计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(6): 1563−1573, 2026

# 基于大语言模型伪标签生成的长视频问答关键帧选择增强方法

宋子杰[1] 关 鑫[2] 胡珍珍[2] 李 佳[2] 洪日昌[2]
[1](安徽大学大数据与统计学院 合肥 230601)
[2](合肥工业大学计算机与信息学院 合肥 230601)
(zjsong@ahu.edu.cn)

# Enhancing Keyframe Selection Method via LLM-Generated Pseudo-Labels for Long-Form Video Question Answering

Song Zijie[1], Guan Xin[2], Hu Zhenzhen[2], Li Jia[2], and Hong Richang[2]
[1](*School of Big Data and Statistics*, *Anhui University*, *Hefei* 230601)
[2](*School of Computer Science and Information Engineering*, *Hefei University of Technology*, *Hefei* 230601)

**Abstract** Keyframe selection is an important method for long-form video question answering, as it accurately identifies key content from redundant information and establishes an interpretable reasoning path. However, existing keyframe selection methods face challenges with insufficient semantic sensitivity during end-to-end training, leading to the introduction of a significant amount of irrelevant frame noise, which affects the model's accuracy and interpretability. To address this, we propose a pseudo-labels guided keyframe selection (PGKS) model. This model first utilizes large language models (LLMs) to semantically integrate questions and answers, generating global descriptive text. It then employs a multimodal alignment model to compute semantic similarity scores between the description and video sampled frames, thereby constructing frame-level pseudo-labels. By guiding the calculation of frame-level similarity scores with these pseudo-labels, the model effectively suppresses irrelevant frame noise, enhancing both the accuracy and interpretability of keyframe selection. The decoupling between the utilization of LLMs and the training of the backbone network is architecturally designed to provide a cost-effective solution for leveraging LLMs in resource-constrained scenarios. Furthermore, our method balances differentiable and non-differentiable sorting results, and introduces a sliding time window mechanism to further improve the model's understanding of temporal relationships in videos. Experimental results demonstrate that PGKS model achieves an accuracy of 62.67% on the long-form video question answering dataset NExT-QA, outperforming existing methods of comparable size and improving by 8.51% over the baseline without pseudo-label guidance.

**Key words** video question answering; keyframe selection; pseudo-labels; large language model; similarity score

**摘 要** 关键帧选择作为长视频问答任务的重要方法,能够从冗余信息中准确定位关键内容,并构建可解释的推理路径。然而,现有的关键帧选择方法在端到端训练中存在语义敏感性不足的问题,导致关键帧选择过程中引入大量无关帧噪声,从而影响模型的准确性和可解释性。为此,提出了一种基于伪标签引导的关键帧选择(pseudo-labels guided keyframe selection, PGKS)方法。该方法首先利用大语言模型对问题与答案进行语义融合,生成全局描述性文本,并结合多模态对齐模型,计算描述与视频采样帧之间的语义相似度分数,从而构建帧级伪标签。然后,通过伪标签引导帧级相似度分数的计算,有效抑制了无关帧

**收稿日期**: 2025-07-20; **修回日期**: 2025-12-15
**基金项目**: 国家自然科学基金项目(62172138, 62202139, U23B2031)
The work was supported by the National Natural Science Foundation of China (62172138, 62202139, U23B2031).
**通信作者**: 胡珍珍(zzhu@hfut.edu.cn)

噪声，增强了关键帧选择的准确性与可解释性。这种大语言模型的利用和训练分离的方法，为资源受限场景中对大语言模型的利用提供了一种低成本的方案。此外，关键帧选择方法平衡了可微性与硬排序的精准性，并通过引入滑动时间窗口机制，进一步提升了模型对视频时序关系的理解能力。实验结果表明，PGKS 模型在长视频问答数据集 NExT-QA 上达到了 62.67% 的准确率，优于现有同体量模型的方法，并较无伪标签引导的基线提升了 8.51 个百分点。

**关键词** 视频问答；关键帧选择；伪标签；大语言模型；相似度分数

**中图法分类号** TP391

**DOI:** 10.7544/issn1000-1239.202550537 **CSTR:** 32373.14.issn1000-1239.202550537

长视频问答(long-form video question answering)[1-4]作为一种强调跨模态推理的任务，要求模型具备较高的时序理解能力，能够从冗余信息中准确定位关键内容，并建立可解释的推理路径。近些年，关键帧选择[1,4-5]作为一种有效的解决视频冗余性问题和提升问答准确性的方法被广泛地研究。关键帧的选择通常依赖于对视频帧的打分和排序。然而，硬排序过程的不可微性使得基于硬排序的关键帧筛选无法实现端到端的训练，进而难以达到学习和优化的目标。可微排序算法[6-8]的引入，使得端到端的关键帧选择成为可能，然而，这类方法的基本原理是将视频帧按其对应的帧分数加权生成关键帧，但这样的松弛化操作会在关键帧的生成过程中引入大量无关帧的噪声，导致选择出来的关键帧偏离真实分布，且无法匹配人类对于视频帧的直觉感知，如图 1 所示。因此，在维持排序过程中的可微性和减弱可微排序过程中引入的噪声之间维持平衡，成为一项严峻的挑战。Hu 等人[9]提出了一种基于双流关键帧选择增强的视频问答方法，在一定程度上缓解了可微排序引入的噪声对模型性能的影响。

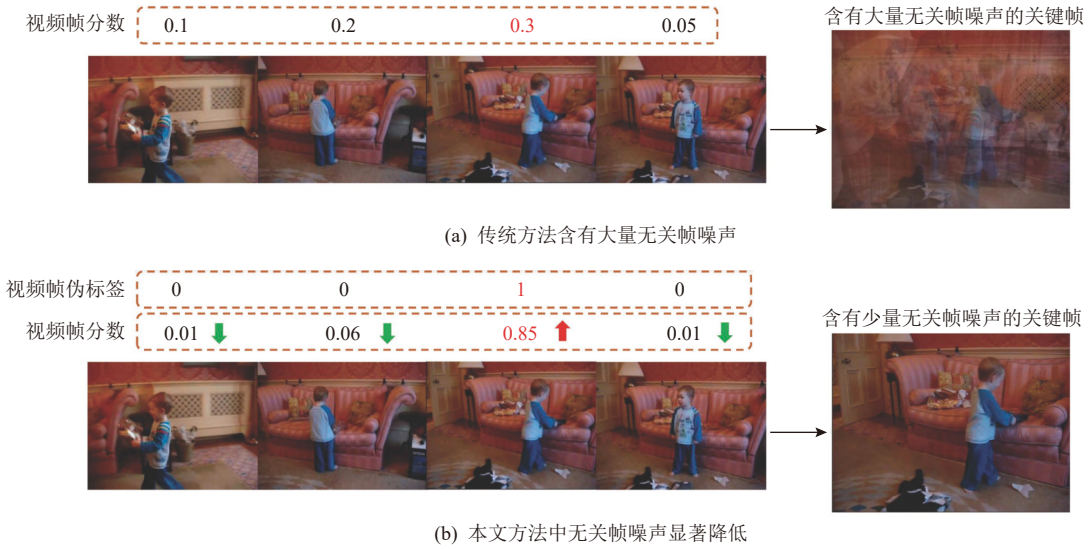



Fig. 1 Keyframes with a large amount of irrelevant frame noise vs keyframes with a small amount of irrelevant frame noise

