.40
UEDVC[34]	**60.32**	59.00	59.65
Streaming GIT[35]	—	—	50.90
Streaming Vid2Seq[35]	—	—	52.90
DIBS[3]	58.39	53.02	55.57
CM^2[2]	56.81	53.71	55.21
本文模型	52.60	74.15	**61.55**

基于真实事件片段时间戳生成的描述（With GT Proposals）和基于事件提案时间戳生成的描述（With LR Proposals）是评价事件描述性能的常用方法。本文将提出的模型在两种条件下分别与其他方法进行了比较。基于 ActivityNet Captions 数据集的模型性能比较见表 2，可以观察到在两种不同条件下，本文 B3 和 METEOR 性能得分取得了最佳结果，证实了多模态记忆知识库的有效性。在与基线模型 BMT 的对比中，本文提出的方法在 B3 指标上实现了 13.8% 的提升，在 METEOR 评价指标上实现了 7.5% 的提升。通过构建多模态记忆知识库，并检索相关主题的视觉、音频和文本线索辅助描述生成，本文模型实现了各项指标的显著提升。在 With LR Proposals 条件下，本文模型同样实现了显著的性能提升，特别是在 METEOR 指标上，与单独引入文本记

忆知识的CM^2比较得到了3.4%的提升。CM^2模型首次将记忆知识的概念应用于密集视频描述领域，通过其设计的记忆读取模块，有效地整合文本记忆知识，相较于基线模型PDVC，同样在各项评价指标上均实现了显著提升，进一步证实了记忆知识对描述生成的帮助。为了保留视频描述提案的多样性，本文在事件定位阶段保留了较多的候选事件提案，这导致部分高置信度事件出现重复描述。这种冗余不利于重视描述全面性的CIDEr指标，从而导致模型在CIDEr指标上表现相对较弱。

表2 基于ActivityNet Captions数据集的模型性能比较

对比项	With GT proposals				With LR proposals			
	B3	B4	C	M	B3	B4	C	M
MDVC[7]	4.52	1.98	—	11.07	2.53	1.01	—	7.38
CMETN[14]	4.70	2.04	—	11.06	3.50	1.64	—	8.58
MSFTN[15]	4.52	1.91	—	10.93	3.47	1.64	—	8.69
PDVC[16]	—	**3.07**	**52.53**	11.27	—	1.78	28.96	7.96
BMT[21]	4.63	1.99	42.00	10.90	3.84	1.88	11.52	8.44
CM[36]	4.69	2.19	—	11.08	3.98	1.84	—	8.93
EMEA[37]	4.43	1.94	—	10.55	3.75	1.87	—	8.33
DIBS[3]	—	—	—	—	—	—	31.89	8.93
CM^2[2]	—	—	—	—	—	**2.38**	**33.01**	8.55
本文模型	**5.27**	2.34	45.56	**11.72**	**4.24**	2.07	13.03	**9.24**

基于YouCook2数据集的模型性能比较见表3，展示了本文提出的方法在YouCook2数据集上事件定位和描述生成的性能表现。本文方法在事件定位模块中实现了显著的性能提升，并且在描述生成方面也展现竞争优势。具体来说，在衡量模型事件定位能力的F1指标上，本文模型相较基线模型BMT和仅引入文本记忆知识的CM^2分别得到了8.5%和2.7%的提升。在METEOR指标上，本文相较基线模型BMT取得了29.2%的大幅提升，而与CM^2模型相比，本文取得了8.3%的性能提升。与在ActivityNet Captions数据集上的结果相似，本文为保留描述事件的多样性，在CIDEr指标上的表现并不突出。

表3 基于YouCook2数据集的模型性能比较

对比项	With LR proposals				
	F1	Pre	Rec	C	M
BMT[21]	26.90	—	—	8.20	5.10
MT[20]	—	—	—	9.30	5.00
E2ESG[38]	—	—	—	25.00	3.50
GIT[39]	17.70	—	—	12.10	3.40
PDVC[16]	26.81	32.37	22.89	29.69	5.56
CM^2[2]	28.43	33.38	**24.76**	**31.66**	6.08
本文模型	**29.20**	**37.75**	23.79	15.56	**6.59**

