计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202221008
61(4): 889−902, 2024

基于多模态知识主动学习的视频问答方案

刘明阳　王若梅　周　凡　林　格
（中山大学计算机学院国家数字家庭工程技术研究中心　广州　510006）
（liumy77@mail2.sysu.edu.cn）

Video Question Answering Scheme Base on Multimodal Knowledge Active Learning

Liu Mingyang, Wang Ruomei, Zhou Fan, and Lin Ge
（*National Engineering Research Center of Digital Life*, *School of Computer Science and Engineering*, *Sun Yat-sen University*, *Guangzhou* 510006）

Abstract　Video question answering requires models to understand, fuse, and reason about the multimodal data in videos to assist people in quickly retrieving, analyzing, and summarizing complex scenes in videos, becoming a hot research topic in artificial intelligence. However, existing methods lack abilities of obtaining the motion details of visual objects in feature extraction, which may lead to false causality. In addition, in data fusion and reasoning, existing methods lack effective active learning ability, making it difficult to obtain prior knowledge beyond feature extraction, which affects the model's deep understanding of multimodal content. To address these issues, we propose a multimodal knowledge-based active learning video question answering solution. The solution acquires the semantic correlation of visual targets in image sequences and the dynamic relationship with the surrounding environment to establish the motion trajectory of each visual target. Further, static content is supplemented with dynamic content to provide more accurate video feature expression for data fusion and reasoning. Then, the solution achieves self-improvement and logical thinking focus of multimodal information understanding through knowledge auto-enhancement multimodal data fusion and reasoning model, filling the gap in deep understanding of multimodal content. Experimental results show that the performance of our scheme is better than the most advanced video question answering algorithm, and a large number of ablation and visualization experiments also verify the rationality of this solution.

Key words　video question answering; data fusion and reasoning; multimodal active learning; video details description extraction; deep learning

摘　要　视频问答是人工智能领域的一个热点研究问题.现有方法在特征提取方面缺乏针对视觉目标运动细节的获取,从而会导致错误因果关系的建立.此外,在数据融合与推理过程中,现有方法缺乏有效的主动学习能力,难以获取特征提取之外的先验知识,影响了模型对多模态内容的深度理解.针对这些问题,首先,设计了一种显性多模态特征提取模块,通过获取图像序列中视觉目标的语义关联以及与周围环境的动态关系来建立每个视觉目标的运动轨迹.进一步通过动态内容对静态内容的补充,为数据融合与推理提供了更加精准的视频特征表达.其次,提出了知识自增强多模态数据融合与推理模型,实现了多模态信息理解的自我完善和逻辑思维聚焦,增强了对多模态特征的深度理解,减少了对先验知识的依赖.最后,提出了一种基于多模态知识主动学习的视频问答方案.实验结果表明,该方案的性能优于现有最先进的视频问答算法,大量的消融和可视化实验也验证了方案的合理性.

收稿日期: 2022-12-16; **修回日期**: 2023-06-26
基金项目: 国家重点研发计划项目(2021YFF0900900)
This work was supported by the National Key Research and Development Program of China(2021YFF0900900).
通信作者: 林格(linge3@mail.sysu.edu.cn)

关键词 视频问答；数据融合与推理；多模态主动学习；视频细节描述提取；深度学习

中图法分类号 TP391

视频问答任务旨在通过问答的形式来帮助人们快速检索、解析和总结视频内容. 相较于基于静态图像的问答任务[1]，视频问答需要处理的信息从图像变成由连续图像序列、音频等多模态信息组成的视频，复杂的人物关系和上下文关联分散在这些多模态信息序列中，蕴含着一个完整的故事情节. 这使得视频问答面临着更为复杂的多模态特征提取、数据融合以及跨模态逻辑推理[2-3]等人工智能关键问题的挑战，成为比图像问答更高层次的人工智能任务.

为了实现视频问答的任务，研究人员使用了一系列的深度神经网络[4-6]来进行视频内丰富的外观信息、空间位置信息、动作信息、字幕、语音和问题文本等多模态信息的特征编码，为数据融合与推理提供必要的上下文语义线索. 为了理解分散在连续视频图像序列内的完整故事情节和获取准确的答案预测，研究人员提出了跨模态注意力机制[7]，动作−外观记忆网络[8]和图神经网络[9]等一系列数据融合与推理模型，尝试通过跨模态语义的计算与推理，从繁杂的多模态特征编码中识别和整合出那些可能在时间上相邻或不相邻的有效特征序列，过滤掉不相关甚至不利于解答问题的多模态信息，为给定问题预测准确的答案.

文献 [7–9] 在多模态特征提取和数据融合与推理方面取得了许多有意义的研究成果. 但是由于视频问答任务的多元性和复杂性，视频问答任务中多模态特征提取以及数据融合和推理的研究仍然是具有挑战性的难点问题. 通过对中外文献的研究与分析，我们发现在视频问答的研究中仍存在 2 点不足：

1)特征提取方法对于视频的细节表示不足. 目前的多模态特征提取方法更注重关于视频图像和视频片段粗粒度的特征提取[10-11]，粗粒度的外观信息或动作信息缺乏对图像序列内视觉目标等细粒度信息的关注，致使在数据融合与推理过程中，视频中重要的视觉目标及其动作细节可能被遗漏，影响了正确的空间位置和时序关系的建立，导致数据融合与推理过程可能建立错误的因果关系.

2)数据融合与推理的主动学习能力不足. 现阶段的数据融合与推理模型主要是针对视觉线索的单向筛选处理[12-13]，缺少主动使用已经掌握的内容来完善多模态信息的能力. 更确切地说，现阶段数据融合与推理模型无法使用已经掌握的知识去主动学习或猜测那些还没有掌握的内容，导致在数据融合与推理过程中只能对特征提取阶段所获取的多模态特征编码进行计算与推理，很难在数据融合与推理阶段获取特征提取之外的多模态先验知识，影响了模型对多模态内容的深度理解，加剧了语义鸿沟对跨模态数据融合与推理的影响.

针对这 2 点不足，本文提出了基于多模态知识主动学习的视频问答方案，如图 1 所示. 该方案由 3 个部分组成：显性多模态特征提取模块、知识自增强多模态数据融合与推理模型、答案解码模块. 首先，为了解决特征提取方法对于视频的细节表示不足的问题，我们设计了一种显性多模态特征提取模块. 该模块通过计算带有语义约束、空间约束和动态约束的显式轨迹，得到每个视觉目标的运动轨迹，从而抑制可能存在的目标位置偏移、重叠或变形所引起的语义偏移，实现了对视觉目标的精准动态特征提取. 接着，该模块借助动态特征对静态内容的补充，有效避免错误时序关联的建立和错误因果关系的推断，为数据融合与推理提供了更加精准的视频特征表达.

为了解决逻辑推理的主动学习能力不足的问题，我们设计了一种知识自增强多模态数据融合与推理(knowledge auto-enhancement multimodal data fusion and reasoning, KAFR)模型. 该模型以显性多模态特征提取模块的外观信息、动作信息和包含了视觉目标、复杂运动轨迹和多维时空交互的视频细节信息作为输入，通过时序表达与推理、多模态表示再学习、聚焦表示学习和汇总表示学习 4 种模块组成的数据融合与推理网络，赋予了视频问答模型从初次审题与推理，到信息的重学习，再到思维聚焦，最后归纳总结的完整逻辑思维能力.

在数据融合与推理过程中，该模型能够利用已经掌握的多模态信息来完善视频问答系统的先验知识，同时通过逻辑思维的聚焦能力，减少视频中需要理解的多模态信息，改善对先验知识的依赖.

为了获取分散在视频片段和图像中的视觉语义线索，我们将 KAFR 按照视频的层次结构如图像、视频片段等进行排列，使得视频问答模型能够自底向上地收集视频所提供的视觉语义线索. 然后通过答案解码模块对分散在不同模态下的答案线索进行汇总，为特定问题提供准确的答案预测.

