中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2025)05-1257-15

论文引用格式:Zeng R H, Li J L, Zhuo Y S, Duan H H, Chen Q and Hu X P. 2025. Iterative optimization for video retrieval data using large language model guidance. Journal of Image and Graphics, 30(5):1257-1271(曾润浩, 李嘉梁, 卓奕深, 段海涵, 陈奇, 胡希平. 2025. 大语言模型引导的视频检索数据迭代优化. 中国图象图形学报, 30(5):1257-1271)[DOI:10.11834/jig.240545]

大语言模型引导的视频检索数据迭代优化

曾润浩[1,2],李嘉梁[3],卓奕深[3],段海涵[1,2],陈奇[4*],胡希平[1,2]

1. 深圳北理莫斯科大学人工智能研究院,深圳　518172; 2. 粤港澳情感智能与普适计算联合实验室,深圳　518172;
3. 深圳大学机电与控制工程学院,深圳　518060; 4. 澳大利亚阿德莱德大学计算机科学学院,阿德莱德　5005

摘　要:**目的**　视频文本跨模态检索旨在从视频库或给定视频中检索出语义上与给定查询文本最相似的视频或视频片段,是视频理解的重要应用之一。现有方法主要聚焦于如何通过跨模态交互提高模态间的语义匹配,但忽略了目前数据集存在一个查询文本对应多个视频片段或视频的问题。该问题在训练过程中可能导致模型混淆,制约模型性能。为此,提出一种大语言模型引导的视频检索数据迭代优化方法。**方法**　通过视觉文本相似度定位出数据集中存在一对多问题的查询文本及对应视频,并提取视频中未被查询文本所描述的对象、详细外观、颜色属性等细粒度信息。将这些信息与原查询文本输入到大语言模型中总结优化为更细粒度的查询文本。通过基于视频文本语义关联的迭代条件判断,自动选择优化当前提示并进行下一轮优化或退出优化过程,从而不断优化查询文本。将优化后的数据用于视频文本跨模态检索模型的训练。**结果**　在视频片段检索任务上,4种神经网络模型在使用了本文方法优化后的Charades文本时序标注(charades-sentence temporal annotations,Charades-STA)数据集进行训练,在交并比(intersection over union, IoU)为0.5时,首一召回率(Recall@Top1,R@1)平均提升2.42%,在基于查询的视频高光时刻检测(query-based video highlights,QVHighlights)数据集上,2种神经网络模型平均提升3.42%。在视频检索中,2种神经网络模型在微软视频文本检索(Microsoft research video to text,MSR-VTT)数据集的R@1指标上平均提升1.4%。**结论**　提出的大语言模型引导的视频检索数据迭代优化方法,缓解了数据集中存在的一对多问题,使模型性能显著提升。

关键词:视频理解;跨模态检索;跨模态特征对齐;大语言模型(LLM);数据优化

Iterative optimization for video retrieval data using large language model guidance

Zeng Runhao[1,2], Li Jialiang[3], Zhuo Yishen[3], Duan Haihan[1,2], Chen Qi[4*], Hu Xiping[1,2]

1. Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen　518172, China;
2. Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen　518172, China;
3. College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen　518060, China;
4. School of Computer and Mathematical Sciences, Adelaide University, Adelaide　5005, Australia

收稿日期:2024-09-04;**修回日期**:2024-11-27;**预印本日期**:2024-12-04

*通信作者:陈奇　qi.chen04@adelaide.edu.au

基金项目:国家自然科学基金项目(62202311);深圳市高等院校稳定支持计划项目(20220809180405001);深圳市优秀科技创新人才培养项目(RCBS20221008093224017);广东省基础与应用基础研究基金项目(2023A1515011512);广东省重点领域研发计划(2018B010107001)

Supported by: National Natural Science Foundation of China (62202311); Shenzhen Natural Science Foundation (the Stable Support Plan Program) (20220809180405001); Excellent Science and Technology Creative Talent Training Program of Shenzhen Municipality (RCBS20221008093224017); Guangdong Basic and Applied Basic Research Foundation (2023A1515011512); Guangdong Provincial Scientific and Technological Funds (2018B010107001)