为了进一步研究CIDEr指标表现相对较弱的原因，本文还对不同的提案生成方式进行了讨论，基于不同提案生成方式的模型性能比较见表4。从表4中可以看到目前主流密集视频描述模型的4种提案生成器架构，即Anchor-based、Anchor-free、Pointer Network以及Event counter。在With GT proposals条件下，本文CIDEr得分与其他方法性能相当。而在With LR proposals条件下，基于Anchor-based和Anchor-free的提案生成器架构CIDEr得分均表现不佳。本文对此现象的理解是，大多数基于Anchor-based或Anchor-free的提案生成方法采用保留置信度得分最高的100个或更多的事件提案的方法，这会产生大量冗余的事件提案，生成重复或泛化的描述，导致更关注描述多样性的CIDEr指标得分下降。基于Pointer Network的方法通过设计的模块自适应确定提案的保留数量与顺序，而基于Event counter的方法则通过Event counter模块预测事件数量，两者均显著减少了冗余提案，CIDEr指标依旧表现良好。值得注意的是，基于Pointer Network的方法在工作前期同样产生了大量的事件提案，为本文后续改进提供了研究方向。

表4　基于不同提案生成方式的模型性能比较

对比项	提案生成器架构	With GT proposals		With LR proposals	
		C	M	C	M
BMT[21]	Anchor-based	42.00	10.90	11.52	8.44
MT[20]	Anchor-based	47.71	11.16	9.25	4.98
MSFTN[15]	Anchor-free	38.35	10.96	7.65	8.16
CMETN[14]	Anchor-free	41.04	10.77	8.47	8.77
SDVC[4]	Pointer Network	43.38	**13.07**	**30.68**	8.82
PDVC[16]	Event counter	**52.53**	11.27	28.96	7.96
MPP-Net[40]	Event counter	50.07	10.59	29.76	7.61
本文模型	Anchor-based	45.56	11.72	13.03	**9.24**

3.5　消融实验

3.5.1　多模态记忆知识库对模型性能的影响

本文对提出的多模态记忆知识库模块进行了消融实验，旨在比较多模态记忆知识库基于真实事件片段生成的描述和基于事件提案生成的描述条件下对性能的影响。各模块性能见表5，分别比较了6种不同模型的变体（1）~（6），包括：基线模型，基线模型分别配置视觉、音频和文本记忆知识库，基线模型同时配置视觉、音频和文本记忆知识库，以及基线模型同时配置音频、视觉、文本记忆库和门控网络的模型。具体设置如下。

（1）Baseline表示使用基线模型。

（2）Baseline+Text表示使用基线模型配置文本记忆知识库。

（3）Baseline+Visual表示使用基线模型配置视觉记忆知识库。

（4）Baseline+Audio表示使用基线模型配置音频记忆知识库。

（5）Baseline+All表示使用基线模型同时配置视觉、音频及文本记忆知识库。

（6）Baseline+All+Gate表示使用基线模型同时配置视觉、音频及文本记忆知识库和门控网络。

从表5中可以看到，每个模块的加入都显著提升了模型在METEOR、BLEU{3，4}以及CIDEr指标上的表现。这一结果证实了本文所提方法的有效性，表明了记忆知识可以为源视频视觉、音频和文本表示提供有益的补充知识。同时，基线模型在同时配置视觉、音频以及文本记忆知识库的性能得分也较配置单模态记忆知识库的模型变体的得分有较大提升，这表明了不同模态的记忆知识能够协同作用于描述生成过程。然而，在记忆知识匹配过程中，模型获取了大量与源视频相关的记忆知识，但也不可避免地引入了冗余、无关的信息。因此，有必要设计门控网络用于限制多模态记忆的影响，提高匹配知识的可信度。通过比较变体（5）和变体（6）的实验结果，可以证明，门控网络能够增加匹配的记忆知识的可信度。通过该门控网络对多模态信息的调节，模型在B3、B4和METEOR指标上分别实现了3%、6%和1.8%的性能提升。