图 1 含有大量无关帧噪声的关键帧对比含有少量无关帧噪声的关键帧

端到端的关键帧选择虽然提供了可学习的训练手段，但其性能完全依赖于视频帧评分的有效性和合理性。目前，模型的训练仅依赖于问答标注的监督信号，这使得该方案在同体量的视频问答模型[10-11]比较中并未展现出明显的优势。此外，采用类似视频片段定位[12]的人工标注成本巨大，且对于问答任务和冗余视频帧，人为标注面临较高的难度和不确定性。Fang 等人[13]创新性地从博弈论视角重新审视视频时序定位，将视频帧和查询词建模为博弈参与者，通过多变量合作博弈理论量化帧词协作关系实现多粒度跨模态对齐和交互。针对弱监督场景下标注数据稀缺的挑战，Zhao 等人[14]提出视频-文本提示方法结合对比学习，在视频对象管状区域绘制视觉标记作为提示，并插入对应文本提示，在保留全局上

下文的同时增强正确候选的区分度。因此，如何引导和约束模型选择准确且合理的关键帧，成为建立有效视频问答推理路径的重中之重。

另一方面，视频预训练的研究工作也取得了突破性进展。通过大规模视频数据集（如 HowTo100M[15]）进行视频语言预训练表示，使得基于 Transformer 包括视频问答模型[16]获得的一系列视频帧编码不再是静态的视觉图像信息，而是具有关联的视觉特征组。这种预训练视频模型在强大算力和庞大的数据量支持下确实涌现出了强大的表征能力[17]，但是隐式的视觉编码并不能解读具有可解释的推理逻辑。从视频时序性[1,5,18]来说，如何体现模型在表征的过程中捕捉到了时序信息；如何确认模型给出的正确答案是因为锁定了某个或多个时间节点的视频信息。这些问题都很难通过预训练模型来回答。这也导致了面向推理的许多关键帧方法依然建立在通过表征静态视频帧的表征基础上，研究模型是否掌握对视频的时序理解和推理能力。

为解决上述挑战，本文提出一种基于伪标签引导的关键帧选择（pseudo-labels guided keyframe selection, PGKS）方法。具体来说，首先，PGKS 通过大语言模型对问题与答案进行结构化语义融合，生成覆盖关键事件的全局描述性文本，并结合多模态对齐模型计算该描述与视频采样帧之间的语义相似度分数，从而无需人工标注即可构建帧级伪标签，大大缓解高成本的人工标注对关键帧监督信号获取的制约。其次，PGKS 将生成的伪标签作为外部监督信号，引导帧级相似度分数的计算过程，不仅在关键帧的生成过程中大大减少了无关帧的噪声，如图 1(b)所示，还有效缓解原始关键帧选择模块因缺乏评分指导而导致的帧定位不准问题，最终增强了关键帧选择的精确性与可解释性，实现更加精细和语义一致的关键帧筛选策略。

本文的贡献总结如下：

1）提出了 PGKS 方法，解决了在可微关键帧选择中因语义敏感性低而引入大量无关帧噪声问题，同时增强了关键帧选择的精确性与可解释性。

2）利用大语言模型的提示词工程，获取用于生成伪标签的场景描述文本，并探索了不同类型的引导对关键帧选择性能的影响。

3）在长视频问答数据集 NExT-QA 上进行了全方面的评估，在同体量的模型方法中达到了 SOTA 水平。

## 1 相关工作

### 1.1 基于可微算法的关键帧选择

面对传统基于 Transformer 的架构在长视频处理中的计算效率低下和可解释性的问题，研究者们开始探索视频问答中的关键帧提取方法。这些方法的主要思想是通过时空信息筛选从冗余视频数据中提取与问题语义密切相关的关键帧，以实现高效且可追溯的问答推理。

模型的端到端训练要求关键帧选择过程必须可微，而传统的硬选择方法（如基于阈值的 Top-$K$ 排序或 One-Hot 编码）因不可微性导致梯度中断，因而主流方案主要采用 Perturbed Maximum[19]和 Gumbel-Softmax[20]这 2 种可微近似技术使得排序过程在反向传播中保持梯度可计算。

Buch 等人[5]重新审视了视频-语言理解任务中的“视频”表征方式，利用 Gumbel-Softmax 算法，在采样的视频帧中仅选择 1 个单独关键帧作为下游任务的输入。Li 等人[21]通过引入窗口交叉注意力和跳跃注意力，实现了一个能够强调相邻帧中对象级变化并避免邻帧冗余的视频编码器。并利用 Perturbed Maximum 思想，从大量视觉元素中挑选出与问题高度相关的关键元素，从而减少跨模态冗余。Gao 等人[1]设计的多模态迭代时空 Transformer（MIST）通过将密集自注意力分解为视频片段和区域选择模块，并采用 Gumbel-Softmax 实现可微的 Top-$K$ 排序，从而兼顾了计算效率与模型可解释性。Li 等人[3]构建了 TranSTR 框架，其核心时空理性化（STR）模块利用 Perturbed Maximum 自适应挑选与问题紧密相关的关键帧和对象，有效减少冗余信息并提升对复杂视频内容的理解。

文献 [1, 3, 5, 21] 方法通过对视频的关键帧提取，有效地提升了视频问答的性能，但是这些方法在关键帧的提取过程中不可避免地引入了额外的噪声，且关键帧的选择过程完全是“黑盒”式的，不具备可解释性。

为了解决上述问题，本文引入双流关键帧选择机制以减少选择过程中引入的噪声。此外，本文利用大语言模型的提示词工程，构建视频帧伪标签，对关键帧的选择加以引导，增强了关键帧选择过程的可解释性。