本文的主要贡献包括 3 个方面：

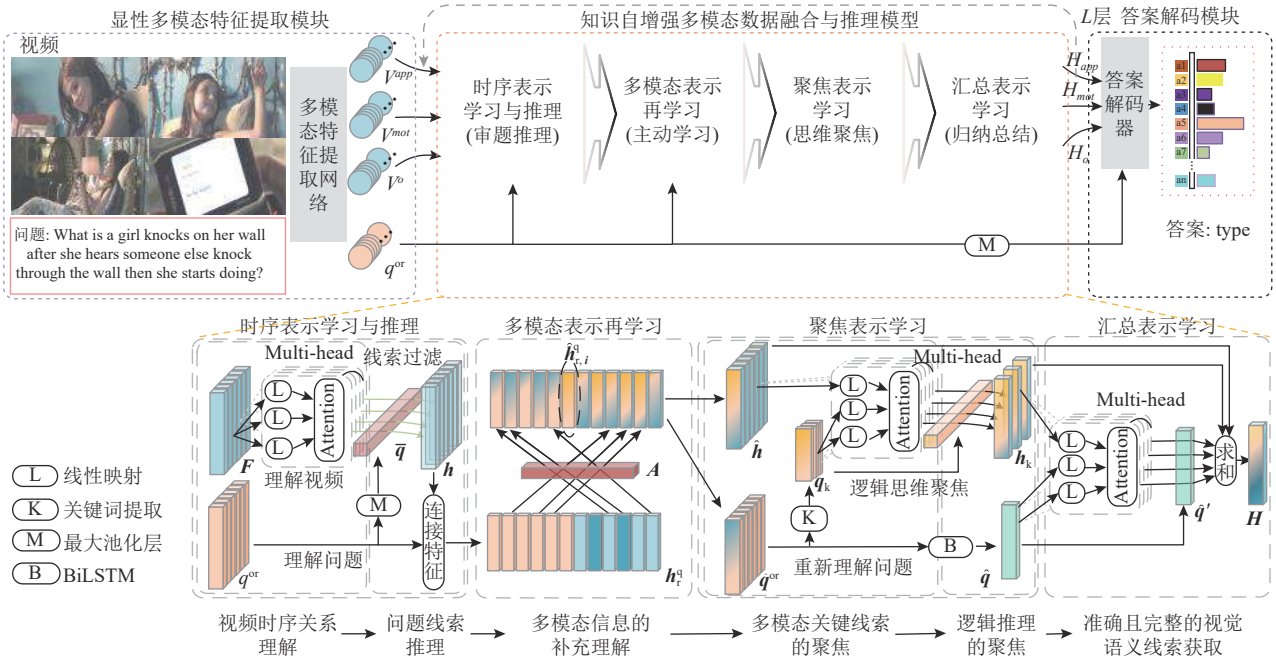


Fig. 1 The overview of our proposed video question answering scheme

图 1 本文提出的视频问答方案概述

1)提出了一种显性的视频细节描述方法. 该方法能够将视频的静态细节描述推广到动态细节描述, 为数据融合与推理提供更精准的视频描述表达.

2)设计了一种 KAFR 模型. 该模块能够在数据融合与推理计算过程中主动完善多模态信息的深度理解, 还能通过思维的聚焦学习, 减少视频中需要理解的多模态信息, 降低数据融合与推理对于先验知识的依赖, 改善特征提取不足所带来的挑战.

3)基于对 1)和 2)的改进, 提出了一种新颖的基于多模态知识主动学习的视频问答解决方案, 该方案能够自底向上地收集视频所提供的视觉语义线索, 有效地完成视频问答任务. 在 TGIF-QA[14], MSVD-QA[15], MSRVTT-QA[16] 视频问答标准数据集的实验表明, 本文提出的解决方案的性能优于现有最先进的视频问答算法.

1 相关工作

视频问答任务需要通过视觉和语言之间的跨模态数据推理来实现对复杂视频场景的理解, 这需要视频问答模型能够对视频内容进行精准编码, 并通过数据融合与推理计算将分散在空间和时间内的多模态语义线索联系起来. 这使得视频特征提取和数据融合与推理成为现阶段视频问答 2 个关键的研究点. 本节将对这 2 个关键研究问题的国内外研究现状进行分析和总结.

1.1 视频特征提取

视频特征提取旨在获取视频中包含的目标、动作、复杂的动态位置关系和上下文关联等丰富的视觉语义, 组成能够反映整个故事情节的特征表达, 为后续的跨模态数据融合与推理提供完整的视觉语义线索. 视频问答的早期方法主要通过 VGG[17], ResNet[4], ResNeXt[5] 等一系列深度网络从原始视频中提取和整合视觉语义特征[10,12,18]. 然而, 文献 [4−5, 10, 12, 17−18] 仅仅利用了图像级或视频片段等粗粒度视觉特征来描述故事情节, 缺乏对视频细节信息的关注. 最近, 针对对象级信息进行视频特征提取展现出卓越的性能[19-20], 为视频问答模型提供了故事情节的细节描述, 增强了视觉关系推理的能力. Huang 等人[19] 通过建立图像帧间与帧内的位置编码来丰富对象特征的时空关系. Seo 等人[20] 将对象级特征提取推广到运动特征的提取, 增强了对象特征的动态表达.

文献 [4−5, 10, 12, 17−20] 方法通过对视频的细节特征提取, 有效地提升了视频问答的性能. 但是这些方法只关注到图像所提供的静态细节特征和时空进行关联, 没有显式地捕获视觉目标的动态细节特征, 这样可能会导致错误的关系理解, 如拥抱和打架, 也可能无法捕获视觉目标的动作细节, 如挥手和亲吻. 为了解决上述问题, 本文显式地计算出每一个视觉目标的运动轨迹, 对每一个视觉目标进行精准的细

节特征提取，同时通过动态信息对静态内容的补充，有效地避免了错误时序关联的建立，纠正了错误的因果关系.

1.2 数据融合与推理

数据融合与推理的目的是从复杂的视频故事情节中获取能够指引出正确答案的视觉线索. 在视频问答的早期发展中，研究人员专注于将视频图像或视频片段作为数据融合与推理的对象，提出了跨模态注意力机制、动作外观记忆网络和图神经网络等一系列数据融合与推理技术，试图通过单个问答模型来获取整个视频的内容[21-22]. 近年来，为了获取对视频细节内容的理解，避免问答模型忽略掉那些影响视频故事走向的重要线索，基于模块化的视频问答模型成为了主流[9,12]，它们将数据融合与推理过程渗透到视频的各个层次，通过多步推理的方式，完成对视频从对象级、图像级到片段级的语义线索整合. Le 等人[12]设计了一种能够重复使用的条件关系模块，并且将这些模块按照视频的时序结构进行排列，以捕获存在于视频帧之间和视频片段之间的时序关系. 为了进一步完善对视频层次行的利用，Dang 等人[9]利用图神经网络对视频内的对象及其轨迹进行关系推理，使得数据融合与推理能够深入到场景目标的时空关系中，获取更精准的视觉语义线索.

文献 [9, 12, 21−22] 方法通过对数据融合与推理模块的结构创新，使视频问答任务的性能方面得到了改进. 进一步分析这些方法的数据融合与推理原理，我们发现这些研究都建立在有限的视频特征提取之上，只能获取基于 Imagenet[23]，Kinetics[24] 等数据集的视频先验知识. 然而相较于复杂的视频内容，这些从数据集中获取的有限先验知识很难对视频内容进行准确的描述，无法为后续的数据融合与推理提供充足的视觉知识，使得文献 [9, 12, 21−22] 方法不得不在缺失信息的情况下进行答案预测，严重限制了这些方法的问答性能. 为了应对这种先验知识不足的问题，Zeng 等人[25]提出了一种先验知识检索模块，旨在从外部知识获取先验知识，并将其整合到问题特征中，以丰富多模态信息的特征表达. 同时，研究人员也使用开放域视觉-文本数据[26]进行网络预训练[27-28]，以改善视频问答模型先验知识不足的问题. 虽然文献 [25−28] 方式获取了不错的性能提升，但是不论是数据的获取和标注，还是信息的检索，都是一种费时费力的方法. 因此在本文中，我们设计了一种 KAFR 模型，使得视频问答模型不仅能够在跨模态数据融合与推理过程中，增强对多模态内容的理解，弥补先验知识不足的缺陷，还能够通过逻辑思维的聚焦能力，将逻辑推理聚焦于与问题相关联的多模态信息，进一步减少对先验知识的依赖.

2 基于多模态知识主动学习的视频问答方案

2.1 问题描述

对于任意视频 V 以及对应的任意自然语言问题 q，视频问答需要设计出一个算法 $\mathcal{F}$，从候选答案空间 $\mathcal{A}$ 中推导出正确答案 a^*. 该过程可以定义为：

$$a^* = \arg\max_{a\in\mathcal{A}} \mathcal{F}(a|q,V). \tag{1}$$

为了实现视频问答任务，本文提出的视频问答方案 $\mathcal{F}$ 被分为 3 个部分进行阐述：1）显式多模态特征提取模块（见 2.2 节）；2）KAFR 模型（见 2.3 节和 2.4 节）；3）答案预测模块（见 2.5 节）.

2.2 显性多模态特征提取模块

为了能够更好地获取视觉目标在静态图像内的语义关系和视觉目标与周围环境的动态关系，我们建立了一种显性的多模态特征提取模块. 该模块主要包括了粗粒度视觉特征提取和显性视频细节描述. 粗粒度视觉特征提取能够获取蕴含在视频图像或片段内的全局静态特征和动态特征，显性视频细节描述能够通过显式轨迹计算得到每一个视觉目标的运动轨迹，从而实现关于视觉目标的精准动态特征提取.

2.2.1 粗粒度视觉特征提取方法

粗粒度视觉特征提取模块的目的是为了获取视频图像和图像序列内蕴含的粗粒度动态特征和静态表观特征，我们首先将视频 V 分割为等长的片段 $C=C_1,C_2,\cdots,C_N$，并从每一个片段 C_i 均匀采样出 T 帧 $\{I_{i,j}\}_{j=1}^{T}$ 表示视频内容. 接着应用 ResNet[4] 和线性投影矩阵 $\boldsymbol{W}_{\text{app}}\in\mathbb{R}^{2\,048\times d}$ 来获取每一段视频 C_i 内的静态表观特征序列 $\boldsymbol{V}_i^{\text{app}}=\left(\boldsymbol{v}_{i,j}^{\text{app}}\in\mathbb{R}^d\right)_{j=1}^{T}$，最后应用 ResNeXt-101[5] 以及线性投影矩阵 $\boldsymbol{W}_{\text{mot}}\in\mathbb{R}^{2\,048\times d}$ 来获取每一段视频 C_i 内的运动特征 $\boldsymbol{V}_i^{\text{mot}}\in\mathbb{R}^d$.