本文分析了引入多模态记忆知识库和门控网络模块后模型参数量和推理速度的变化，模型计

表5　各模块性能

对比项	With GT proposals				With LR proposals			
	B3	B4	C	M	B3	B4	C	M
Baseline	4.87	2.06	43.69	11.18	4.06	1.90	13.09	8.91
Baseline+Text	5.15	2.18	43.62	11.48	4.12	1.91	12.59	9.02
Baseline+Visual	5.03	2.16	**47.09**	11.27	4.18	1.97	**13.59**	8.95
Baseline+Audio	4.94	2.14	44.40	11.32	4.08	1.91	12.74	9.01
Baseline+All	5.11	2.20	45.27	11.51	4.22	2.00	13.03	9.08
Baseline+All+Gate	**5.27**	**2.34**	45.56	**11.72**	**4.24**	**2.07**	13.03	**9.24**

算效率性能见表6。其中，Params表示模型参数量，Inference Speed为相同硬件条件下模型对单视频的平均推理速度。本文提出的方法通过引入多模态记忆知识，并为每种模态特征设计额外的编码器，导致了模型参数量的增加。此外，模型在生成描述时需与记忆知识库进行匹配，推理速度受提案数量和硬件读写速度的影响，导致模型推理速度的下降。然而，引入多模态记忆知识库和门控网络模块也使模型METEOR指标实现了4.8%的性能提升。未来工作需进一步聚焦冗余提案的优化与去除以及编码器的共享设计来提升模型的运行效率。

表6 模型计算效率性能

对比项	Params（Mill）	Inference Speed（sec/video）	With GT proposals
			M
Baseline	**50.5**	**0.55**	11.18
Baseline+All+Gate	86.5	0.87	**11.72**

3.5.2 多模态信息对模型性能的影响

本文探讨了不同模态特征对模型性能的影响并进行了消融实验，且进一步探讨了多模态特征融合策略的作用和效果。本文分别对比了单一模态特征生成描述、基于多模态特征生成描述，以及构建跨模态共享表示并利用共享模态编码器促进多模态融合并生成描述这3种方式下的模型性能，各模态及融合方式作用见表7。本文同样采用了直方图来直观展示各个模块的性能表现。各模态性能可视化如图8所示，仅使用单一模态特征进行描述生成的效果弱于基于多模态特征生成描述的方法，这也进一步证明了音频线索在密集视频描述任务的重要性。同时，构建跨模态共享表示并使用共享模态编码器生成描述的方法相较于对多模态特征进行简单拼接来融合多模态信息的方法能够取得更好的效果，这表明了构建跨模态共享表示对于辅助描述生成具有积极作用。

表7 各模态及融合方式作用

音频	视觉	共享模态编码器	With GT proposals			
			B3	B4	C	M
√			2.34	1.02	8.61	6.55
	√		4.59	1.95	41.83	11.18
√	√		4.87	2.06	43.69	11.18
√	√	√	**5.11**	**2.20**	**45.27**	**11.51**

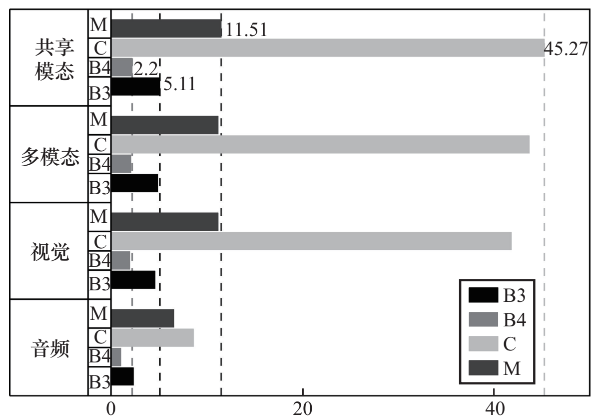


图8 各模态性能可视化

3.5.3 各模态记忆知识匹配数量对模型性能的影响

为了探究匹配不同数量的视觉、音频和文本记忆知识对模型描述生成训练和学习的影响，本文设定了m_v、m_a和$m_t \in [50, 100, 150, 200]$的范围，其中$m_v$、$m_a$和$m_t$分别表示匹配记忆知识的数量。具体来说，本文对比了仅集成单独模态记忆知识的3个变体以及集成所有模态记忆知识的模型，并在ActivityNet Captions数据集上针对模型性能的METEOR指标进行了实验比较。各模态记忆知识匹配数量对模型METEOR性能的影响如图9所示。结果表明，适度增加匹配的记忆知识数量能够提升模型描述生成的性能。然而，进一步增加匹配记忆知识数量时会导致模型性能下降，这可能是过多的记忆知识数量引起了冗余噪声，削弱了模型性能。