Abstract: Objective In recent years, video-text cross-modal retrieval has garnered widespread attention from academia and industry due to its significant application value in areas such as video recommendation, public safety, sports analysis, and personalized advertising. This task primarily involves video retrieval (VR) and video moment retrieval (VMR), aiming to identify videos or video moments from a video library or a specific video that are semantically most similar to a given query text. The inherent heterogeneity between video and text, as they belong to different modalities, makes direct feature matching highly challenging. Thus, the key challenge in video-text cross-modal retrieval lies in effectively aligning these two cross-modal data types in the feature space to achieve precise semantic relevance calculation. Current methods primarily focus on enhancing semantic matching across modalities through cross-modal interactions on existing datasets to improve retrieval performance. Although improvement in modeling has seen significant progress, issues inherent to datasets remain unexplored. In the context of video-text cross-modal retrieval, this study observes an ill-posed problem during training with existing datasets, manifested as a single query text corresponding to multiple videos or video moments, leading to nonunique retrieval results. These one-to-many samples frequently lead to model confusion during training, hinder the alignment of cross-modal feature representations, and degrade overall model performance. For instance, if a query text describes a target video and a nontarget video, then retrieving the latter during training is penalized as incorrect, thereby artificially increasing the distance between the query text and the nontarget video in the feature space, despite their high semantic relevance. This paper defines these problematic one-to-many samples as hard samples, whereas one-to-one samples are defined as easy samples. To address this issue, this paper proposes an iterative optimization method for VR data using large language model guidance. By leveraging the built-in knowledge of large language models, this method augments one-to-many video-text pairs with fine-grained information and iteratively refines them into one-to-one mappings. **Method** Initially, the dataset is divided into easy and hard sample sets based on video-text similarity. Specifically, the similarity between the query text and all videos is calculated. If the similarity between the query text and the target video is not the highest, then the data pair is classified into the hard sample set; otherwise, it is classified into the easy sample set. For videos in the difficult sample set, several frames are uniformly sampled and inputted into an image-text generation model to produce frame-level descriptive texts. This process aims to capture fine-grained information, such as objects not described by the query text, detailed appearances, and color attributes in the video. However, given that multiple frames may contain similar scenes and objects, the extracted fine-grained textual descriptions are often redundant and noisy. To address this, an iterative optimization module based on video-text semantic associations is introduced. This module combines the original query text with fine-grained information extracted from the target video and integrates it with a carefully designed prompt template, which is inputted into a large language model. The model then generates a refined, fine-grained, and unique query text. The quality of the optimization results depends significantly on the design of the prompt templates. The templates include the following key elements: 1) clear task descriptions; 2) relevant examples that meet specified conditions; and 3) specific requirements, such as extracting co-occurring content across multiple frames during summarization. The emphasis on co-occurring content is justified by two key reasons: first, such content often carries critical and essential information; second, summarizing shared elements effectively reduces the likelihood of introducing erroneous descriptions. High-quality outputs from large language models typically result from multiple interactions with the user, as these models can refine their responses based on user feedback. Inspired by this, the study aims to automate the optimization process without requiring predefined interaction rounds. To further optimize the fine-grained query text, an iterative condition based on video-text semantic association is designed. Specifically, the optimized query text and corresponding video are encoded through an encoder. If the similarity of the extracted features in the feature space meets a predefined condition, then the optimized query text is deemed satisfactory, and the optimization process is terminated. Otherwise, if the condition is not met, then the current optimization results are used to update the prompt information, and the query text is further refined iteratively until the dataset no longer contains one-to-many issues for any query text. Finally, the optimized data are used to train the video-text cross-modal retrieval model. **Result** The effectiveness of the proposed method was validated on multiple mainstream video-text cross-modal retrieval datasets. In the VMR task, four neural network models trained on the Charades-STA dataset and optimized using the proposed method showed an average improvement of 2.42% in the R@1, IoU = 0.5 metric, with a maximum improvement of 3.23%. When IoU = 0.7, performance

improvements reached up to 4.38%. In the QVHighlights dataset, the performance of MomentDETR and QDDETR improved by 5.48% and 1.35%, respectively, with an average improvement of 3% when IoU = 0.7. In the VR task, two methods demonstrated an average improvement of 1.4% in the R@1 metric on the MSR-VTT dataset, with a maximum improvement of 1.6%. These results demonstrate the proposed method's effectiveness and its generalizability across different datasets. **Conclusion** The proposed iterative optimization method for VR data using large language model guidance effectively alleviates the one-to-many issue in datasets. A single optimization of the dataset can enhance the retrieval performance of multiple methods. This approach offers a novel perspective for video-text cross-modal retrieval research and promotes advancements in related technologies.