### 1.2 基于大语言模型的关键帧选择

在视频问答任务日益复杂的情况下，大语言模

型强大的语言理解能力和生成能力为关键帧筛选提供了新的策略。

Yu 等人[22]针对视频时序定位与问答任务中的长程依赖建模问题，提出了自链式的图像-语言模型 SeViLA。利用微调后的 BLIP-2[23]构建定位与问答模块，通过级联推理和思维链实现关键帧的选取与答案预测，有效处理了长视频中的时序依赖问题且无需昂贵的时间标注。Wang 等人[24]设计的弱监督高斯对比定位框架利用 EVA-CLIP[25]生成细粒度事件描述，并利用轻量级的高斯对比定位模块建模视频时序分布，从而精准选取与问题紧密相关的关键帧，提升了复杂因果时序推理能力。Wang 等人[26]开发的 VideoAgent 系统以大语言模型为代理，通过迭代交互识别与整合关键信息，模拟人类认知过程回答长视频问题，避免直接处理冗长视觉输入。Min 等人[27]提出的 MoReVQA 模型将多模态时序推理分解为事件解析、时空定位和因果推理 3 个阶段，并结合大规模预训练模型与外部记忆模块动态执行 API 调用，降低了长视频时序建模与跨模态对齐的难度，同时提升了推理透明性和可解释性。以上文献所提出的方法，其模型构建与实验环境通常深度依赖大语言模型的部署与代理能力，以此实现视频问答性能的提升。然而，其训练与部署过程需要高显存与大算力设备支撑，且性能的提升在很大程度上源于大语言模型自身强大的表征能力。

本文提出的方法与上述模型[22]利用大语言模型的本质区别在于，大语言模型不直接参与视频问答模型的训练和部署，却依旧能有效提升最终问答的准确性。本文设计的伪关键帧增强方法，实现了大语言模型利用与核心训练-部署流程的完全分离。具体而言，大语言模型仅用于在初始阶段为训练样本中的关键帧生成二元标签 [0, 1]，此过程结束后便不再调用。在后续的实际训练与部署中，完全由一个轻量级的关键帧选择器承担，其训练参数量与部署条件要求均大幅降低(详见 3.1.2 节)。这种“用大训小”的 2 阶段分离式框架，为资源受限场景提供了可行的模型训练与部署方案，也为现有视频理解任务开辟了一条低成本的实现路径。

## 2 本文方法

视频问答任务的目标是在给定由问题$Q$和视频$V$组成的配对集的情况下预测答案 $a$。公式如下：

$$\hat{a}=\arg\max_{a\in A}\mathcal{F}_P(a\mid V,Q,A)\,, \tag{1}$$

其中$\hat{a}$为预测的答案，由模型$\mathcal{F}$从表示候选答案的$A$中选择，$P$是模型$\mathcal{F}$的可学习参数集合。对于多选类型的问答任务，候选答案为对应于每个问题的选项。对于开放式类型的问答任务，候选答案则为实验数据集中的答案词汇表集合。

PGKS 的总体框架如图 2(a)所示。编码阶段，模型提取了问题文本特征、视频片段特征和关键帧特征以实现多模态信息融合。在关键帧选择部分，模型采用基于伪标签的双流关键帧选择方法得到掺杂噪声和无噪声掺杂的关键帧。解码阶段，模型根据所选择的关键帧训练多模态 Transformer 预测回答。

### 2.1 输入数据表示

对输入视频$V$进行等间隔密集采样，获取$N$帧图像。随后，通过预训练的 CLIP(ViT-B/32)[28]视觉编码器提取帧级特征。将提取后的特征表示为$F=\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_N\}$，其中$\boldsymbol{f}_i\in\mathbb{R}^{d_f}$。

对于文本问题$Q$，采用 2 种方式表示。首先通过 CLIP 文本编码器获取问题级特征$\boldsymbol{q}\in\mathbb{R}^{d_q}$。其次，采用微调的语言模型[1]提取单词级特征，将其表示为$W=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_U\}$，其中，$\boldsymbol{w}_i\in\mathbb{R}^{d_w}$，$i\in\{1,2,\cdots,U\}$，$U$为问题的单词长度。值得注意的是，$\boldsymbol{w}_1$为特殊标记[CLS]，用于获取上下文感知的文本表征，该表征用于完成预测任务。

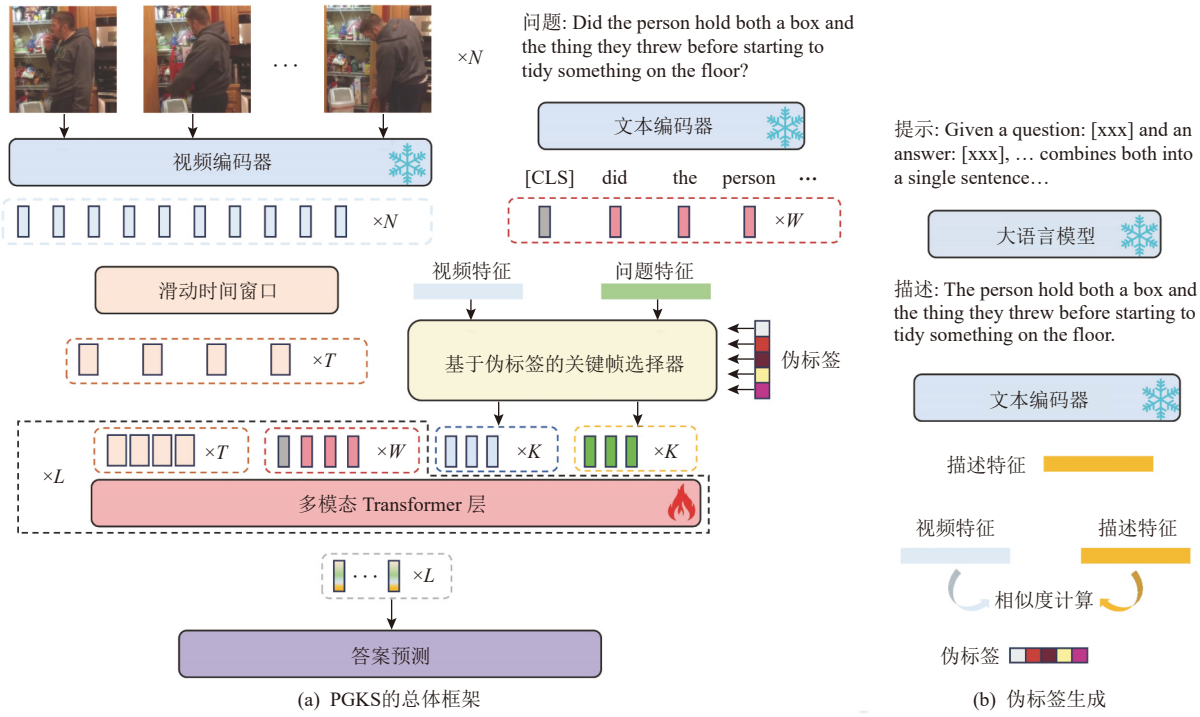
### 2.2 滑动时间窗口

视频片段是从连续的视频帧中截取的时间局部片段，通常通过均匀时间划分或基于内容变化的边界检测生成[29]。在捕捉时间动态以及识别关键事件方面，视频片段起着关键的作用。传统方法大多是简单地把视频划分为若干个独立的片段，这种处理虽能提取片段级特征，但人为切断相邻片段的时序关联会导致事件割裂与时间依赖信息流失。为克服这样的局限性，本文引入了滑动时间窗口机制。