2.2.2 显性视频细节描述方法

粗粒度的视觉特征能够为后续的数据融合和推理提供视频内丰富的全局信息，但是高度耦合的信息表达不利于视频细节的获取. 为了补充视频内的细节信息，更好地获取视觉目标在静态图像内的语义关系和与周围环境的动态关系，我们设计了一种显性视频的细节描述方法，方法流程如图 2 所示.

具体来说，我们首先利用目标检测器[6]从视频片

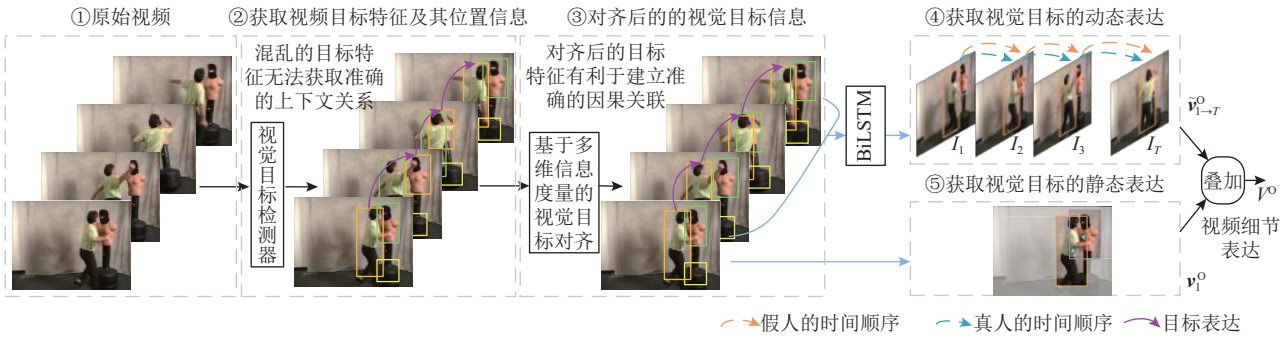


Fig. 2 Detail description method of explicit video

图 2 显性视频的细节描述方法

段C_i的每一帧图像$I_{i,j}$中提取K个视觉目标特征$V_i^{\text{or}}=\left\{\boldsymbol{v}_{i,j}^{\text{or}}\middle|\boldsymbol{v}_{i,j}^{\text{or}}\in\mathbb{R}^{K\times 2\,048}\right\}_{j=1}^{T}$和相应的空间位置信息$V_i^{\text{ob}}=\left\{\boldsymbol{v}_{i,j}^{\text{ob}}\middle|\boldsymbol{v}_{i,j}^{\text{ob}}\in\mathbb{R}^{K\times 4}\right\}_{j=1}^{T}$. 由于目标检测结果可能存在由于目标位置偏移、重叠或变形所引起的语义偏移，使得目标检测的结果顺序无法被预测，这就需要我们对这些检测目标重新排序，以避免获取错误的上下文关系和动态信息. 为了对齐每一个视觉目标的特征序列，我们定义了一种相似度得分 *score* 来衡量相邻帧之间的视觉目标相似度：

$$score_{j,j+1}^{k_1,k_2}=\cos(\boldsymbol{v}_{i,j}^{\text{or},k_1},\boldsymbol{v}_{i,j+1}^{\text{or},k_2})+IoU(\boldsymbol{v}_{i,j}^{\text{ob},k_1},\boldsymbol{v}_{i,j+1}^{\text{ob},k_2})+\tanh\left(\frac{z_{i,j+1}^{k_2}-z_{i,j}^{k_1}}{z_{i,j}^{k_1}-z_{i,j-1}^{k_1}}\right),\tag{2}$$

其中 cos() 表示余弦相似度，用于评估相邻帧的视觉目标之间的语义相似度，以区分不同视觉目标，避免由于错误的时序关联而造成的语义偏移；*IoU*() 表示交并比，用于计算视觉目标之间的空间位置关联，以区分在相同位置或大小不同的视觉目标之间的语义相似性，避免由于错误的空间关联而造成的语义偏移；z 表示视觉目标位置的中心位置，tanh() 表示激活函数，用于限制每个视觉目标的运动范围，评估视觉目标的运动趋势，以避免目标重叠时产生的语义偏移，$j\in\{1,2,\cdots,T-1\}$，$k_1\in\{1,2,\cdots,K\}$，$k_2\in\{1,2,\cdots,K\}$. 借助于这些度量方法，我们可以以每一个视频片段C_i的第 1 帧检测到的K个视觉目标作为基准目标，逐帧计算相邻帧之间的 *score* 得分，接着应用贪心算法获取最大化的 *score* 得分，将相似视觉目标连接起来，从而捕获视觉目标在视频片段中的运动轨迹，实现视觉目标的对齐. 上述方式有效地避免错误的时序关联，为视频问答模型提供了对齐后的视觉目标特征序列$\hat{\boldsymbol{V}}_i^{\text{or}}=\left\{\hat{\boldsymbol{v}}_{i,j}^{\text{or}}\in\mathbb{R}^{K\times 2\,048}\right\}_{j=1}^{T}$和空间位置序列$\hat{\boldsymbol{V}}_i^{\text{ob}}=\left\{\hat{\boldsymbol{v}}_{i,j}^{\text{ob}}\in\mathbb{R}^{K\times 4}\right\}_{j=1}^{T}$.

接下来，我们从每一段视频的第 1 帧图像中获取每一个视觉目标的空间静态特征$\boldsymbol{v}_{i,1}^{\text{ob}}\in\mathbb{R}^{K\times d}=ELU([\hat{\boldsymbol{v}}_{i,1}^{\text{or}};\hat{\boldsymbol{v}}_{i,1}^{\text{ob}}]\boldsymbol{W}_{\text{ob}})$，其中 [;] 表示特征拼接，$\boldsymbol{W}_{\text{ob}}\in\mathbb{R}^{(2\,048+4)\times d}$表示线性投影矩阵，*ELU*() 是指数线性激活函数. 针对每一个视觉目标的动态信息，我们使用双向长短期记忆网络 *BiLSTM*（bi-directional long short-term memory）沿着时间方向将每一个视觉目标的动态变化嵌入到 d 维空间内，获取视频片段C_i的视觉目标的细节动态表达$\tilde{\boldsymbol{v}}_{i,1\to T}^{\text{ob}}=[BiLSTM(\{\hat{\boldsymbol{v}}_{i,j}^{\text{ob}}\}_{j=1}^{T})]\hat{\boldsymbol{W}}_{\text{ob}}$，其中$\hat{\boldsymbol{W}}_{\text{ob}}\in\mathbb{R}^{d\times d}$. 最终我们使用动态信息完成对静态信息的补充，得到每一段视频C_i的视频细节表达$\boldsymbol{V}_i^{\text{ob}}\in\mathbb{R}^{K\times d}=\tilde{\boldsymbol{v}}_{i,1\to T}^{\text{ob}}+\boldsymbol{v}_{i,1}^{\text{ob}}$. 这种方法能够有效地避免错误时序关联的建立，纠正错误的因果关系，为数据融合与推理提供更精准的视频特征表达.

为了获得一种良好的问题文本与候选答案文本的特征表达，我们使用了微调后的 BERT 模型[29]来提取问题中每个词的特征表达$q^{\text{or}}=\left\{\boldsymbol{q}_i^{\text{or}}\in\mathbb{R}^d\right\}_{i=1}^{len}$，以及多项选择题任务中的候选答案特征表达$\left\{\boldsymbol{a}_i\in\mathbb{R}^d\right\}_{i=1}^{N_{\text{ca}}}$，其中$N_{\text{ca}}$表示候选答案的数量，*len* 表示问题的句长.

显性多模态特征提取模块为数据融合与推理模型提供了具有全局静态信息的$V^{\text{app}}=\left\{\boldsymbol{V}_i^{\text{app}}\right\}_{i=1}^{N}$、全局动态信息$V^{\text{mot}}=\left\{\boldsymbol{V}_i^{\text{mot}}\right\}_{i=1}^{N}$和视频的细节信息$V^{\text{ob}}=\left\{\boldsymbol{V}_i^{\text{ob}}\right\}_{i=1}^{N}$，还提供了问题文本和候选答案文本的特征表达$q^{\text{or}}=\left\{\boldsymbol{q}_i^{\text{or}}\in\mathbb{R}^d\right\}_{i=1}^{len}$与$\left\{\boldsymbol{a}_i\in\mathbb{R}^d\right\}_{i=1}^{N_{\text{ca}}}$. 接下来，我们将介绍如何从这些复杂的多模态信息中获取问题语义线索.