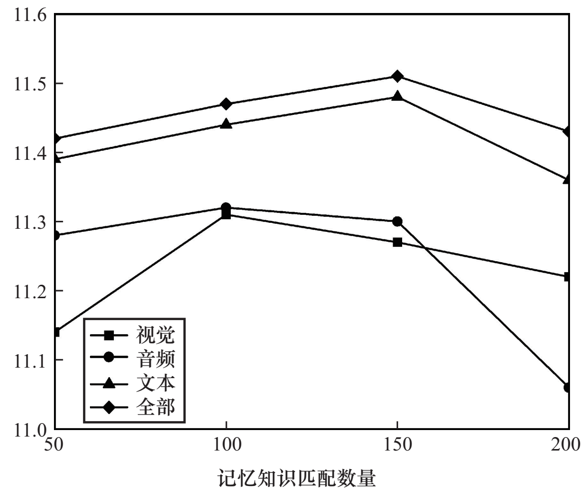


图9　各模态记忆知识匹配数量对模型METEOR性能的影响

3.5.4　提案数量对模型性能的影响

为了探究保留不同数量的事件提案对模型性能的影响，本文分别对比了K为10、25、50、100条件下模型基于事件提案生成描述的性能得分以及事件定位能力，提案数量对模型性能影响见表8，可以看到，随着保留的事件提案数量增加，模型的F1、B3、B4以及METEOR指标均有所提升，而CIDEr指标却下降明显。对此的理解是，随着事件提案保留数量的增加，事件覆盖率提高，因此，模型的F1分数提高，注重n-gram匹配的BLEU指标和注重语义匹配的METEOR得分同样上升。但是，保留更多的事件提案数量也导致事件提案出现冗余的情况，CIDEr指标更关注生成描述的多样性和独特性，事件提案的大量冗余，本文方法导致模型CIDEr指标下降。

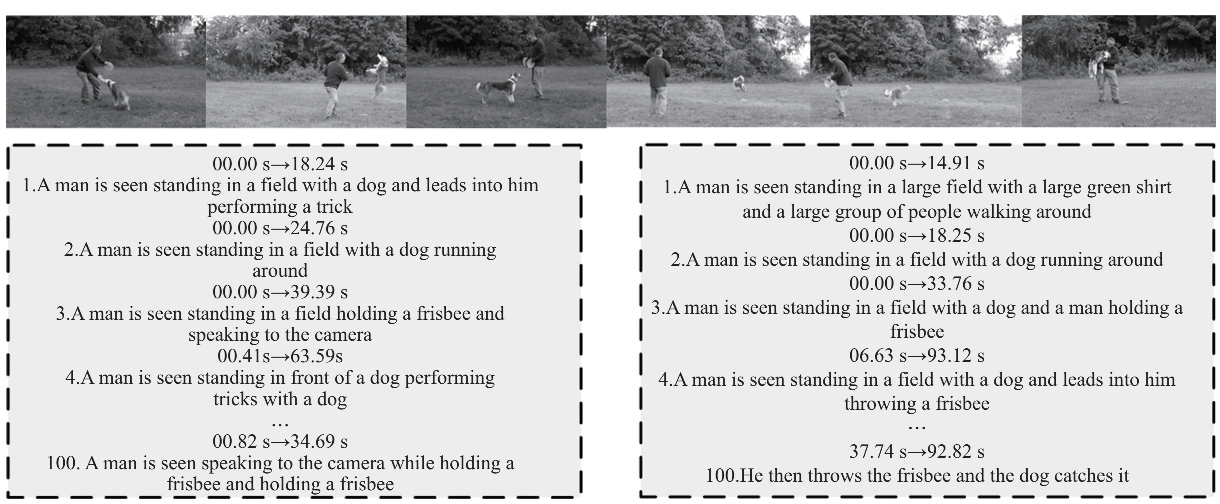
表8　提案数量对模型性能影响

提案数量	F1	With LR proposals			
		B3	B4	C	M
10	53.52	3.62	1.63	**14.7**	8.69
25	58.52	3.80	1.73	13.3	8.92
50	60.97	3.93	1.81	12.7	9.01
100	**61.55**	**4.24**	**2.07**	13.0	**9.24**