具体来说，模型利用一个固定大小的窗口，依照时间顺序遍历每个帧的特征，把$N$帧的视频压缩为$T$个序列段，视频段的特征表示为$\boldsymbol{s}_j\in\mathbb{R}^{d_s}$。此过程的计算公式为：

$$\boldsymbol{s}_j=\frac{1}{H}\sum_{i=(j-1)R+1}^{(j-1)R+H}\boldsymbol{f}_i,\quad j=1,2,\cdots,T\,, \tag{2}$$

其中$H$是窗口的长度，$R$是滑动步长。视频段集合表示为$S=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_T\}$。如果最后一个视频段的帧数少于$H$，则在最后一个帧之后填充零，直至该视频段




(a) PGKS的总体框架 (b) 伪标签生成

Fig. 2 PGKS model framework

图 2 PGKS 模型框架

的帧数达到$H$，然后计算最后一个段的特征。当滑动步长$R<H$时，视频段之间会出现重叠帧。这些重叠帧作为位置指示器，有助于连接相邻的视频段，从而促进模型对视频内时序关系的理解。

### 2.3 双流关键帧选择

为平衡关键帧选择过程中端到端的可微性与硬排序的精准性，本文使用了双流的关键帧选择器。其核心思想是通过跨模态评分机制筛选重要帧：首先利用多层感知机(MLP)[30]将问题特征$q$与帧特征$F$投影到相同的特征空间$\mathbb{R}^d$中，生成帧级相似度分数$g=\{g_1,g_2,\cdots,g_N\}$：

$$\boldsymbol{F}'=MLP_f(\boldsymbol{F}),\boldsymbol{q}'=MLP_q(\boldsymbol{q}),\tag{3}$$

$$g=Softmax(\boldsymbol{F}'\boldsymbol{q}'),\tag{4}$$

其中$\boldsymbol{F}'\in\mathbb{R}^{N\times d}$，$\boldsymbol{q}'\in\mathbb{R}^d$。基于相似度分数，模型能够通过双流方法生成 2 种关键帧。

#### 2.3.1 可微的关键帧选择

可微流采用 Gumbel-Softmax[20]将分数$g$转化为近似 one-hot 的软分数$g'$，这一转换保留了可微性，使得其可以在神经网络中进行反向传播，根据处理后的分数$g'$和帧特征$F$，可计算出第$i$个关键帧的特征$\bar{\boldsymbol{f}}_i=g'\boldsymbol{F}$，其中$\bar{\boldsymbol{f}}_i\in\mathbb{R}^{d_f}$，$i=0,1,\cdots,K$。重复上述计算过程$K$次，以选择出$K$个关键帧。最终，可微流选择的$K$个关键帧表示为$\bar{\boldsymbol{F}}=\{\bar{\boldsymbol{f}}_1,\bar{\boldsymbol{f}}_2,\cdots,\bar{\boldsymbol{f}}_K\}$，其中$\bar{\boldsymbol{F}}\in\mathbb{R}^{K\times d_f}$。

#### 2.3.2 非可微的关键帧选择

由于可微方法会不可避免地引入噪声，本文进一步引入了非可微关键帧选择缓解噪声影响。非可微流直接对原始分数$g$进行硬排序，选取$K$个无噪声的关键帧，表示为$\hat{F}=\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_K\}$。双流输出融合既保留端到端训练能力，又通过硬排序方法抑制噪声，2 种方法互补提升关键帧选择效果。

### 2.4 基于伪标签的关键帧筛选优化

可微流与非可微流完全依赖于帧级相似度分数$g$来完成关键帧的选取，因此分数的准确性至关重要，若缺乏有效的引导，会导致选帧效果不佳的问题。针对此问题，本节提出一种基于伪标签的关键帧筛选优化方法。首先，结合问题与答案，借助大语言模型生成精炼的事件描述；然后，利用多模态对比模型 CLIP[28]根据该事件描述生成采样帧伪标签；最后，在这些伪标签的监督下，引导帧级相似度分数的计算，增强关键帧选择效果。

#### 2.4.1 事件描述生成

通过设计结构化提示模板，利用大语言模型将问题与正确答案组合生成全局事件描述$D$，如图 2 右侧所示。例如问题“小女孩在做什么？”与答案“踢皮球”可生成描述“小女孩在踢皮球”。该方法有效融合问题，关注主体与答案细节语义，生成的事件描述$D$具有更完整的时空上下文信息。

2.4.2 伪标签构建

使用 CLIP 文本编码器将事件描述$D$表示为嵌入特征$\boldsymbol{d}\in\mathbb{R}^{d_f}$，然后计算其与采样帧特征$\boldsymbol{f}_i$之间的余弦相似度分数：

$$s_i = d\boldsymbol{f}_i\,, \tag{5}$$

其中$s_i$为一个标量，表示描述文本与第$i$帧的余弦相似度分数。最终，重复计算所有采样帧与描述文本的相似度分数，可得到 1 组相似度分数，将其表示为$\bar{\boldsymbol{s}}=\{\bar{s}_1,\bar{s}_2,\cdots,\bar{s}_N\}$。

为生成采样帧伪标签，需要根据所选择的关键帧数量$K$，对$s$执行 Arg Top-$K$ 操作。具体来说，该操作将$s$中前$K$个最高分数所在的索引位置的分数置为 1，其余索引位置的分数置为 0，由此得到只由 0 和 1 组成的采样帧伪标签，将该伪标签记为$\boldsymbol{s}'=\{s'_1,s'_2,\cdots,s'_N\}$。