2.3 KAFR 模型

现阶段的数据融合与推理模型主要是针对视觉线索的单向筛选处理[12-13]，缺少主动获取特征提取之外先验知识的手段，影响了模型对多模态内容的深度理解和跨模态数据融合与推理的能力. 为此，本文提出了 KAFR 模型. 该模块的输入是长度为 X 的视频特征序列$F=\left\{\boldsymbol{f}_j\middle|\boldsymbol{f}_j\in\mathbb{R}^d\right\}_{j=1}^{X}$和问题特征$q^{\text{or}}$，通过 4 个跨模态数据融合与推理过程：时序表示学习与推理、多模态表示再学习、聚焦表示学习和汇总表示学习

赋予视频问答模型从初次审题与推理，到信息的重学习，再到思维聚焦，最后归纳总结的完整逻辑思维能力. 使得数据融合与推理过程中不仅能够利用所收集的视觉线索填补对多模态信息的理解，还能通过逻辑思维的聚焦能力，改善逻辑推理对于先验知识的依赖.

2.3.1 时序表示学习与推理

时序表示学习与推理旨在建立视觉特征的上下文关系，以理解视频内容并整理与问题相关联的视觉语义线索，例如从视觉目标中获取与问题所关注的视觉对象及其动态轨迹. 为了实现这样的目的，我们首先使用多头注意力模型[30]来捕获视频特征序列 F 中各个特征向量之间的语义关系，使得 F 中每个特征向量能够在多个维度上共享其特征，赋予模型理解视频的能力. 该过程如式(3)(4)所示：

$$\boldsymbol{F}' = [\boldsymbol{head}_1;\boldsymbol{head}_2;\cdots;\boldsymbol{head}_H]\boldsymbol{W}^{\mathrm{O}'}, \tag{3}$$

$$\boldsymbol{head}_i = softmax\left(\frac{\boldsymbol{F}\boldsymbol{W}_i^{\mathrm{Q}}(\boldsymbol{F}\boldsymbol{W}_i^{\mathrm{K}})^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{F}\boldsymbol{W}_i^{\mathrm{V}}, \tag{4}$$

其中 $\boldsymbol{W}_i^{\mathrm{Q}} \in \mathbb{R}^{d\times\frac{d}{H}}$，$\boldsymbol{W}_i^{\mathrm{K}} \in \mathbb{R}^{d\times\frac{d}{H}}$，$\boldsymbol{W}_i^{\mathrm{V}} \in \mathbb{R}^{d\times\frac{d}{H}}$，$\boldsymbol{W}^{\mathrm{O}'} \in \mathbb{R}^{d\times d}$ 是不同的线性投影矩阵，$\sqrt{d}$ 是缩放因子，$i = 1,2,\cdots,H$，H 表示多头注意力模型中注意力模型的个数，$F = [\boldsymbol{f}_1;\boldsymbol{f}_2;\cdots;\boldsymbol{f}_X]$. 通过式(3)(4)，为视频问答模型获取了具有上下文关系的视觉语义表达 $F' = \left\{\boldsymbol{f}_j' \,\middle|\, \boldsymbol{f}_j' \in \mathbb{R}^d\right\}_{j=1}^{X}$. 接着，我们利用问题的整体表达 $\bar{\boldsymbol{q}} \in \mathbb{R}^d = maxpool(q^{\mathrm{or}})$ 从视频 F' 的各个情节 $\boldsymbol{f}_j'$ 中推理出问题线索，$maxpool()$ 表示最大池化操作，获取语境 F' 中存在的隐含语义线索 $\boldsymbol{h}$.

$$\begin{cases}\boldsymbol{h}_j = ELU([\boldsymbol{f}_j;\bar{\boldsymbol{q}}]\boldsymbol{W}_j),\\ \boldsymbol{h} = [\boldsymbol{h}_1;\boldsymbol{h}_2;\cdots;\boldsymbol{h}_X],\end{cases} \tag{5}$$

其中 $\{\boldsymbol{W}_j \in \mathbb{R}^{2d\times d}\}_{j=1}^{X}$ 是不同的线性投影矩阵. 通过时序表达学习和推理，视频问答模型获取了解读视频内容的能力，整理出与问题相关联的视觉线索. 但是特征提取模块所提供的有限先验知识很难保证视频问答模型能够完全理解视频或文本内容，多模态信息理解的缺失可能导致错误因果关系的学习，影响后续的数据融合与推理计算. 为此我们设计了“多模态表示再学习”模块.

2.3.2 多模态表示再学习

多模态表示再学习的目的是利用已经获取的视觉语义线索，增强对多模态信息的深度理解，并弥补先验知识的不足. 例如，该模块可以利用已经明确的视觉目标及其轨迹信息，来强化或补充那些在特征提取阶段无法获取的视觉目标先验知识. 为此，我们首先使用式(6)获取视觉语义特征和文本特征之间的复杂语义关系 $\boldsymbol{A}$，以便指导后续的多模态信息之间的语义补充理解.

$$\boldsymbol{A} \in \mathbb{R}^{(X+len)\times(X+len)} = ((\boldsymbol{h}_{\mathrm{r}}^{\mathrm{q}})\boldsymbol{W}_{\mathrm{r1}})((\boldsymbol{h}_{\mathrm{r}}^{\mathrm{q}})\boldsymbol{W}_{\mathrm{r2}})^{\mathrm{T}}, \tag{6}$$

其中 $\boldsymbol{W}_{\mathrm{r1}} \in \mathbb{R}^{2d\times d}$ 和 $\boldsymbol{W}_{\mathrm{r2}} \in \mathbb{R}^{2d\times d}$ 是线性投影矩阵，$\boldsymbol{h}_{\mathrm{r}}^{\mathrm{q}} = [\boldsymbol{h};q^{\mathrm{or}}]$ 将视觉特征 $\boldsymbol{h}$ 与问题原始特征 q^{or} 组合到同一向量中. 接着在关系网络 $\boldsymbol{A}$ 的引导下，利用已经掌握的多模态语义补充每一个视觉信息和问题词汇的深度理解 $\hat{\boldsymbol{h}}_{\mathrm{r}}^{\mathrm{q}}$：

$$\hat{\boldsymbol{h}}_{\mathrm{r},i}^{\mathrm{q}} = ReLU\left(\boldsymbol{h}_{\mathrm{r},i}^{\mathrm{q}} + \sum_{j\in N_i}\alpha_{i,j}(\boldsymbol{h}_{\mathrm{r},j}^{\mathrm{q}}\boldsymbol{W}_{\mathrm{r3}})\right), \tag{7}$$

其中 $\boldsymbol{W}_{\mathrm{r3}} \in \mathbb{R}^{2d\times d}$ 是线性投影矩阵，N_i 表示除第 i 个特征节点外的节点特征集合，$\boldsymbol{A} = \left(\alpha_{i,j}\right)_{i=1,j=1}^{(X+len),(X+len)}$ 表示特征之间的关联程度，$ReLU$ 表示修正线性单元激活函数. 经过上述的迭代操作，重复地对多模态语义进行补充与被补充，最终获取到充分理解后的视频 $\hat{\boldsymbol{h}}$ 和问题序列 $\hat{q}^{\mathrm{or}}$. 接着应用 $BiLSTM()$ 进行针对问题的重新审阅，获取理解更为准确的问题表达 $\hat{\boldsymbol{q}} = BiLSTM(\hat{q}^{\mathrm{or}})$. 通过对多模态信息的再学习，实现了模型对多模态特征的深度理解，填补了多模态先验知识的不足.

2.3.3 聚焦表示学习

为了进一步实现对多模态内容关键点的聚焦，减少与问题弱相关或无关的视觉信息对数据融合与推理的干扰，从复杂的视频场景中找出与问题强相关视觉语义线索，例如蕴含着答案的潜在视觉目标以及其运动轨迹更有利于问题的解答. 为此，一种聚焦表示学习模块被提出，旨在实现逻辑思维的聚焦能力. 该模块的目的是利用问题的关键词，使视频问答模型能够聚焦多模态内容中的关键内容，减少推理过程中可能造成混淆的无关或弱相关的内容. 在该模块的设计中，我们首先使用关键词检测技术①从问题 $\hat{\boldsymbol{q}}^{\mathrm{O}}$ 中获取每个关键词的语义表达 $q_{\mathrm{k}} = \left\{\boldsymbol{q}_{\mathrm{k},i}\right\}_{i=1}^{n}$，其中 n 表示关键词的个数. 视频问答模型借助关键词 q_{k} 从隐藏的语义线索 $\hat{\boldsymbol{h}}$ 中准确地识别出与关键信息相关的视觉信息，以总结出与问题强相关的视觉语义线索.