3.6　定性比较

本文在ActivityNet Captions数据集上将所提方法与基线模型BMT以及现有最先进方法CM^2进行了定性比较。ActivityNet Captions TOP 100事件候选提案生成描述如图10所示，展示了所提方法与基线模型BMT TOP K个候选事件提案生成的部分事件描述。值得注意的是，本文模型在输出描述时展现出了更高的多样性，“frisbee”“dog”和“man”等关键词的出现频率显著增加。

同时对于每个带有人工描述的真实事件，本文分别从BMT、CM^2以及提出的模型预测的提案



（a）BMT　　（b）本文模型

图10　ActivityNet Captions TOP 100事件候选提案生成描述

列表中选择与真实事件持续时间重叠最多的事件提案进行比较。本文模型与BMT、CM^{2}定性比较示例如图11所示，通过比较事件提案预测的事件边界和描述，本文方法比基线模型BMT生成更高质量的事件描述、在描述生成的准确性和质量上与CM^{2}达到了同等水平。并且本文生成的描述中关键词、短语出现的频次更高。图11分别呈现了人工描述、BMT预测描述、CM^{2}预测描述、本文模型预测描述，以及记忆知识库中所匹配到与描述生成相关且有意义的记忆知识，并采用更加关注单词、短语以及同义词匹配的方法来对比生成的描述，对匹配程度较高的单词、短语进行加粗展示。通过观察本文模型与BMT、CM^{2}生成的描述，本文模型生成的描述质量更高、更准确。例如，在演示事件2中，BMT中没有正确理解视频中人与狗的互动，将事件2描述为“performing tricks with a dog”。本文模型则提供了视频详细且更准确的描述，正确地描述了视频中的男子用飞盘与狗玩耍的场景。在演示事件3中，BMT同样没有理解视频中人与狗玩飞盘的行为，CM^{2}描述事件为“The man continues to catch the frisbee with the dog”，错误地认为视频中男子也在接飞盘，而本文模型则与人工描述一致，正确地生成了描述“The man throws the frisbee”。同时图11还展示了所匹配到的记忆知识库中对描述生成有实质性帮助的部分记忆知识，这些文本记忆是通过预测的事件提案检索记忆知识库获得的。通过检索多模态记忆知识库获得的额外多模态信息，本文方法更有助于描述生成器进行更准确的预测。

同时，本文对不同记忆知识库在模型描述生成中的作用进行了分析。记忆知识库对模型性能贡献示例如图12所示，基线模型由于缺乏记忆知识，生成的描述无法准确理解视频中人们正在进