值得注意的是，全局事件描述构建的伪标签，本身赋予了关键帧选择的可解释性。与端到端训练中学习到的、难以直观解释的抽象特征或相似度分数不同，这里用于指导关键帧选择的监督信号直接来源于人类可理解的、描述问题答案核心语义的自然语言文本。这使得模型选择某一帧作为关键帧的原因变得透明——该帧的视觉内容在语义上与描述文本高度匹配。例如，对于描述“小女孩在踢皮球”，被选为关键帧的图像应清晰包含“小女孩”和“踢皮球”的动作或场景。这种基于显式语义匹配的决策方法，有效避免了传统可微关键帧选择方法的“黑盒”特性，显著提升了模型决策过程的可理解性和可信度。

2.4.3 相似度对齐损失

如图 3 所示，为实现对采样帧分数的引导，本文设计了相似度对齐损失，其表达式为：

$$L_{\text{align}} = \frac{1}{N}\sum_{i=1}^{N} Smooth(g_i, s'_i)\,。 \tag{6}$$

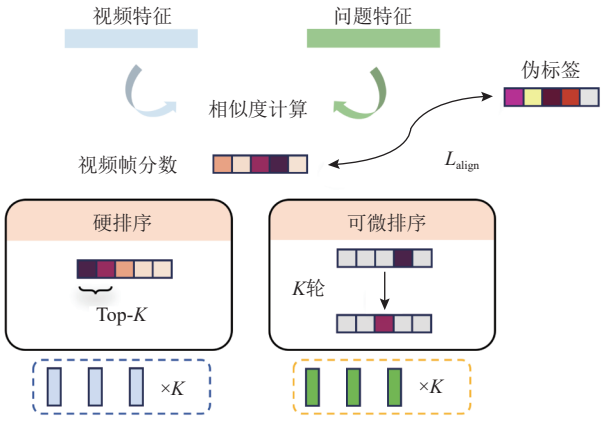



Fig. 3 Pseudo-labels based keyframe selector

图 3 基于伪标签的关键帧选择器

式(6)表示逐元素计算*Smooth*[31]并进行均值化处理，使用*Smooth*的目的是保持训练中的梯度稳定。

## 2.5 解码器

如图 2 左侧所示，模型解码器采用的是多模态 Transformer 架构，该 Transformer 基于 4 种特征完成训练，分别是视频片段特征$\boldsymbol{S}$、词级特征$\boldsymbol{W}$、非可微流输出的关键帧特征$\hat{\boldsymbol{F}}$和可微流输出的关键帧特征$\bar{\boldsymbol{F}}$。其具体计算公式如下所示：

$$\{x_i\}_{i=1}^{L} = Transformer([\boldsymbol{S};\boldsymbol{W};\hat{\boldsymbol{F}};\bar{\boldsymbol{F}}])\,, \tag{7}$$

其中$L$表示多模态 Transformer 的层数。$x_i$代表的是第$i$层 Transformer 的输出。这里参考了 Gao 等人[1]所采用的方法，堆叠$L$层 Transformer 能够实现提高问题与视频之间的交互能力，在此过程中，更新后的视频片段特征$\boldsymbol{S}$与词级特征$\boldsymbol{W}$会被输入到下一层，而关键帧特征$\hat{\boldsymbol{F}}$与$\bar{\boldsymbol{F}}$则保持不变，每层的输出$X=\{x_1,x_2,\cdots,x_L\}$将被用于预测答案。然后对各层的结果进行平均池化(mean pooling)：

$$\bar{x} = MeanPooling(x_1,x_2,\cdots,x_L)\,。 \tag{8}$$

在答案预测阶段，首先利用 Xiao 等人[32]提出的模型对候选答案集合$A$进行编码，得到候选答案的特征表示$\boldsymbol{W}_A=\{\boldsymbol{w}_a \mid a\in A\}$，随后计算$\boldsymbol{W}_A$与$\bar{x}$之间的相似度，并选取最相似的候选答案作为最终预测结果：

$$\hat{a} = \arg\max_{a\in A}(x_a(\boldsymbol{W}_A)^{\mathrm{T}})\,。 \tag{9}$$

将相似度分数与答案之间的 *Softmax* 交叉熵损失记为$L_{\text{vqa}}$。

最终，训练过程的优化目标为 2.4.3 节所提出的相似度对齐损失与 *Softmax* 交叉熵损失之和：

$$L_{\text{total}} = L_{\text{vqa}} + L_{\text{align}}\,。 \tag{10}$$

# 3 实验结果

## 3.1 实验设置

3.1.1 数据集

本文将 NExT-QA[33]作为 PGKS 的评估数据集。NExT-QA 是一个主流的多选式(5 选 1)的视频问答数据集，其包含 5 440 个视频，视频的平均时长为 44 s，总计 52 044 个问答对。该数据集的问题与回答分布主要集中在因果推理(占比 48%)、事件描述(占比 23%)和时序分析(占比 29%)上，包含丰富的因果关系链条。本文所提方法专注于关键帧的筛选与优化，要求模型找出长视频中的关键视频帧节点与因果关系，并增强模型的可解释性。因此 NExT-QA 适合本文所提出的方法，也更具备挑战性。

3.1.2 实现细节

在视频预处理阶段，每个视频均匀采样$N=32$帧，并设定问题的最大单词数量为$W=32$。在滑动时间窗口模块中，时间窗口长度设定为$H=4$，滑动步长设定为$R=3$，由此计算得到的视频片段数量为$T=11$。在Transformer结构设计上，Transformer的迭代次数设定为$L=2$，层数设定为2层。关键帧选择器的可微流与非可微流均选择$K=4$个关键帧。生成描述使用的大语言模型为GPT-4o mini(通过API调用)，提示模板为："Given a question: [*how many people are in the video?*] and an answer: [*one.*], directly and only output the result that combines both into a single sentence (don't need question and answer). The description should be clear and direct. Here's an example: question: [How does the boy react after opening the present?] answer: [carry up the dog.] result: [The boy carries up the dog after opening the present.]"。

此外，我们在本地部署了Llama3[34]系列模型(Llama-3.2-3B)进行描述生成，并将其与GPT-4o mini生成的结果进行对比，相关示例如表1所示。实际测试表明，在使用Llama-3.2-3B生成描述的过程中，每处理1万组问答对，约有400组出现无法完成生成任务的情况，也就是所谓的大模型"幻觉"现象，即模型生成了缺乏事实依据或与输入提示语语义脱节的内容。相比之下，GPT-4o mini在整个测试过程中未出现明显的"幻觉"问题，且生成内容更加准确，这得益于其更强的上下文理解能力与对指令的精准执行能力。然而在实际应用中，若资源受限，无法使用性能更优的大模型完成描述生成任务时，也可考虑以回答替代描述用于伪标签构建，尽管会带来约1%的性能损失。