为了实现上述功能，我们首先将 $\hat{\boldsymbol{h}}$ 和 q_{k} 作为一个多头注意力模型的输入，经过在多维度语义空间的

① https://github.com/maartengr/keybert

关键词筛选，获取与问题相关的聚焦表达 $\boldsymbol{h}_{\mathrm{k}}\in\mathbb{R}^{n\times d}$：

$$\begin{aligned}\boldsymbol{h}_{\mathrm{k}}=&MultiHead(q_{\mathrm{k}},\hat{\boldsymbol{h}},\hat{\boldsymbol{h}})=\\&[\boldsymbol{head}'_1;\boldsymbol{head}'_2;\cdots;\boldsymbol{head}'_H]\hat{\boldsymbol{W}}^{O'},\end{aligned}\tag{8}$$

$$\boldsymbol{head}'_i=softmax\left(\frac{q_{\mathrm{k}}\hat{\boldsymbol{W}}_i^{\mathrm{Q}}(\hat{\boldsymbol{h}}\hat{\boldsymbol{W}}_i^{\mathrm{K}})^{\mathrm{T}}}{\sqrt{d}}\right)\hat{\boldsymbol{h}}\hat{\boldsymbol{W}}_i^{\mathrm{V}},\tag{9}$$

其中 $\hat{\boldsymbol{W}}_i^{\mathrm{Q}}\in\mathbb{R}^{d\times\frac{d}{H}}$，$\hat{\boldsymbol{W}}_i^{\mathrm{K}}\in\mathbb{R}^{d\times\frac{d}{H}}$，$\hat{\boldsymbol{W}}_i^{\mathrm{V}}\in\mathbb{R}^{d\times\frac{d}{H}}$，$\hat{\boldsymbol{W}}^{O'}\in\mathbb{R}^{d\times d}$ 是不同的线性投影矩阵. 接着，为了从 $\boldsymbol{h}_{\mathrm{k}}$ 的语义空间中获取更为有效的语义线索 $\hat{\boldsymbol{h}}_{\mathrm{k}}$，我们使用了关键词的特征之和 $\hat{\boldsymbol{q}}_{\mathrm{k}}=sum(q_{\mathrm{k}})$ 来对 $\boldsymbol{h}_{\mathrm{k}}$ 进行 2 次筛选：

$$\begin{cases}\hat{\boldsymbol{h}}_{\mathrm{k},i}=ELU([\boldsymbol{h}_{\mathrm{k},i};\hat{\boldsymbol{q}}_{\mathrm{k}}]\boldsymbol{W}_i^{\mathrm{k}}),\\\hat{\boldsymbol{h}}_{\mathrm{k}}=[\hat{\boldsymbol{h}}_{\mathrm{k},1};\hat{\boldsymbol{h}}_{\mathrm{k},2};\cdots;\hat{\boldsymbol{h}}_{\mathrm{k},n}],\end{cases}\tag{10}$$

其中 $\{\boldsymbol{W}_i^{\mathrm{k}}\in\mathbb{R}^{2d\times d}\}_{i=1}^{n}$ 是不同的线性投影矩阵. 通过上述的模块的处理，问答模型已经从视频特征序列 F 中收集了更为详尽的语义线索 $\hat{\boldsymbol{h}}_{\mathrm{k}}$.

2.3.4 汇总表示学习

汇总表示学习模块是对整个数据融合与推理过程的整合，推导出最终的结论，获得准确和完整的高层次浓缩视觉语义表达. 首先使用完全理解的问题 $\hat{\boldsymbol{q}}$ 来分析视觉关键语义线索 $\hat{\boldsymbol{h}}_{\mathrm{k}}\in\mathbb{R}^{n\times d}$，并通过线性映射来获得与问题强相关的浓缩视觉语义线索 $\hat{\boldsymbol{H}}'$：

$$\begin{cases}\hat{\boldsymbol{H}}\in\mathbb{R}^{d}=MultiHead(\hat{\boldsymbol{q}},\hat{\boldsymbol{h}}_{\mathrm{k}},\hat{\boldsymbol{h}}_{\mathrm{k}}),\\\hat{\boldsymbol{H}}'\in\mathbb{R}^{n\times d}=ELU([\hat{\boldsymbol{H}};\hat{\boldsymbol{q}}]\hat{\boldsymbol{W}}),\end{cases}\tag{11}$$

其中 $\hat{\boldsymbol{W}}\in\mathbb{R}^{2d\times d}$ 为线性投影矩阵. 最后，为了能够保留完整的视觉语义线索，防止弱语义线索的丢失，将不同推理步骤获取的视觉语义线索进行汇总，来表示视频特征序列 $F=\left\{\boldsymbol{f}_j\middle|\boldsymbol{f}_j\in\mathbb{R}^d\right\}_{j=1}^{X}$ 中与问题强相关的浓缩视觉线索表达 H：

$$\boldsymbol{H}=\hat{\boldsymbol{h}}+\hat{\boldsymbol{h}}_{\mathrm{k}}+\hat{\boldsymbol{H}}'.\tag{12}$$

上述特征表达不仅涵盖了充足且准确的多模态先验知识，还包含了对多模态信息的深层次理解，为答案解码提供了丰富的视觉语义线索. 高度浓缩的视觉语义限线索也为视频问答模型获取更高层次的视觉语义线索提供了便利.

2.4 基于多模态知识主动学习的多层次视频问答网络

2.3 节提出的 KAFR 模型能够在数据融合与推理过程中主动完善多模态信息的深度理解，还能通过思维的聚焦学习，减少视频中需要理解的多模态信息，降低数据融合与推理过程对于先验知识的依赖，改善特征提取不足所带来的挑战. 接着我们将 KAFR 按照视频的层次结构，如图像、视频片段等进行排列，搭建了静态外观与语言、动态信息与语言和视觉目标与语言等多层次视频问答网络，进一步从视频中理解完整的故事情节，获取视频层级所提供的多层次视觉语义线索，为视频问答提供更加准确的答案预测.

这些多层次视频问答网络分别以全局静态信息 $V^{\mathrm{app}}=\left\{\boldsymbol{V}_i^{\mathrm{app}}\right\}_{i=1}^{N}$、全局动态信息 $V^{\mathrm{mot}}=\left\{\boldsymbol{V}_i^{\mathrm{mot}}\right\}_{i=1}^{N}$ 和视频的细节信息 $V^{\mathrm{ob}}=\left\{\boldsymbol{V}_i^{\mathrm{ob}}\right\}_{i=1}^{N}$ 作为视觉信息输入，通过多层次的 KAFR 模型的融合与推理计算，自底向上地从视频的不同片段、不同图像获取与问题相关的浓缩视觉线索表达.

我们按照视频层次结构排列了 L 层的 KAFR 模型，每一层以上一层 KAFR 模型的特征输出集合 $\{\boldsymbol{H}_i^{l-1}\}_{i=1}^{ns}$ 和问题 q^{or} 作为输入，并通过 KAFR 模型的数据融合与推理计算来获取更高层次的浓缩语义表达.

$$\begin{cases}\boldsymbol{H}_j^{l}\in\mathbb{R}^{d}=KAFR^{l}(\boldsymbol{V}_i,q^{\mathrm{or}}),l=1,\\\boldsymbol{H}_j^{l}\in\mathbb{R}^{d}=KAFR^{l}(\{\boldsymbol{H}_j^{l-1}\}_{j=1}^{ns},q^{\mathrm{or}}),l>1,\end{cases}\tag{13}$$

其中 $l\in\{1,2,\cdots,L\}$，ns 表示上一层所处理的视频特征集合数量，$\boldsymbol{V}_i=(\boldsymbol{V}_i^{\mathrm{app}},\boldsymbol{V}_i^{\mathrm{mot}},\boldsymbol{V}_i^{\mathrm{ob}})$. 经过上述过程，我们能够将数据融合与推理渗透到视频的各个层次，自底向上地从视频中获得不同层次的视觉语义线索，为视频问答模型最终的答案解码提供高度凝练且完整的静态语义线索 $\boldsymbol{H}_{\mathrm{app}}$，动作语义线索 $\boldsymbol{H}_{\mathrm{mot}}$ 以及包含了每个与问题相关联的视觉目标、复杂运动轨迹和多维时空交互的视觉目标语义线索 $\boldsymbol{H}_{\mathrm{ob}}$.

我们在后续的实验中对于所提出方案中的网络结构的合理性以及多层次设计方案进行了严格的消融实验（见 3.4.1 节），实验结果表明，多层次网络设计的问答性能优于单层次的网络设计，证实了多层次结构网络结构的优越性.

2.5 答案解码

本节针对多项选择任务、开放性任务和重复计数任务等不同类型的视频问题设计了不同的解码器，使视频问答模型能够应对不同类型任务的挑战.

针对多项选择任务，本文以 $\boldsymbol{H}_{\mathrm{ob}}$，$\boldsymbol{H}_{\mathrm{app}}$，$\boldsymbol{H}_{\mathrm{mot}}$，$\bar{\boldsymbol{q}}$，$\{\boldsymbol{a}_i\}_{i=1}^{N_{\mathrm{ca}}}$ 作为输入. 利用式（14）完成答案 $\boldsymbol{\delta}_i$ 的解码.

$$\boldsymbol{\delta}_i=(ELU([\boldsymbol{H}_{\mathrm{ob}};\boldsymbol{H}_{\mathrm{app}};\boldsymbol{H}_{\mathrm{mot}};\bar{\boldsymbol{q}}\boldsymbol{W}_{\mathrm{q}};\boldsymbol{a}_i\boldsymbol{W}_{\mathrm{a}}]\boldsymbol{W}_{\mathrm{m}}))\boldsymbol{W}_{\mathrm{multi}},\tag{14}$$

其中 $\boldsymbol{W}_{\mathrm{multi}}\in\mathbb{R}^{N_{ca}\times d}$，$\boldsymbol{W}_{\mathrm{m}}\in\mathbb{R}^{5d\times d}$，$\boldsymbol{W}_{\mathrm{q}}\in\mathbb{R}^{d\times d}$，$\boldsymbol{W}_{\mathrm{a}}\in\mathbb{R}^{d\times d}$ 是不同的线性投影矩阵，将所有的输入映射到最终的语义空间内，推导出最终的问题答案 a. 最后选择得分最高的作为预测答案.

$$a=\arg\max_i(\boldsymbol{\delta}_i).\tag{15}$$

在这类视频问答中，交叉熵损失函数被用于网络模型的优化.