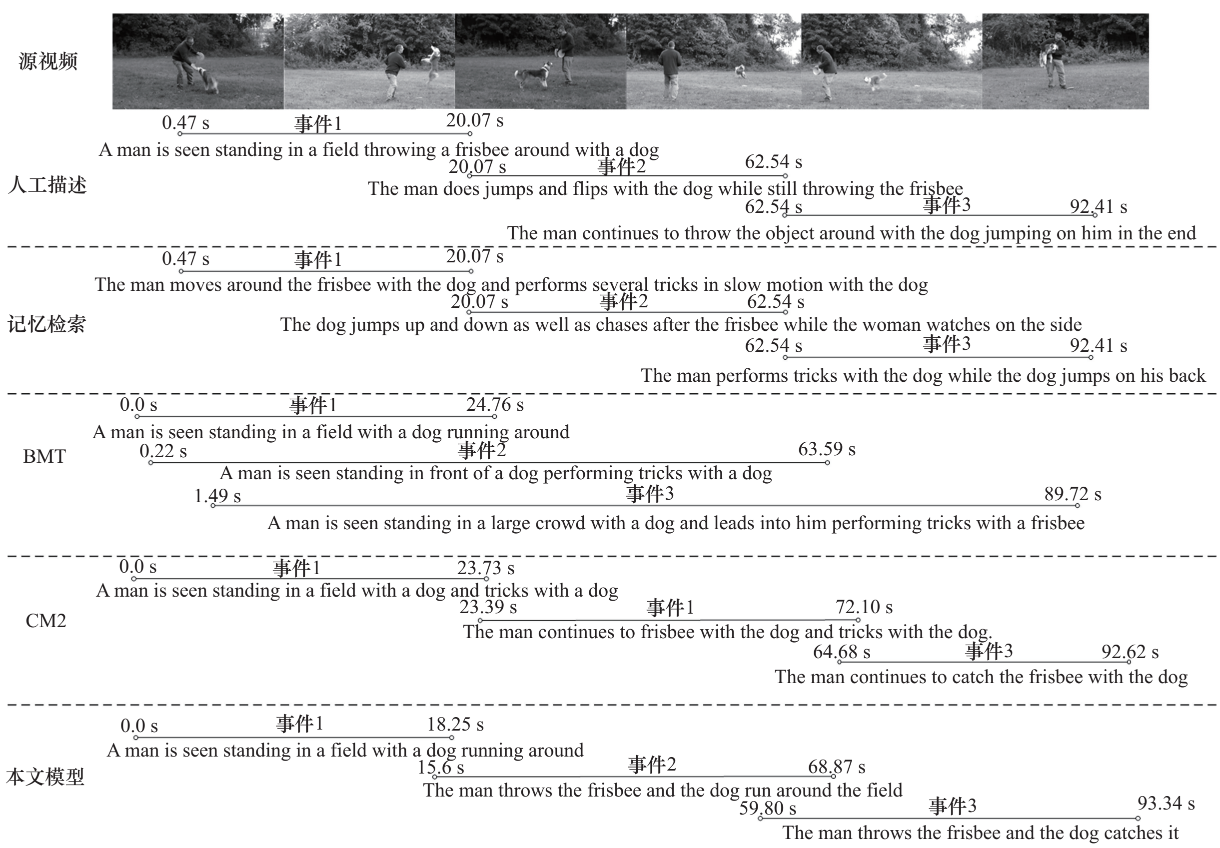


图11　本文模型与BMT、CM^{2}定性比较示例

行的球类活动，将运动场景描述成“A man is seen standing in a field and holding a stick”，而通过引入多模态记忆知识库，本文方法使模型能够根据不同模态的知识匹配结果生成更加准确和多样化的描述。具体而言，当引入音频记忆知识，模型能够捕捉音频线索并在生成的描述中体现相关语义信息，即“A man is seen speaking to the camera”；当引入视觉记忆知识，模型正确理解视频中的球类活动“lacrosse”。此外，文本通过引入记忆知识促进模型生成更符合人工描述的句子，如短语“A group of”和单词“around”，这为消融实验中变体（3）~（5）对比基准模型在B3、B4指标上的提升提供了合理的解释。这表明，多模态记忆知识库在描述生成中发挥了显著的作用，视觉和音频知识增强了模型语义理解的准确性，文本记忆知识提升了生成描述的可读性。

尽管本文模型的性能取得了显著提升，但在某些方面仍存在局限性。如图12演示视频的事件3中，本文模型未能生成如人工描述中的“in the end”、CM^{2}中的“continue”“then”等表述，这一情况说明本文模型在构建事件间时序联系及理解事件先后顺序方面存在相对不足，模型难以生成如“In the end”“then”“after that”等时序连接词，这限制了生成描述的连贯性和可读性。

4　结束语

本文提出了一种基于多模态记忆知识的密集视频描述方法。该方法通过构建多模态记忆知识库，使候选事件提案能够检索与源视频相关的视觉、音频和文本线索，从而增强最终生成描述的准确性和多样性。为了有效整合多模态记忆知识避免冗余，本文设计了一种高效的门控网络来动态平衡多模态记忆知识。在多模态融合方面，本文引入跨模态共享表征，并通过层级化注意力机制实现了深层次模态特征对齐。基于ActivityNet Captions和YouCook2数据集的大量实验结果验证了本文所提方法的有效性。未来将进一步探索如何在保留更多候选事件提案的同时更好地去除冗余事件提案，以解决CIDEr指标相对偏低的问题。同时也将深入研究基于建模视频事件间关系的密集视频描述方法，以更好地捕捉生成描述的时序关系。

参考文献：

[1] KRISHNA R, KENJI H T, REN F, et al. Dense-captioning events in videos[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 706-715.