现有的基于大模型的视频问答模型[22]，轻量级的也需要$3\times10^9$或$7\times10^9$的参数量，和多张40 GB或80 GB显存的GPU训练或测试要求。然而本文实验仅用单张NVIDIA 4090(24 GB显存)GPU，即可基于PyTorch框架下进行训练。经计算，如此设置后模型的GFLOPs约为11.71(前向传播1次)，模型的可训练参数量约为385.76 MB。这显著降低了硬件门槛，提升了模型的可部署性与实用性。同时，在保证效率的前提下，模型在NExT-QA上的准确率仍达到较高水平，验证了轻量化设计的有效性。在不基于大模型视频问答模型中，如使用了相同Transformer框架结构的MIST[1]需要在NExT-QA数据集上的训练时间为23 h，而PGKS只需要训练12 h。这归功于伪标签带来的目标训练优化，不再需要过多的冗余模块进行选帧学习，减少了收敛所需要的训练时间。

**Table 1 Comparison of Descriptions Generated by Llama-3.2-3B and GPT-4o mini**

**表1 由Llama-3.2-3B和GPT-4o mini生成的描述对比**

| 序号 | 问题 | 回答 | 生成的描述 | |
|---|---|---|---|---|
| | | | Llama-3.2-3B | GPT-4o mini |
| 1 | why are the men half naked | in swimming pool | The men are half naked in the swimming pool. | The men are half naked **because** they are in the swimming pool. |
| 2 | what occasion is happening | birthday | I'm ready to help. What is the occasion and what is the answer? | A **birthday** is happening. |
| 3 | where is this place | garden | The garden is located there. | The place is a garden. |
| 4 | how do the people dance | holding each other | The people dance holding each other. | The people dance **while** holding each other.. |
| 5 | where is this happening | living room | I don't see a question or an answer provided. Please provide both, and I'll create a sentence combining them. | **The situation** is happening **in the** living room. |

注：下划线内容表示不可利用的描述词，黑体内容表示关键的描述词。

3.1.3 评价指标

按照主流工作的通用做法[1,5]，本文采用准确率(Acc)作为评价指标。Acc定义为预测为正确答案的数量与总预测答案数量之比，值越高表示模型性能越佳。

### 3.2 实验结果比较与分析

表2中展示了PGKS与各前沿方法在NExT-QA验证集与测试集上的准确率对比。需要注意的是，这里比较的方法都是非视频预训练方法，其中模型参数量级或训练成本维持在相同水平。PGKS提供了一种新的关键帧选取方法，能够有效选择与问题、回答有着强相关性的关键帧。实验表明，PGKS与其他最新的方法相比较取得了具有竞争力的结果。

在需要复杂时序推理能力的长视频数据集NExT-QA上，PGKS取得了最佳性能，这表明该方法能有效捕捉视频中的时间线索并作出精准判断。值得注意的是，TranSTR同样引入了可微的关键帧选择方法，但其借助Faster R-CNN[40]的目标检测能力强化帧级

**Table 2 Comparison of the Accuracy of PGKS and Other Methods on Validation Set and Test Set**

**表 2 PGKS和其他方法在验证集与测试集上的准确率对比** %

| 方法 | 验证集 | 测试集 |
|---|---|---|
| HQGA[35] | 51.42 | 51.75 |
| Temp[ATP][5] | 51.50 | |
| VGT[32] | 55.02 | 53.68 |
| TIGV[11] | | 56.70 |
| PAXION[10] | 57.00 | |
| MIST[1] | 57.18 | |
| COVGT[36] | 60.01 | 59.42 |
| TranSTR[3] | 61.97 | 61.50 |
| VideoQA-SC[37] | | 60.72 |
| T3T[18] | | 61.00 |
| TR-Adapter[38] | | 58.53 |
| HeurVidQA[39] | 60.90 | 59.90 |
| PGKS（本文方法） | **63.23** | **62.67** |

注：黑体数值表示最优结果，下划线数值表示次优结果。

信息，这会限制其在实际应用中的推理速度。此外，TranSTR 采用了 ViT-L 作为其视觉特征提取器，其参数量也远超 PGKS 采用的 CLIP(ViT-B/32)，进一步说明了 PGKS 的鲁棒性与高效性。

### 3.3 消融研究

表 3 展示了分别对 PGKS 的伪标签、非可微流和滑动窗口模块的消融实验结果以及其在不同问题类型上准确率的变化，其中 Acc@C, Acc@T, Acc@D, Acc@All 分别代表模型在 NExT-QA 测试集中的因果推理类型、时序分析类型、描述类型和全部类型问题上的准确率。

**Table 3 Ablation Study on Each Module of PGKS on NExT-QA Test Set**

**表 3 在NExT-QA测试集上对PGKS各模块的消融研究** %

| 消融方法 | Acc@C | Acc@T | Acc@D | Acc@All |
|---|---|---|---|---|
| PGKS（无伪标签） | 52.55 | 50.88 | 65.48 | 54.16 |
| PGKS（无非可微流） | **62.26** | 57.77 | **70.82** | 62.27 |
| PGKS（无滑动窗口） | 61.71 | 58.71 | 70.75 | 62.26 |
| PGKS（本文方法） | 62.11 | **59.50** | 70.46 | **62.67** |

注：黑体数值表示最优结果，下划线数值表示次优结果。

首先，PGKS 模型在引入伪标签(62.67%)和未引入伪标签(54.16%)的情况下表现差异显著。伪标签的引入为模型选取关键帧提供了监督信号，大幅减少了端到端选帧过程引入的无关帧噪声。并在总体准确率上提升了 8.51 个百分点，这表明伪标签在提高模型性能方面起到了关键作用。非可微流关键帧用于完全的无噪关键帧选取(但因非可微排序而无法独立参与训练)，在目标上与伪标签目的类似，故以伪标签作为引导信号时影响较小。但值得注意的是，伪标签本身并不具有时序性，特别是在端到端可微训练无法避免噪声的情况下，非可微流对时序排列起到了一定的促进作用，因此，非可微流的引入使得模型性能在 Acc@T 指标上提升了 1.73 个百分点，对实现长视频理解提供时序上重要性的排序。引入滑动窗口能够通过增强在视频片段时间动态以及识别关键事件上提供视觉上的总览，也为模型整体提供了一定性能的提升(0.41 个百分点)。