Key words: video understanding; cross-modal retrieval; cross-modal feature alignment; large language model (LLM); data optimization

0 引　言

视频文本跨模态检索技术主要包括视频检索和视频片段检索，目标是通过用户输入的文本在视频库中找到最符合描述的视频或者在特定视频中寻找与文本最匹配的片段。近年来，因其在视频推荐、公共安全、体育赛事分析和个性化广告等领域具有极高的应用价值，引起了学术界和工业界的广泛关注。由于视频与文本属于两种不同模态，其存在的显著异构性使得直接进行特征匹配极具挑战性（尹奇跃等，2021；陈磊 等，2024）。因此，视频文本跨模态检索的核心挑战在于如何在特征空间有效实现这两种跨模态数据的精准对齐，从而确保语义相关度的精确计算（刘华峰 等，2023；刘颖 等，2020）。现有的跨模态检索算法通常设计复杂的多模态交互模块来实现跨模态特征对齐表示，以提升检索性能（王亚鸽等，2020）。尽管模型改进已取得显著进展，但数据集中存在的问题仍然未被探索。

对于视频文本跨模态检索，本文观察到利用现有数据集进行训练时存在不适定问题（ill-posed problem），其表现为一个查询文本对应多个视频片段或者视频，即检索视频时解不唯一（Hadamard，1902）。这些一对多的样本在训练时会引起模型的混淆。如图1所示，当数据集中存在标注为“一个人在开车”的视频f和标注为“一个人在操控汽车的中控台”的视频e时，现有方法在训练过程中若输入查询文本“一个人在开车”得到的检索结果为视频e时，会被判定为错误，导致特征空间中两者的距离被不恰当地拉远。尽管该查询文本实际上也与视频e高度相关，但由于这类解不唯一的情况会引起模型的混淆，因而不利于跨模态特征的对齐。本文将这些引起不适定问题的一对多样本称为困难样本，将一对一的样本称为简单样本，例如图1中查询文本“一个人在接受采访”仅与视频d匹配。为此，本文基于视频文本相似度对视频片段检索数据集Charades-STA（charades-sentence temporal annotations）（Sigurdsson 等，2016）和视频检索数据集MSR-VTT（Microsoft research video to text）（Xu 等，2016）的测试集划分为简单和困难样本，并将训练集上训练好的模型在这两种样本上测试。实验结果如图2所示，当检索困难样本时，模型性能相较于简单样本的检索显著下降。结果表明直接用已有数据集训练模型，会出现难以将文本和视频进行特征对齐的情况，进而导致检索性能受限。

为解决这一问题，本文将困难样本优化为一对一的简单样本以缓解模型在训练时产生的混淆。鉴于生成式预训练Transformer第4代（generative pre-trained Transformer 4，GPT4）等大语言模型（large language model，LLM）在大规模数据训练后具有出色的推理能力，本文提出一种大语言模型引导的视频检索数据迭代优化方法。该方法首先利用视觉语言预训练模型初步筛选数据集中的困难样本，随后从视频中提取更丰富的文本描述。最后，利用这些更丰富的文本描述生成提示，并利用大语言模型不断提炼提示中的细粒度信息，以迭代优化这些困难样本。实验结果表明，该方法有效地提高了模型性能。

本文主要贡献如下：1）揭示了主流视频文本检索算法在直接利用现有数据集训练时面临的不适定问题，并通过实验验证了其对模型性能的负面影响；2）提出了一种大语言模型引导的视频检索数据迭代优化方法，通过增加额外细节信息将粗粒度文本转化为细粒度文本，并设计了基于视频文本语义关联的迭代优化模块，自动决策是否需要优化提示并调

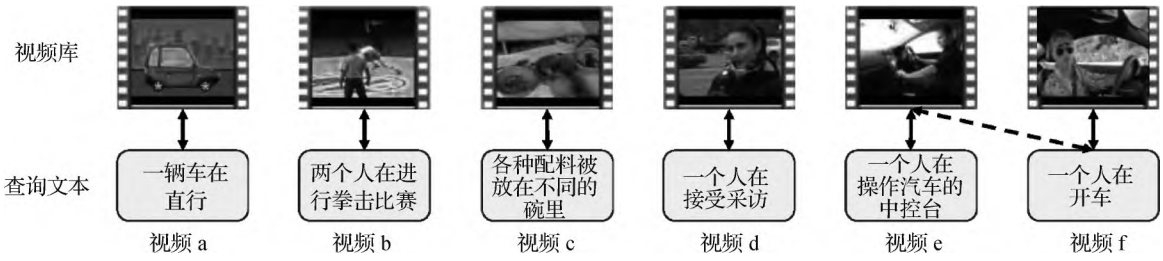

图1　现有跨模态数据集存在不适定问题：同一查询文本可能对应多个视频（虚线箭头）

Fig. 1　The existing cross modal dataset has an ill posed problem: the same query text may correspond to multiple videos (dashed arrow)