针对开放性任务，特征 $\boldsymbol{H}_{\mathrm{ob}}$，$\boldsymbol{H}_{\mathrm{app}}$，$\boldsymbol{H}_{\mathrm{mot}}$，$\bar{\boldsymbol{q}}$ 作为输

入, 式(16)被用于得到每个候选答案的最终得分$\boldsymbol{\delta}_{\text{open}} \in \mathbb{R}^{N_{\text{open}}}$,

$$\boldsymbol{\delta}_{\text{open}} = softmax((ELU([\boldsymbol{H}_{\text{o}};\boldsymbol{H}_{\text{app}};\boldsymbol{H}_{\text{mot}};\bar{\boldsymbol{q}}\boldsymbol{W}_{\text{q}}]\boldsymbol{W}_{\text{open}'}))\boldsymbol{W}_{\text{open}}), \tag{16}$$

其中$\boldsymbol{W}_{\text{open}} \in \mathbb{R}^{d\times N_{\text{open}}}$, $\boldsymbol{W}_{\text{open}'} \in \mathbb{R}^{4d\times d}$是不同的线性投影矩阵, N_{open}表示答案空间$|\mathcal{A}|$的长度. 最后我们选择得分最高的答案作为预测答案.

$$a = \arg\max_{a^*}(\boldsymbol{\delta}_{\text{open}}). \tag{17}$$

在这类视频问答任务中, 交叉熵损失函数被用于优化网络模型.

针对重复计数任务, 线性回归函数被用来预测整数值的答案$\delta_{\text{count}} \in \mathbb{R}^1$:

$$\delta_{\text{count}} = (ELU([\boldsymbol{H}_{\text{ob}};\boldsymbol{H}_{\text{app}};\boldsymbol{H}_{\text{mot}};\bar{\boldsymbol{q}}\boldsymbol{W}_{\text{q}}]\boldsymbol{W}_{\text{c}}))\boldsymbol{W}_{\text{count}}, \tag{18}$$

其中$\boldsymbol{W}_{\text{count}} \in \mathbb{R}^{d\times 1}$, $\boldsymbol{W}_{\text{c}} \in \mathbb{R}^{4d\times d}$是不同的线性投影矩阵. 在这类视频问答任务中, 均方误差损失被用于优化网络模型.

3 实验结果及分析

为了能够客观公正地评估本文的方法, 我们选取了3个现阶段广泛使用且极具挑战性的视频问答数据集进行了实验测试.

3.1 数据集介绍

1) TGIF-QA[14]. 该数据集包含有16.5万个问题对, 按照问题的独特属性将数据集划分为4类子任务: Repeating Action, Transition, Repeating counting, Frame QA.

2) MSVD-QA[15]. 该数据在1 970个视频片段中标注了5万个开放性视频问题对, 其中训练集、验证集、测试集中分别有3.09万、0.64万、1.3万个问题对, 答案空间的长度为1 852.

3) MSRVTT-QA[16]. 该数据在10万个视频片段中标注了24.3万个问题对, 其中训练集、验证集、测试集中分别有15.8万、1.22万、7.28万个问题对, 答案空间的长度为4 000. 相较于前2种视频问答数据集, 该数据集拥有10~30 s的视频序列, 这使得视频内的场景更加复杂, 对数据融合与推理能力提出了更高的挑战.

3.2 实施细节

本文方法是基于Pytorch深度学习框架实现. 在实验设置中, 视频片段数N=8, 并在每个片段中采样, T=16帧表示该片段的内容, 在每一帧图像中提取K=10个视觉目标特征. 针对每一个问题, 关键字数n=3. 对于外观特征、运动特征和目标特征, 我们分别使用了L=2, L=2, L=1层的KAFR模型. 设置在每一个模块内的多头注意力网络的头数均为H=8, 设置特征维度d=512. 在训练过程中, 模型被训练25轮. Adam优化器被用来优化模型参数, 数据的批大小设置为32, 学习率设置为0.5E−4.

3.3 评价标准

为了便于与现有方法进行比较, 我们使用均方误差(mean square error, MSE)对TGIF-QA数据集中的Repeating counting任务进行评估. MSE值越小, 性能越好. 对于数据集的其他任务, 采用准确率来评估模型的性能. 准确率越高, 性能越好.

3.4 消融实验

为了验证本文所做出的贡献, 我们在所提出的基于多模态知识主动学习的视频问答方案上进行了广泛的消融实验, 以验证网络结构及其模块的合理性、显性细节特征提取的有效性和超参数的合理性.

3.4.1 网络结构及其模块的合理性

在本文中, KAFR模型按照视频的层次结构如图像、视频片段等构建了不同层次的数据融合与推理计算网络, 以获取分散在视频内不同层次的浓缩视觉语义线索. 为了验证这种网络结构的合理性, 我们在MSRVTT-QA和MSVD-QA中比较了网络结构对于性能的影响. 从表1可以看出, 当使用单个KAFR模型时, 算法的性能有明显的下降. 而多层次的网络设计展现了优异的问答性能, 这展示了多层次结构网络结构的优越性.

Table 1 Verify the Rationality of the Network Structures and Their Modules

表1 验证网络结构及其模块的合理性 %

多模态特征提取	MSVD-QA	MSRVTT-QA
w/o 目标对齐	42.2	38.1
w/o 静态特征	41.2	37.6
w/o 动作特征	41.5	37.8
w/o 视觉目标特征	39.8	37.3
only 静态特征	38.6	36.5
only 动作特征	36.5	35.9
only 视觉目标特征	39.5	37.2

注: w/o表示去除相应模块, only表示只使用当前特征.

除此之外, 本节还在每一个KAFR模型中, 尝试引入主动学习和思维聚焦来帮助视频问答模型应对先验知识不足的问题, 进一步深化模型对多模态信息的理解, 并收集归纳与问题强相关的视觉语义线

索. 为了验证该模型的有效性，我们在表 1 进行了详细的消融实验. 可以看出，KAFR 的所有模块都很重要，删除其中任何一个都会降低相应的性能. 值得注意的是，传统的数据融合与推理过程缺乏思维聚焦和主动学习，其性能明显低于 KAFR，这有力证明了 KAFR 的优越性，并支持了本文对于视频问答存在先验知识不足的猜想. 同时，这也进一步表明在数据融合与推理过程中，增加主动学习能力和思维聚焦能力是提升问答性能和增强视频理解能力的有效策略. 此外，通过对逻辑思维过程顺序的消融实验结果分析可以发现，主动学习能够为聚焦学习提供正确的多模态语义理解，指导思维聚焦过程，这进一步证实了本文所设计的数据融合与推理模型的合理性.

3.4.2 显性细节特征提取有效性验证

在本节中，显性细节特征提取模块提取了视觉目标、静态和动作等多模态特征信息，以期望为视频问答提供完整的视觉语义线索. 为了验证不同模态特征对性能的影响，本文比较了在 MSRVTT-QA 和 MSVD-QA 中以不同模态信息作为输入对性能的影响. 从表 2 可以看出，所提出的模型都能够有效地对每一种模态信息进行数据融合和推理计算，证明了本文提出的显性细节特征提取方法的有效性. 同时，通过进一步比较可以发现，去掉视觉目标的对齐会导致性能下降，这也证明了本文提出的显性视频细节特征提取方法能够有效地减少视觉目标混乱所造成的性能损失，完善视频的特征表达，提高问答性能.

Table 2 Verify the Effectiveness of Explicit Detail Feature Extraction

表 2 验证显性细节特征提取的有效性 %

	网络组成	MSVD-QA	MSRVTT-QA
层次结构	1 层 KAFR 模块	40.8	37.7
	2 层 KAFR 模块	43.1	38.7
	3 层 KAFR 模块	42.9	38.8
KAFR 模块的组成	传统数据融合与推理过程	39.8	36.8
	w/o 时序表示学习与推理	38.9	35.6
	w/o 多模态表示再学习	40.8	37.6
	w/o 聚焦表示学习	42.4	38.1
	w/o 汇总表示学习	41.6	37.9
	聚焦思维→主动学习	42.3	38.1
	主动学习→聚焦思维	43.1	38.7

注：w/o 表示去掉相应特征.

3.4.3 超参数合理性验证

本节使用了 K=10 个的视觉目标特征来描述视频的细节信息. 为了验证这种设置的合理性，我们在 MSVD-QA 数据集上比较了不同 K 值对性能和模型参数的影响. 从图 3 中可以看出，性能与 K 值不存在正相关关系，并且在 K=10 处获取了最优的问答性能. 这是因为过多的目标采样导致视频细节冗余，影响了正常的数据融合与推理计算，从而降低了性能. 同时，KAFR 与现阶段流行的模型 HCRN 的比较结果可以看出，KAFR 虽然参数增加了 2×10^6，但性能提升明显，这证明了 KAFR 设计的合理性.