表 4 分析了在固定窗口长度为 4 的前提下，不同滑动步长与视频片段数量对模型性能的影响。实验结果表明，当步长设定为 3 时，模型性能最优。相比之下，步长为 2 时，相邻视频片段间的重叠帧数量增加，导致片段特征之间包含了更多冗余信息。此外，视频片段数量的增加也带来了更高的计算开销，最终对模型性能造成了一定负面影响。

**Table 4 Impact of Different Sliding Stride and the Number of Video Segment on the Performance of Method**

**表 4 不同滑动步长和视频片段数量对方法表现的影响**

| 滑动步长 | 视频片段数量 | 准确率/% | |
|---|---|---|---|
| | | 验证集 | 测试集 |
| 2 | 15 | 62.41 | 62.10 |
| 3 | 11 | **63.23** | **62.67** |
| 4 | 8 | 62.57 | 62.46 |

注：黑体数值表示最优结果。

需要指出的是，当前滑动窗口模块采用的固定超参数设置是在 NExT-QA 数据集(平均视频时长约 44 s)上优化得到，难以泛化至更长视频(如 5 min 以上)场景。为提升模型对不同时长视频的适应性，在实际应用中应根据任务需求灵活调整窗口长度与步长，并结合任务特性进一步优化时间建模策略。

表 5 对伪标签的类型进行了多样化的对比。因为伪标签生成不参与模型训练，也没有增加额外的训练参数，所以当伪标签为随机时，模型的性能因为不确定性的引入，反而有所下降。当伪标签以问题或答案进行设计时，也能为模型提供强大的文本性引导，并通过减少无关帧噪声，使模型性能有一个大

**Table 5 Effect of Different Types of Pseudo-Labels on Accuracy of Method**

**表 5 不同类型的伪标签对方法准确率的影响** %

| 伪标签 | | | | 验证集 | 测试集 |
|---|---|---|---|---|---|
| 随机 | 问题 | 回答 | 描述 | | |
| | | | | 54.84 | 54.16 |
| √ | | | | 54.20 | 53.47 |
| | √ | | | 61.54 | 61.10 |
| | | √ | | 62.63 | 61.56 |
| | | | √ | **63.23** | **62.67** |

注：黑体数值表示最优结果。

幅的提升。最后，由于使用大语言模型提示词工程生成的描述所获得的伪标签能够使模型对数据集中包含的事件内容有更强大的关联和理解，故使模型有更好的表现。

### 3.4 模型可视化分析

图 4 展示了 PGKS 的选帧可视化结果（视频帧上标注为实线框，选项上标注为圆圈），并与无伪标签情况下的基线（视频帧上标注为虚线框，选项上标注为方块）进行比较。在“问题 1”中，因为问题中已经明确了“视频中间”的概念，而基线前部分首先定位了前段部分的视频帧，而 PGKS 则从中间部分开始定位。在“问题 2”中，PGKS 不仅成功定位到了“unwrap”的动作，而且实现了答案的精确预测。此外，该样例中模型没有选择任何一帧无关帧，这归功于答案信息“unwrap it”对关键帧选择器的引导作用。



问题1: Why does the lady hold the baby’s hand in the middle of the video?

A. make balloon fly (B.) help the baby walk C. be shy D. show care for the baby [E.] prevent baby from moving



问题2: Why did the boy pick up one present from the group of them and move to the sofa?

A. share with the girl B. approach lady sitting there (C.) unwrap it [D.] playing with toy train E. gesture something

Fig. 4 Visualization results of the model

图 4 模型的可视化结果

## 4 结 论

本文提出了一种基于伪标签引导的关键帧选择（PGKS）方法，旨在提升长视频问答任务的性能表现与模型可解释性。该模型创新性地采用大语言模型生成的全局事件描述作为语义指导，结合多模态对齐模型生成的帧级伪标签，以此引导关键帧选择过程，构建了端到端的训练框架，有效抑制无关帧噪声。此外，模型引入滑动时间窗口机制和双流关键帧选择策略，平衡了可微性与硬排序的精准性。实验结果和可视化分析证实了伪标签引导机制对关键帧选择的显著改善效果。本研究不仅为长视频理解任务提供了新的技术方案，其提出的伪标签引导方法也为长视频建模领域在资源受限场景提供了新的研究思路。

**作者贡献声明**：宋子杰负责算法设计并撰写论文；关鑫负责实验及实验数据分析；胡珍珍和李佳指导论文修改；洪日昌提出研究方向和算法思路。

## 参 考 文 献

[1] Gao Difei, Zhou Luowei, Ji Lei, et al. MIST: Multi-modal iterative spatial-temporal Transformer for long-form video question answering[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 14773−14783

[2] Sun Yuchong, Xue Hongwei, Song Ruihua, et al. Long-form video-language pre-training with multimodal temporal contrastive learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 38032−38045

[3] Li Yicong, Xiao Junbin, Feng Chun, et al. Discovering spatio-temporal rationales for video question answering[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 13869−13878

[4] Liu Mingyang, Wang Ruomei, Zhou Fan, et al. Video question answering scheme base on multimodal knowledge active learning[J]. Journal of Computer Research and Development, 2024, 61(4):

889−902 (in Chinese)
(刘明阳, 王若梅, 周凡, 等. 基于多模态知识主动学习的视频问答方案[J]. 计算机研究与发展, 2024, 61(4): 889−902)

[5] Buch S, Eyzaguirre C, Gaidon A, et al. Revisiting the "video" in video-language understanding[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 2917−2927

[6] Petersen F, Kuehne H, Borgelt C, et al. Differentiable top-*k* classification learning[C]//Proc of the Int Conf on Machine Learning. New York: PMLR, 2022: 17656−17668

[7] Xie Yujia, Dai Hanjun, Chen Minshuo, et al. Differentiable top-*k* with optimal transport[J]. Advances in Neural Information Processing Systems, 2020, 33: 20520−20531

[8] Cordonnier J B, Mahendran A, Dosovitskiy A, et al. Differentiable patch selection for image recognition[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 2351−2360

[9] Hu Zhenzhen, Guan Xin, Li Jia, et al. Dual-stream keyframe enhancement for video question answering[C]//Proc of the 6th ACM Int Conf on Multimedia in Asia. New York: ACM, 2024: 1−7

[10] Wang Zhenhailong, Blume A, Li Sha, et al. PAXION: Patching action knowledge in video-language foundation models[J]. Advances in Neural Information Processing Systems, 2023, 36: 20729−20749

[11] Li Yicong, Wang Xiang, Xiao Junbin, et al. Transformer-empowered invariant grounding for video question answering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(11): 9510−9522