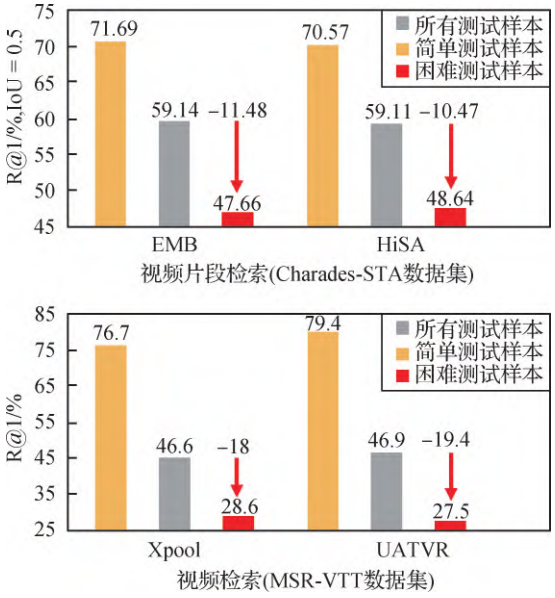

图2　现有方法在不同难度样本上的检索性能

Fig. 2　The retrieval performance of existing methods on samples of different difficulty levels

用大语言模型进行下一轮优化，确保优化后文本在语义上与视频一一对应。为缓解数据集中存在不适定问题提供一种解决思路；3）在Charades-STA、QVHighlights（guery-based video highlights）、MSR-VTT数据集上的实验表明，本文方法在多个具有代表性的视频文本跨模态检索方法的基础上显著提升了其性能。

1　相关工作

1.1　视频检索

视频检索是视频理解的基本任务之一，在日常生活中有着广泛的应用。该任务的目标是在庞大的视频数据库中，根据给定的查询文本检索出最相关的视频内容。本文将现有研究大致分为两类，并对它们进行简要回顾。在视频检索领域，早期的一种趋势是采用现成的视频（Carreira和Zisserman，2017）和文本（Pennington等，2014）表示方法，结合复杂的特征编码技术或多模态融合机制，将文本和视频转换为固定维度的特征向量，使得计算相似性更高效。例如，Chen等人（2020）研究将视频文本匹配分解为全局到局部级别，提出了一种用于细粒度视频文本检索的模型。Yu等人（2018）利用分层注意力机制提出了一种不同模态顺序交互的联合序列融合模型。

由于对比式语言图像预训练（contrastive language-image pretraining，CLIP）模型（Radford等，2021）通过大量的图像文本对进行训练，展现了强大的泛化能力，许多工作开始将CLIP的知识迁移到视频检索任务。这些方法能够有效地利用图像和文本之间的语义关系，从而取得良好的性能。由于CLIP是图像文本模型，无法利用视频的时序信息，因此，现有方法通常还会设计额外的时序建模模块。具体而言，Gorti等人（2022）设计了一个跨模态注意力模型，通过文本缩放的点积注意力，关注其语义上最相似的帧，该模块使模型只关注以给定文本为条件的相关视频帧。Jin等人（2023）创造性地从生成的角度通过逐步生成文本和视频的联合概率分布，解决了当前判别式检索方法的局限性。Liu等人（2023）提出了一种时空辅助分支结构，可以同时考虑高级和低级的知识传递，有效地将CLIP模型扩展到不同的视频任务中。Fang等人（2023）在编码器中添加了额外的可学习令牌，以自适应地聚合多粒度语义，从而实现灵活的高层次推理。

1.2　视频片段检索

视频片段检索目的是从未修剪的视频中，根据给定的查询文本找到与这条查询文本语义最为相似

的目标片段。现有方法主要分为两类：基于预定义候选片段的方法和直接定位的方法。

1.2.1　基于预定义候选片段的方法

在早期的视频片段检索方法中，研究者通常依赖手工预定义的候选框架来提取视频片段。这些方法首先定义视频中的候选片段，候选片段通常是由行为检测器、目标检测器或分割器等工具自动生成的，用以代表视频的主要内容。随后，查询文本与候选片段进行对比，应用评分机制或注意力模型，从所有匹配的片段中筛选出最相关的部分。这一系列操作不仅需要精确地标定视频内容，还要高效地处理查询文本与视频数据之间的复杂关联。例如，Gao等人（2017）和Hendricks等人（2017）通过使用不同规模的滑动窗口来生成候选框，从有限粒度的滑动窗口中准确地定位动作。Zhang等人（2020）设计了一个二维时间图来枚举候选框，以覆盖不同长度的不同视频时刻，同时表示它们的相邻关系。然而，这些密集的候选框虽然可以减少搜索空间，但是同时引入了大量的冗余计算。