[2] KIM M, KIM H B, MOON J, et al. Do you remember? Dense video captioning with cross-modal memory retrieval[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2024: 13894-13904.

[3] WU H, LIU H B, QIAO Y, et al. DIBS: enhancing dense video captioning with unlabeled videos via pseudo boundary enrichment and online refinement[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2024: 18699-18708.

GT:	Men wearing white uniforms are playing cricket on a field
B:	A man is seen standing in a field and holding a stick
B+V:	A man is seen standing on a field playing a game of lacrosse
B+A:	A man is seen speaking to the camera and leads into people playing a game
B+T:	A group of men are seen standing around a field playing a game
本文模型：	A group of people are seen standing around a field and playing a game of lacrosse on the field

图12　记忆知识库对模型性能贡献示例

[4] MUN J, YANG L, REN Z, et al. Streamlined dense video captioning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Piscataway: IEEE Press 2019: 6588-6597.

[5] CHOI W, CHEN J S, YOON J. Step by step: a gradual approach for dense video captioning[J]. IEEE Access, 2023, 11: 51949-51959.

[6] LI P, ZHANG P, WANG T, et al. Time–frequency recurrent transformer with diversity constraint for dense video captioning[J]. Information Processing & Management, 2023, 60(2): 103204.

[7] LASHIN V, RAHTU E. Multi-modal dense video captioning[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2020: 4117-4126.

[8] RAHMAN T, XU B C, SIGAL L. Watch, listen and tell: multimodal weakly supervised dense event captioning[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 8907-8916.

[9] JING S Q, ZHANG H N, ZENG P P, et al. Memory-based augmentation network for video captioning[J]. IEEE Transactions on Multimedia, 2023, 26: 2367-2379.

[10] CAO S, WANG B, ZHANG W, et al. Visual consensus modeling for video-text retrieval[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Los Altos: AAAI, 2022, 36(1): 167-175.

[11] 马苗, 王伯龙, 吴琦, 等. 视觉场景描述及其效果评价[J]. 软件学报, 2019, 30(4): 867-883.
MA M, WANG B L, WU Q, et al. Visual scene description and its performance evaluation[J]. Journal of Software, 2019, 30(4): 867-883.

[12] 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[J]. 自动化学报, 2022, 48(2): 375-397.
TANG P J, WANG H L. From video to language: survey of video captioning and description[J]. Acta Automatica Sinica, 2022, 48(2): 375-397.

[13] CHANG Z, ZHAO D X, CHEN H L, et al. Event-centric multimodal fusion method for dense video captioning[J]. Neural Networks, 2022, 146: 120-129.

[14] NIU J J, XIE Y L, ZHANG Y, et al. Tri-modal dense video captioning based on fine-grained aligned text and anchor-free event proposals generator[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2022, 36(12).

[15] XIE Y L, NIU J J, ZHANG Y, et al. Global-shared text representation based multi-stage fusion transformer network for multi-modal dense video captioning[J]. IEEE Transactions on Multimedia, 2023, 26: 3164-3179.

[16] WANG T, ZHANG R M, LU Z C, et al. End-to-end dense video captioning with parallel decoding[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 6827-6837.

[17] SELVI T K, S N, THOMPSON M, et al. Quasi-parallel dense video captioning: a novel approach to achieving ground truth event captions using hierarchical attention[C]//Proceedings of the 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT). Piscataway: IEEE Press, 2024: 1278-1283.

[18] CHEN F Y, XU C, JIA Q, et al. Egocentric vehicle dense video captioning[C]//Proceedings of the 32nd ACM International Conference on Multimedia. New York: ACM Press, 2024: 137-146.