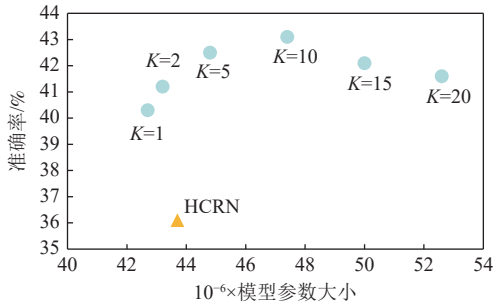


Fig. 3 Verify the rationality of K value

图 3 验证 K 值的合理性

除此之外，为了实现跨模态的语义融合，本文使用了大量的映射矩阵. 为了验证投影矩阵维度 d=512 的合理性，我们在 MSVD-QA 比较了不同 d 值对性能和网络参数的影响. 结果如图 4 所示，d=512 时的问答性能优于 d=256 或（d=1 024）时的问答性能. 这是由于高维度的特征投影（d=1 024）虽然有助于建立跨模态语义的稳定映射关系，但是也带来冗余的网络参数，从而导致网络难以收敛，影响了问答的性能. 而低维度的特征映射（d=256）无法提供稳定的语义的稳定映射关系，影响了问答的性能. 因此，我们所选取的投影矩阵参数设置是合理的.

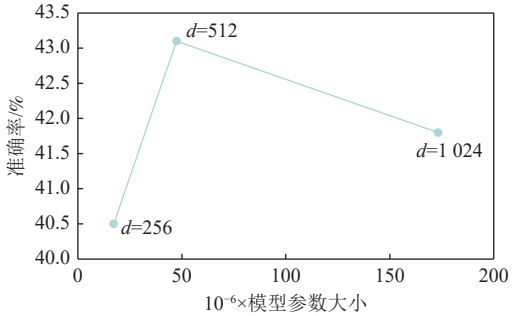


Fig. 4 Verify the rationality of d value

图 4 验证 d 值的合理性

3.5 性能比较

为了更好地评估本文的工作，我们将本文提出的 KAFR 与近几年的算法进行比较.

1)L-GCN[19]. 该模型通过位置感知图来构建视频问答任务中检测到的对象之间的关系，将对象的位置特征融入列图和构建中.

2)HGA[21]. 该模型设计了一个深度异构图对齐网络，从表示、融合、对齐和推理4个步骤来推断答案.

3)HCRN[12]. 该模型是一种条件关系网络，作为构建块来构建更复杂的视频表示和推理结构.

4)HOSTR[9]. 该模型是一种面向视频内对象的视频问答方法，利用位置信息对视频内实体关系进行建模，获取细粒度的时空表达和逻辑推理能力.

5)MASN[20]. 该模型是一种运动外观协同网络，以融合和创建运动外观特征与静态外观特征之间的协同融合.

6)HRNAT[31]. 该模型是一个带有辅助任务的分层表示网络，用于学习多层次表示并获得句法感知的视频字幕.

7)DualVGR[11]. 该模型是一种用于视频问答的双视觉图推理单元，该单元通过迭代堆叠来模拟视频片段之间与问题相关的丰富时空交互.

8)PKOL[25]. 该模型是一种面向视频问答的先验知识探索和目标敏感学习方法，探索了先验知识对数据融合与推理性能的影响.

9)ClipBERT[27]. 该模型是一种用于端到端的视频问答框架，在训练过程中使用图像-文本的预训练.

10)CoMVT[28]. 该模型是一种基于双流多模态视频transformer的数据融合与推理框架，它能有效地联合处理文本中的单词和视觉对象，利用网络中的在线教学视频数据集进了预训练.

KAFR与多个视频问答数据集上最先进的方法进行比较，结果如表3所示. KAFA在所有任务中都优于现有未经预训练的方法. 具体来说，在Action, Transition, FrameQA, Count, MSVD-QA, MSRVTT-QA测试中，相较于未经预训练的模型，KAFR分别提高了0.8%, 2.7%, 1.3%, 0.04%, 2.0%, 1.8%. 而相较于那些预训练模型，KAFA也能获取与之相匹配的性能，甚至除MSRVTT-QA测试之外，都有性能的提升. 这说明KAFA能够获取更为准确的视频表达，而数据融合与推理模型能够通过逻辑推理计算过程中的思维聚焦与主动学习，有效地完善了视频问答系统的先验知识，降低了对先验知识的依赖，获取了更为合理、充分的视觉语义线索和高性能的视频问答能力.

3.6 结果可视化

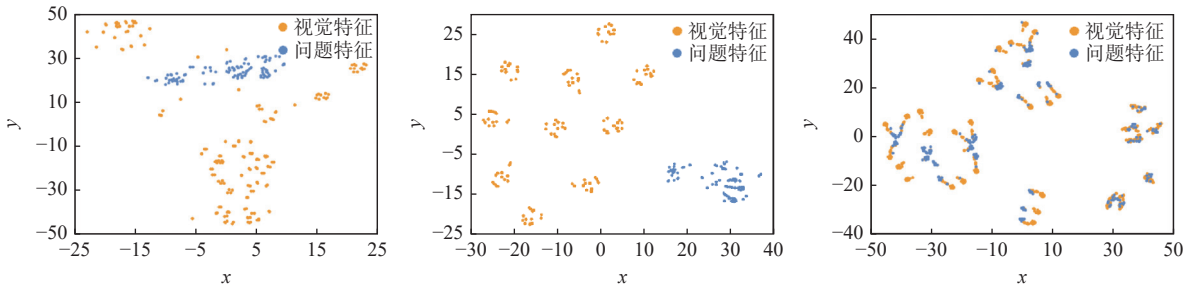
为了更好地理解我们在数据融合与推理方面所做出的贡献，本节在图5中给出了一些特征分布的可视化结果. 从图5(a)中可以看出，视觉特征与问题特征序列非均匀地分布在原始特征空间内，存在着明显的语义鸿沟问题. 而在图5(b)中，视觉特征和问题特征通过时序表示学习与推理计算后，特征空间缩小了近50%，视觉特征与问题特征在空间中相互接近，但语义鸿沟依旧存在，多模态特征依旧分布在不同的子空间，阻碍了数据融合与推理的进行. 在图5(c)中，视觉特征与问题特征通过多模态表示再学习的自主学习过程后，补充后的子问题与填充后的视觉信息能够彼此纠缠，分布于相同的语义空间内，

Table 3 Comparison of Our Method with the Most Advanced Methods on Multiple Video Question Answering Datasets

表3 本文方法与多个视频问答数据集上最先进的方法的比较 %

模型	预训练	TGIF-QA				MSVD-QA ↑	MSRVTT-QA ↑
		Action ↑	Transition ↑	FrameQA ↑	Count ↓		
L-GCN[19]		74.3	81.1	56.3	3.95	34.3	
HGA[21]		75.4	81.0	55.1	4.09	34.7	35.5
HCRN[12]		75.0	81.4	55.9	3.82	36.1	35.6
HOSTR[9]		75.6	83.0	58.2	3.65	39.4	35.9
MASN[20]		84.4	87.4	59.5	3.75	38.0	35.2
HRNAT[31]						38.2	34.9
DualVGR[11]						39.0	35.5
PKOL[25]		74.6	82.3	61.8	3.67	41.1	36.9
ClipBERT[27]	√	82.8	87.8	60.3			37.4
CoMVT[28]	√					42.6	**39.5**
KAFR（本文方法）		**85.2**	**90.1**	**63.1**	**3.61**	**43.1**	38.7

注：Count以均方误差为评价指标，其他以准确率为评价指标. 加粗数字表示性能最优. ↑表示越大越好，↓表示越小越好.



(a) 原始特征空间 (b) 时序表达与推理后的特征空间 (c) 多模态表示再学习后的特征空间

Fig. 5 Visual t-SNE graph for multimodal embedding distribution

图 5 用于多模态嵌入分布的可视化 t-SNE 图

有效克服了语义鸿沟的问题，为接下来的数据融合与推理计算提供了有利的条件. 上述结果表明，KAFR 能够很好地利用已经掌握的视觉内容填补对多模态特征的深度理解，减小了语义鸿沟对跨模态数据融合与推理计算的影响，提升了模型的问答性能.

接着，我们还给出了一些视频问答预测结果的演示，如图 6 所示，包括 3 个视频问答问题. 在图 6(a)中，KAFR 通过对视觉细节的特征提取与视觉目标的对齐，深入理解了视频场景内所发生的故事情节“竞争(race)”，而缺少视觉目标对齐的结果只能浅显地理解每个所做的动作“跑步(run)”. 在图 6(b)中，缺少视觉目标对齐的结果缺少对视频景深的理解，只能片面地理解 2 维平面的“behind”，而将“lady”也考虑在答案中. 而通过视觉目标运动信息对静态信息的纠正，修正了模型对于“lady”位置的理解，使得 KAFR 能够准确预测出了答案“two”. 在图 6(c)中，KAFR 只理解了由人、马、植被和草地所组成的复杂场景，未能准确地识别出沙地和山峰等复杂要素，致使模型将深层次的复合场景语义“desert”被错误认定为了“yard”.