1.2.2　直接定位的方法

与基于预定义候选片段的方法不同，直接定位的方法不依赖于预定义的视频候选片段，而是直接从整个视频序列中学习匹配查询文本的内容。为了减少计算成本，近期工作主要围绕以下两种思路开展：1）直接预测每个视频帧作为目标片段开始和结束位置的概率。例如，Xu等人（2022）引入了特征解耦，以分离视频中的动作和背景因素，有效地减少了视频内部信息的相互干扰。Huang等人（2022）提出引导式注意机制，优化了基于段落内容的逐帧端点选择，增强了模型对标签不确定性的鲁棒性。2）利用可学习的查询预测出多个候选框。例如，Lei等人（2021）引入了一个端到端的基于Transformer的模型，将其视为候选框预测任务，消除了传统的预处理和后处理步骤的需要。Moon等人（2024）使用了带有虚拟令牌的自适应交叉注意力层，以选择性地增强视频片段和文本查询之间的相关性。

综上，随着深度学习技术的发展，许多研究在跨模态检索上取得了长足的进展。然而，这些方法都是针对模型改进、设计有效的模态交互模块或改进时序建模能力，却忽略了本文所指出的数据集本身存在的问题。即除了原始标注视频文本对之外，还存在其他视频与查询文本语义匹配的情况。在训练过程中，这类数据会在模型学习跨模态信息时产生混淆，从而限制模型性能。针对上述问题，本文研究并提出大语言模型引导的视频检索数据迭代优化方法，以获取更高质量的数据。

2　方　法

2.1　任务定义

本文定义包含N个样本对的视频文本检索数据集为$\boldsymbol{D}=\{(\boldsymbol{v}_i,\boldsymbol{q}_i)\}_{i=1}^{N}$，其中$\boldsymbol{v}_i$表示视频，$\boldsymbol{q}_i$则为$\boldsymbol{v}_i$对应的查询文本，即$\boldsymbol{q}_i$描述了视频中包含的内容或发生的事件。在视频检索任务中，给定一个查询文本$\boldsymbol{q}_i$，需要在视频集合$\boldsymbol{V}=\{\boldsymbol{v}_i\}_{i=1}^{N}$中，找到与查询文本$\boldsymbol{q}_i$语义最匹配的视频$\boldsymbol{v}_i$。而对于视频片段检索任务，给定一个视频和查询文本$\boldsymbol{q}$，目标是定位出与查询文本$\boldsymbol{q}$最相关的视频片段$\boldsymbol{m}=(t_\mathrm{s},\ t_\mathrm{e})$，其中$t_\mathrm{s}$和$t_\mathrm{e}$分别为开始和结束时间。

现有的视频文本检索方法通常认为数据集中的$\boldsymbol{v}_i$和$\boldsymbol{q}_i$存在一一对应关系。但实际上，由于人类标注的文本通常只关注视频动作，缺少对视频中场景、物体的细节描述，会存在$\boldsymbol{v}_j(j\neq i)$与$\boldsymbol{q}_i$在语义上更匹配的情况。在这种情况下，直接对模型进行训练，通常只考虑在特征空间中拉近$\boldsymbol{q}_i$与$\boldsymbol{v}_i$的距离，而将$\boldsymbol{v}_j$视为不匹配的样本，因此推远了$\boldsymbol{q}_i$与$\boldsymbol{v}_j$特征之间的距离，使模型产生语义混淆。本文的目标是将所有对应多个视频的查询文本$\boldsymbol{q}_i$进行优化，使其仅与$\boldsymbol{v}_i$匹配，且与所有的$\boldsymbol{v}_j(j\neq i)$在语义上都不匹配。最终，将优化后的文本用于模型训练。

2.2　总体算法框架

如图3所示，为解决在现有视频文本跨模态检索数据集训练时存在的不适定问题，本文提出一种大语言模型引导的视频检索数据迭代优化方法。首先，本文将数据集$\boldsymbol{D}$输入到视频文本困难样本对的自动筛选模块中得到困难样本集$\boldsymbol{D}_\mathrm{h}$和简单样本集$\boldsymbol{D}_\mathrm{e}$。对于$\boldsymbol{D}_\mathrm{h}$中的任意一个视频—查询文本对$(\boldsymbol{v}_i,\boldsymbol{q}_i)$，至少存在一个混淆视频$\boldsymbol{v}'\in\boldsymbol{D}$与$\boldsymbol{q}_i$的相似度大于$\boldsymbol{v}_i$和$\boldsymbol{q}_i$的相似度（为了方便阅读，下文将下标$i$省去）。接着将$\boldsymbol{D}_\mathrm{h}$中的$\boldsymbol{v}$和$\boldsymbol{v}'$输入到空间域场景物体细粒度信息提取模块，得到细粒度描述$\boldsymbol{c}_v$和$\boldsymbol{c}_{v'}$。最后，基于视频文本语义关联的迭代优化模块根据