[19] SHOMAN M, WANG D D, ABOAH A, et al. Enhancing traffic safety with parallel dense video captioning for end-to-end event analysis[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2024: 7125-7133.

[20] ZHOU L W, ZHOU Y B, CORSO J J, et al. End-to-end dense video captioning with masked transformer[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8739-8748.

[21] IASHIN V, RAHTU E. A better use of audio-visual cues: Dense video captioning with bi-modal transformer[C]//Proceedings of the 2020 British Machine Vision Conference. Durham: BMVA, 2020.

[22] YU W M, XU H, YUAN Z Q, et al. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Los Altos: AAAI, 2021, 35(12): 10790-10797.

[23] YANG D K, KUANG H P, HUANG S, et al. Learning modality-specific and-agnostic representations for asynchronous multimodal language sequences[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM Press, 2022: 1708-1717.

[24] WANG D H, ZHAO T, YU W H, et al. Deep multimodal complementarity learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(12): 10213-10224.

[25] YOU Q Z, JIN H L, WANG Z W, et al. Image captioning with semantic attention[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 4651-4659.

[26] CHEN Z F, ZHOU Q H, SHEN Y K, et al. Visual chain-of-thought prompting for knowledge-based visual reasoning[J].

Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(2): 1254-1262.

[27] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the International Conference on Machine Learning. New York: PMLR, 2021: 8748-8763.

[28] HERSHEY S, CHAUDHURI S, ELLIS D P W, et al. CNN architectures for large-scale audio classification[C]//Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2017: 131-135.

[29] HEILBRON F C, ESCORCIA V, GHANEM B, et al. ActivityNet: a large-scale video benchmark for human activity understanding[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2015: 961-970.

[30] ZHOU L W, XU C L, CORSO J. Towards automatic learning of procedures from web instructional videos[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1).

[31] BANERJEE S, LAVIE A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Cambridge: MIT Press, 2005: 65-72.

[32] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[J]. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2002, 7: 311-318.

[33] YANG A, NAGRANI A, SEO P H, et al. Vid2seq: large-scale pretraining of a visual language model for dense video captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 10714-10726.

[34] ZHANG Q, SONG Y Q, JIN Q. Unifying event detection and captioning as sequence generation via pre-training[C]//Proceedings of the Computer Vision-ECCV 2022. Cham: Springer, 2022: 363-379.

[35] ZHOU X Y, ARNAB A, BUCH S, et al. Streaming dense video captioning[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2024: 18243-18252.

[36] HAN S X, LIU J, ZHANG J, et al. Lightweight dense video captioning with cross-modal attention and knowledge-enhanced unbiased scene graph[J]. Complex & Intelligent Systems, 2023, 9(5): 4995-5012.

[37] WEI Y W, YUAN S Z, CHEN M, et al. Enhancing multimodal alignment with momentum augmentation for dense video captioning[C]//Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2023: 1-5.

[38] ZHU W R, PANG B, THAPLIYAL A V, et al. End-to-end dense video captioning as sequence generation[EB]. 2022.

[39] WANG J F, YANG Z Y, HU X W, et al. GIT: a generative image-to-text transformer for vision and language[EB]. 2022.

[40] WEI Y, YUAN S, CHEN M, et al. MPP-net: multi-perspective perception network for dense video captioning[J]. Neurocomputing, 2023, 552: 126523.

[作者简介]

方豪杰（2000− ），男，浙江理工大学计算机科学与技术学院（人工智能学院）硕士生，主要研究方向为计算机视觉和密集视频描述等。

李永刚（1979− ），男，博士，嘉兴大学人工智能学院、嘉兴大学全省多模态感知与智能系统重点实验室副教授、硕士生导师，主要研究方向为计算机视觉、视频图像处理、机器学习等。

曹宗瑞（2000− ），男，浙江理工大学计算机科学与技术学院（人工智能学院）硕士生，主要研究方向为计算机视觉和密集视频描述等。

叶利华（1978− ），男，博士，嘉兴大学人工智能学院、嘉兴大学全省多模态感知与智能系统重点实验室讲师、硕士生导师，主要研究方向为计算机视觉、视频图像处理等。