(a) 问题: What are several runners doing?
KAFR w/o 目标对齐: run ×
KAFR: race √
正确答案: race

(b) 问题: How many people are hiding behind a bush?
KAFR w/o 目标对齐: three ×
KAFR: two √
正确答案: two

(c) 问题: What did the cowboy ride his horse in?
KAFR w/o 目标对齐: yard ×
KAFR: yard ×
正确答案: desert

Fig. 6 Video question answering result demonstration

图 6 视频问答结果演示

最后，还展示了思维聚焦的可视化演示结果，以 2 个视频问答问题为例，结果如图 7 所示. 在图 7(a)中，缺少思维聚焦功能的注意力热图缺少焦点. 但经过对关键信息“tears, piece, paper”的定位后，逻辑推理聚焦到与问题密切相关的橙色虚线标注视频片段，准确地找出了包含正确答案的视觉线索，正确预测了答案“man”. 而在图 7(b)中，KAFR 通过定位关键信息“woman, scoop, ice cream”，准确找出了与问题紧密相关的 2 个紫色虚线标注视频片段，正确预测了答案“two”. 以上结果表明，KAFR 通过思维聚焦能够缩小特征空间，减少需要理解的多模态信息，改善了对先验知识的依赖，从而提高了算法的性能.

4 总 结

本文针对视频问答任务中视频细节提取不足和模型主动学习能力不足的问题，提出了一种基于多模态知识主动学习的视频问答方案 KAFR. 在该方案中，显性细节表达提取模块首先通过将视频的静态细节表达推广到动态细节描述，以防止由于视频细节内容的缺失导致的错误因果关系，建立了更为准确的视频模型. 接着，KAFR 模型通过多模态信息深度理解的自我完善以及思维的聚焦，为数据融合与推理计算提供更准确和精炼的多模态特征表达. 在多个公开视频问答数据集上的实验结果表明：显性

Fig. 7 Visualization of the thinking focus process

图 7 思维聚焦过程的可视化

细节表达提取模块能够有效获取视频的细节表达和更为完整的视频多模态表达. 同时, 带有自主学习和思维聚焦能力的 KAFR 模型能够有效缓解特征提取阶段先验知识不足的问题, 从而提高了模型的性能.

结合人工智能技术的视频问答研究不仅具有重要的理论研究意义, 更重要的是具有广泛的应用价值. 通过视频问答技术与机器人技术的结合, 未来机器人将能够更好地理解人类的语言和意图, 并通过观察和分析视频内容来获取更多的环境信息, 在我们的日常生活中发挥更加重要的作用. 特别是在未来的数字家庭和智慧社区中, 这些配备视频问答技术的机器人将成为我们生活中的智能伙伴, 提供个性化、便捷和智能化的服务和支持.

作者贡献声明: 刘明阳提出算法思路, 完成实验并撰写论文; 王若梅提出指导意见; 周凡参与论文校对和实验方案指导; 林格提出指导意见和审核论文.

参 考 文 献

[1] Yu Jun, Wang Liang, Yu Zhou. Research on visual question answering techniques[J]. Journal of Computer Research and Development, 2018, 55(9): 1946−1958 (in Chinese)
(俞俊, 汪亮, 余宙. 视觉问答技术研究 [J]. 计算机研究与发展, 2018, 55(9): 1946−1958)

[2] Zhang Lu, Cao Feng, Liang Xinyan, et al. Cross-modal retrieval with correlation feature propagation[J]. Journal of Computer Research and Development, 2022, 59(9): 1993−2002 (in Chinese)
(张璐, 曹峰, 梁新彦, 等. 基于关联特征传播的跨模态检索 [J]. 计算机研究与发展, 2022, 59(9): 1993−2002)

[3] Li Zhixin, Wei Haiyang, Zhang Canlong, et al. Research progress on image captioning[J]. Journal of Computer Research and Development, 2021, 58(9): 1951−1974 (in Chinese)
(李志欣, 魏海洋, 张灿龙, 等. 图像描述生成研究进展 [J]. 计算机研究与发展, 2021, 58(9): 1951−1974)

[4] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]// Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770–778

[5] Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imageNet[C]// Proc of the 36th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6546–6555

[6] Peter A, He Xiaodong, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]// Proc of the 36th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6077–6086

[7] Lu Jiasen, Yang Jianwei, Batra D, et al. Hierarchical question-image co-attention for visual question answering[C]// Proc of the 30th Int Conf on Neural Information Proc Systems. New York: ACM, 2016: 289–297

[8] Gao Jiyang, Ge Runzhou, Chen Kan, et al. Motion appearance co-memory networks for video question answering[C]// Proc of the 36th

IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6576–6585

[9] Dang L H, Le T, Le V, et al. Hierarchical object-oriented spatio-temporal reasoning for video question answering[C]// Proc of the 30th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2021: 636–642

[10] Jiang Jianwen, Chen Ziqiang, Lin Haojie, et al. Divide and conquer: Question-guided spatio-temporal conrmual attention for video question answering[C]// Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 11101–11108

[11] Wang Jianyu, Bao Bingkun, Xu Changsheng. DualVGR: A dual-visual graph reasoning unit for video question answering[J]. IEEE Transactions on Multimedia, 2022, 24: 3369−3380

[12] Le T M, Le V, Venkatesh S, et al. Hierarchical conditional relation networks for video question answering[C]// Proc of the 38th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 9969–9978

[13] Liu Fei, Liu Jing, Wang Weining, et al. HAIR: Hierarchical visual-semantic relational reasoning for video question answering[C]// Proc of the 18th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 1678–1687

[14] Jang Y, Song Y, Yu Y, et al. TGIF-QA: Toward spatiotemporal reasoning in visual question answering[C]// Proc of the 33rd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 1359–1367

[15] Xu Dejing, Zhao Zhou, Xiao Jun, et al. Video question answering via gradually refined attention over appearance and motion[C]// Proc of the 25th ACM Int Conf on Multimedia. New York: ACM, 2017: 1645–1653

[16] Xu Jun, Mei Tao, Yao Ting, et al. MSRVTT: A large video description dataset for bridging video and language[C]// Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 5288–5296

[17] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint, arXiv: 1409.1556v6, 2015

[18] Zhao Zhou, Zhang Zhu, Xiao Shuwen, et al. Open-ended long-form video question answering via adaptive hierarchical reinforced networks[C]// Proc of the 27th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2018: 3683–3689

[19] Huang Deng, Chen Peihao, Zeng Runhao, et al. Location-aware graph convolutional networks for video question answering[C]// Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 11021–11028

[20] Seo A, Kang G, Park J, et al. Attend what you need: Motion-appearance synergistic networks for video question answering[C]// Proc of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing. Stroudsburg, PA: ACL, 2021: 6167–6177

[21] Jiang Pin, Han Yahong. Reasoning with heterogeneous graph alignment for video question answering[C]// Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 11109–11116

[22] Park J, Lee J, Sohn K. Bridge to answer: Structure-aware graph interaction network for video question answering[C]// Proc of the 39th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 15526–15535

[23] Russakovsky O, Deng Jia, Su Hao, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211−252

[24] Will K, Carreira J, Simonyan K, et al. The Kinetics human action video dataset[J]. arXiv preprint, arXiv: 1705.06950, 2017

[25] Zeng Pengpeng, Zhang Haonan, Gao Lianli, et al. Video question answering with prior knowledge and object-sensitive learning[J]. IEEE Transactions on Image Processing, 2022, 31: 5936−5948

[26] Chen Xinlei, Fang Hao, Lin Tsung-Yi, et al. Microsoft COCO captions: Data collection and evaluation server[J]. arXiv preprint, arXiv: 1504.00325, 2015

[27] Lei Jie, Li Linjie, Zhou Luowei, et al. Less is more: ClipBERT for video-and-language learning via sparse sampling[C]// Proc of the 39th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 7331–7341

[28] Seo P H, Nagrani A, Schmid C. Look before you speak: Visually conrmualized utterances[C]// Proc of the 39th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 16877–16887

[29] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// Proc of the 14th Conf of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 4171–4186

[30] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]// Proc of the 31st Int Conf on Neural Information Processing Systems. New York: ACM, 2017: 5998–6008

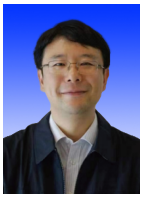
[31] Gao Lianli, Lei Yu, Zeng Pengpeng, et al. Hierarchical representation network with auxiliary tasks for video captioning and video question answering[J]. IEEE Transactions on Image Processing, 2022, 31: 202−215

Liu Mingyang, born in 1994. PhD. His main research interests include video comprehension, video question answering, and machine learning.

刘明阳, 1994 年生. 博士. 主要研究方向为视频理解、视频问答、机器学习.

Wang Ruomei, born in 1961. PhD, professor, PhD supervisor. Her main research interests include computer graphics, computer-aided design, and multimedia processing.

王若梅, 1961 年生. 博士, 教授, 博士生导师. 主要研究方向为计算机图形学、计算机辅助设计、多媒体处理.



Zhou Fan, born in 1976. PhD, professor, PhD supervisor. His main research interests include intelligent home systems, intelligent question answering, and the processing of cross-media data.

周　凡, 1976年生. 博士，教授，博士生导师. 主要研究方向为智能家居系统、智能问答、跨媒体数据处理.

Lin Ge, born in 1980. PhD, associate professor. His main research interests include intelligent home systems and artificial intelligence.

林　格, 1980年生. 博士，副研究员. 主要研究方向为智慧家庭、人工智能.