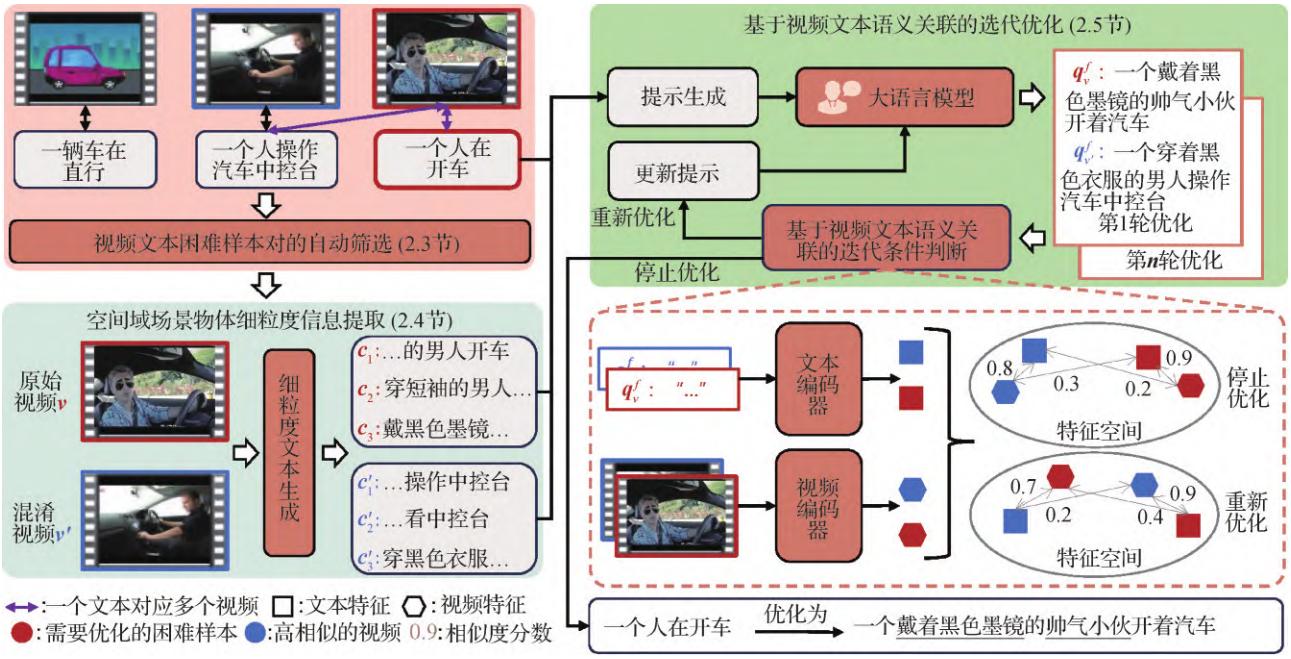

图3　大语言模型引导的视频检索数据迭代优化方法整体框架

Fig. 3　The overall framework of large language model-guided iterative optimization for video retrieval data

$\boldsymbol{c}_v$、$\boldsymbol{c}_{v'}$和原始标注$\boldsymbol{q}$生成提示并调用大语言模型得到优化文本$\boldsymbol{q}_v^f$，利用$\boldsymbol{q}_v^f$，$\boldsymbol{v}$，$\boldsymbol{v}'$间的语义关联进行自动决策。若认为优化后的文本满足预设定的迭代条件则停止优化文本，若不满足条件则利用当前优化结果更新提示并反馈给大语言模型重新优化。该模块充分利用大语言模型的推理总结和上下文学习能力将$\boldsymbol{q}_v^f$不断迭代直优化，得到更丰富且更独特的文本。