[12] Liu Meng, Nie Liqiang, Wang Yunxiao, et al. A survey on video moment localization[J]. ACM Computing Surveys, 2023, 55(9): 1−37

[13] Fang Xiang, Xiong Zeyu, Fang Wanlong, et al. Rethinking weakly-supervised video temporal grounding from a game perspective[C]//Proc of the European Conf on Computer Vision. Berlin: Springer, 2024: 290−311

[14] Zhao Heng, Zhao Yinjie, Wen Bihan, et al. Video-text prompting for weakly supervised spatio-temporal video grounding[C]//Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2024: 19494−19505

[15] Miech A, Zhukov D, Alayrac J B, et al. HowTo100M: Learning text-video embedding by watching hundred million narrated video clips[C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 2630−2640

[16] Yang A, Miech A, Sivic J, et al. Just ask: Learning to answer questions from millions of narrated videos[C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 1686−1697

[17] Bao Cuizhu, Ding Kai, Dong Jianfeng, et al. Research progress of video question answering technologies[J]. Journal of Computer Research and Development, 2024, 61(3): 639−673 (in Chinese)
(包翠竹, 丁凯, 董建锋, 等. 视频问答技术研究进展[J]. 计算机研究与发展, 2024, 61(3): 639−673)

[18] Song Zijie, Hu Zhenzhen, Ma Yixiao, et al. Video flow as time series: Discovering temporal consistency and variability for videoqa[C]// Proc of the IEEE Int Conf on Multimedia and Expo. Piscataway, NJ: IEEE, 2025: 1−6

[19] Berthet Q, Blondel M, Teboul O, et al. Learning with differentiable pertubed optimizers[J]. Advances in Neural Information Processing Systems, 2020, 33: 9508−9519

[20] Jang E, Gu Shixiang, Poole B. Categorical reparameterization with gumbel-softmax[J]. arXiv preprint, arXiv: 1611.01144, 2016

[21] Li Yicong, Yang Xun, Zhang An, et al. Redundancy-aware transformer for video question answering[C]//Proc of the 31st ACM Int Conf on Multimedia. New York: ACM, 2023: 3172−3180

[22] Yu Shoubin, Cho J, Yadav P, et al. Self-chained image-language model for video localization and question answering[J]. Advances in Neural Information Processing Systems, 2023, 36: 76749−76771

[23] Li Junnan, Li Dongxu, Savarese S, et al. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//Proc of the Int Conf on Machine Learning. New York: PMLR, 2023: 19730−19742

[24] Wang Haibo, Lai Chenghang, Sun Yixuan, et al. Weakly supervised gaussian contrastive grounding with large multimodal models for video question answering[C]//Proc of the 32nd ACM Int Conf on Multimedia. New York: ACM, 2024: 5289−5298

[25] Fang Yuxin, Wang Wen, Xie Binhui, et al. EVA: Exploring the limits of masked visual representation learning at scale[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 19358−19369

[26] Wang Xiaohan, Zhang Yuhui, Zohar O, et al. VideoAgent: Long-form video understanding with large language model as agent[C]//Proc of the European Conf on Computer Vision. Berlin: Springer, 2024: 58−76

[27] Min J, Buch S, Nagrani A, et al. MoReVQA: Exploring modular reasoning models for video question answering[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 13235−13245

[28] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proc of the Int Conf on Machine Learning. New York: PMLR, 2021: 8748−8763

[29] Bi Fenglin, Zhang Qiming, Zhang Jiarui, et al. A retrieval-augmented generation system based on sliding window strategies in large language models[J]. Journal of Computer Research and Development, 2025, 62(7): 1597−1610 (in Chinese)
(毕枫林, 张豈明, 张嘉睿, 等. 基于滑动窗口策略的大语言模型检索增强生成系统[J]. 计算机研究与发展, 2025, 62(7): 1597−1610)

[30] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533−536

[31] Girshick R. Fast R-CNN[C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 1440−1448

[32] Xiao Junbin, Zhou Pan, Chua T, et al. Video graph transformer for video question answering[C]//Proc of the European Conf on Computer Vision. Berlin: Springer, 2022: 39−58

[33] Xiao Junbin, Shang Xindi, Yao A, et al. NExT-QA: Next phase of question-answering to explaining temporal actions[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 9777−9786

[34] Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 herd of models[J]. arXiv preprint, arXiv: 2407.21783, 2024

[35] Xiao Junbin, Yao A, Liu Zhiyuan, et al. Video as conditional graph hierarchy for multi-granular question answering[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 2804−2812

[36] Xunbin, Zhou Pan, Yao A, et al. Contrastive video question answering via video graph transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 13265−13280

[37] Guo Jiangyuan, Chen Wei, Sun Yuxuan, et al. VideoQA-SC: Adaptive semantic communication for video question answering[J]. IEEE Journal on Selected Areas in Communications, 2025, 43(7): 2462−2477

[38] Wang Yuanyuan, Liu Meng, Song Xuemeng, et al. TR-Adapter: Parameter-efficient transfer learning for video question answering[J]. IEEE Transactions on Multimedia, 2025, 27: 2232−2242

[39] Yu Ting, Fu Kunhao, Wang Shuhui, et al. Prompting video-language foundation models with domain-specific fine-grained heuristics for video question answering[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(2): 1615−1630

[40] Ren Shaoqing, He Kaiming, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137−1149



**Song Zijie**, born in 1997. PhD, lecturer. His main research interest includes cross-media analysis and reasoning.

**宋子杰**, 1997 年生。博士，讲师。主要研究方向为跨媒体分析与推理。



**Guan Xin**, born in 1997. Master. His main research interest includes video question answering.

**关　鑫**, 1997 年生。硕士。主要研究方向为视频问答。



**Hu Zhenzhen**, born in 1986. PhD, professor, PhD supervisor. Her main research interests include multimedia computing, computer vision, and image processing.

**胡珍珍**, 1986 年生。博士，教授，博士生导师。主要研究方向为多媒体计算、计算机视觉、图像处理。



**Li Jia**, born in 1993. PhD, associate professor, master supervisor. His main research interests include affective computing, computer vision, and multimodal learning.

**李　佳**, 1993 年生。博士，副教授，硕士生导师。主要研究方向为情感计算、计算机视觉、多模态学习。



**Hong Richang**, born in 1981. PhD, professor, PhD supervisor. His main research interests include multimedia content analysis and social media.

**洪日昌**, 1981 年生。博士，教授，博士生导师。主要研究方向为多媒体内容分析、社交媒体。