2.3　视觉文本困难样本对的自动筛选

视频文本检索数据集中通常存在多个视觉内容相似的视频，它们之间仅有细微的差别。故当查询文本$\boldsymbol{q}$为粗粒度文本描述时，除了与人工标注的目标视频$\boldsymbol{v}$相似度高外，还可能与其他视频相似度高，这种情况在训练时容易导致模型混淆。考虑到人工检查查询文本是否对应多个视频的过程既费时又成本高昂，本文提出一种视觉文本困难样本对的自动筛选方法。利用$\boldsymbol{q}$与所有候选视频$\boldsymbol{V}=\{\boldsymbol{v}_i\}_{i=1}^{N}$的相似度，将原始数据集$\boldsymbol{D}$划为难区分的困难样本集$\boldsymbol{D}_{\mathrm{h}}$和易区分的简单样本集$\boldsymbol{D}_{\mathrm{e}}$。主要思路如图4所示：如果$\boldsymbol{q}$与候选视频中相似度最高的不是目标视频$\boldsymbol{v}$，那么该数据对被归为$\boldsymbol{D}_{\mathrm{h}}$；反之，则归为$\boldsymbol{D}_{\mathrm{e}}$。本文采用余弦相似度$S$来计算相似度分数，筛选过程可表示为

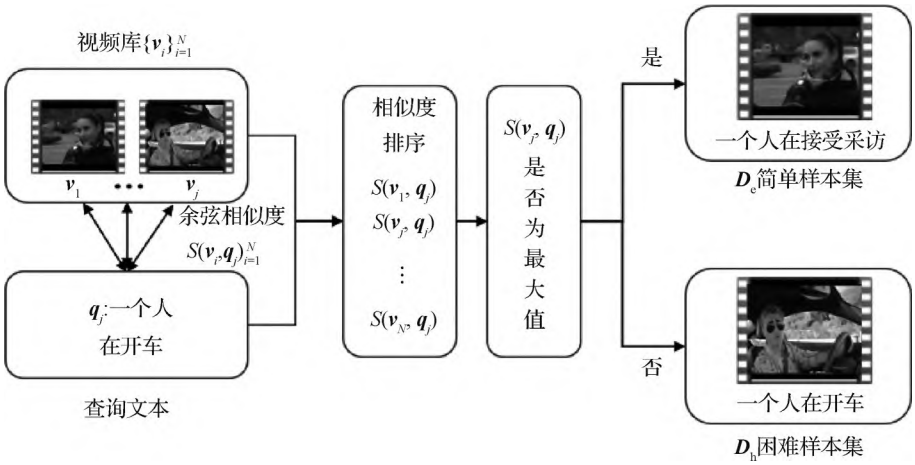

图4　视觉文本困难样本对的自动筛选机制

Fig. 4　Automatic screening mechanism for difficult samples pairs of visual text

$$\boldsymbol{D}_{\mathrm{h}}=\left\{\left(\boldsymbol{v}_i,\boldsymbol{q}_i\right)\right\}i,j\in[1,N];\exists i\neq j,S\left(\boldsymbol{v}_i,\boldsymbol{q}_i\right)\leqslant S\left(\boldsymbol{v}_j,\boldsymbol{q}_i\right)\Big\} \tag{1}$$

$$\boldsymbol{D}_{\mathrm{e}}=\left\{\left(\boldsymbol{v}_i,\boldsymbol{q}_i\right)\right\}i,j\in[1,N];\forall i\neq j,S\left(\boldsymbol{v}_i,\boldsymbol{q}_i\right)>S\left(\boldsymbol{v}_j,\boldsymbol{q}_i\right)\Big\} \tag{2}$$

2.4 空间域场景物体细粒度信息提取

如前文所述，粗粒度查询文本可能会对应多个视频，从而引起语义歧义问题。为解决此问题，希望能够挖掘目标视频中额外的信息，获取更详细的描述，并借助大语言模型的强大推理与总结能力，将这些细节描述整合到原始的粗粒度查询文本中，进而优化为细粒度查询文本。通过这一方式，可以将数据集中原本一对多的数据对转化为一一对应的关系。在本研究中，只对难以区分的困难样本集$\boldsymbol{D}_{\mathrm{h}}$进行优化处理。

为了从视频中捕捉更多细节描述，现有方法大多利用描述生成模型生成细粒度文本描述。图像描述生成模型倾向于描述图片中物体和场景等细粒度信息，而视频描述生成模型倾向于描述物体的动作信息及它们之间的时序关系。事实上，困难样本中的查询文本已经描述了视频的动作，只是由于其对物体和场景描述不足，从而导致其对应了多个视频。故希望描述生成模型能够捕获目标视频中细粒度的物体和场景信息，以便与混淆视频区分开来，将困难样本转化为一一对应的视频文本对。此外，相对于视频描述生成模型，图像描述生成模型计算量更低，推理速度更快。

因此，本文选择使用图像描述生成模型来为每个视频生成详细的帧级别描述。为了确保生成的文本描述能充分反映视频的细节，将目标视频$\boldsymbol{v}$均匀采样M帧输入到图像描述生成模型中，生成一系列详细文本描述$\boldsymbol{c}_v=\left\{\boldsymbol{c}_m\right\}_{m=1}^{M}$。困难样本存在混淆视频$\boldsymbol{v}'$比目标视频更匹配，其中$\boldsymbol{v}'\coloneqq\boldsymbol{v}_j,j=\underset{j\in[1,N]}{\operatorname{argmax}}S(\boldsymbol{v}_j,\boldsymbol{q})$，（“$\coloneqq$”表示定义为）。提取$\boldsymbol{v}'$的视觉信息，后续与$\boldsymbol{c}_v$、$\boldsymbol{q}$一起送入到大语言模型中推理总结，可使大语言模型更好地区分$\boldsymbol{v}'$和$\boldsymbol{v}$的区别。故对于$\boldsymbol{v}'$，同样均匀采样M帧输入到图像描述生成模型中得到相应的文本描述$\boldsymbol{c}_{v'}=\left\{\boldsymbol{c}_m\right\}_{m=1}^{M}$。

2.5 基于视频文本语义关联的迭代优化

由于视频中多帧的场景和物体信息可能是相同的，故提取到的细粒度文本描述$\boldsymbol{c}_v$是杂乱且冗余的。大语言模型（LLM）在自然语言处理领域取得了显著的进展，表现出强大的推理总结能力。因此，本文提出基于视频文本语义关联的迭代优化模块，其将大语言模型作为优化工具并用从海量语料中学习的知识提炼$\boldsymbol{c}_v$中关键信息，最后将这些细粒度文本信息整合到原始查询文本中。该模块主要包含提示生成、调用大语言模型以及更新优化提示功能。

2.5.1 查询文本优化

由于大语言模型的提示会影响优化结果的质量，因此本文设计的提示模板Q主要包含以下关键元素：1）清晰的任务描述：设置适当的背景可以让大语言模型了解优化文本的目的，进而提高语言模型的表现。2）必须满足的具体要求：在总结过程中提取由图像描述生成模型生成的多帧细粒度文本$\boldsymbol{c}_v$中的共性内容。其主要因为两点：首先，多帧中共现的内容往往承载着重要且关键的信息；其次，总结共现的内容可以有效避免引入错误的描述。例如，细粒度文本描述“戴黑色墨镜的人看着窗外”和“戴黑色墨镜的人正在开车”，戴黑色墨镜更有可能是较为重要的信息。3）提供满足条件的相关例子：提供几个优化示例是一种隐式的人类指导，可显著提高优化数据的质量。调用大语言模型优化查询文本可表示为

$$\boldsymbol{q}_v^f,\boldsymbol{q}_{v'}^f=LLM\left(\boldsymbol{q},\boldsymbol{c}_v,\boldsymbol{c}_{v'}\right) \tag{3}$$

需要注意的是，对于目标视频$\boldsymbol{v}$，由于原始查询$\boldsymbol{q}$与视频$\boldsymbol{v}$之间的高度相关性和准确性，因此要求大语言模型在总结时不改变$\boldsymbol{q}$的原意，在此基础上补充$\boldsymbol{c}_v$的共性信息。而对于混淆视频，由于不存在对应的查询文本，则通过直接总结详细文本描述$\boldsymbol{c}_{v'}$来生成混淆视频的新查询文本$\boldsymbol{q}_{v'}^f$。

2.5.2 基于视频文本语义关联的迭代条件

尽管大语言模型具备上下文理解和生成能力，但要在单轮的优化中获得高质量的查询文本仍面临较大挑战。在实际情况下，质量较高的输出通常来源于大语言模型与用户的多次交互，因为大语言模型可以根据用户反馈修改完善响应。同时，本文希望能够自动停止优化而不需要人为设定具体的对话轮次。受此启发，提出一种基于视频文本语义关联的迭代条件判断，以提高优化过程的可控性。该策略根据预先定义的规则来评判模型的输出结果，若结果符合要求则停止优化；若不满足条件，则基于当

前结果更新提示 Q 并反馈给大语言模型，利用大语言模型强大的整合上下文信息和学习能力将其重新优化为符合条件的文本。本文的目标是通过引入具有区分度的细粒度描述，实现视频—查询文本对的一一对应。为此，如图5所示，设计一个基于视频文本语义关联的迭代条件。主要思想是：1）优化后的文本 $\boldsymbol{q}_v^f$ 与原始视频 $\boldsymbol{v}$ 更匹配，而与混淆视频 $\boldsymbol{v}'$ 形成较大的语义差异；2）$\boldsymbol{v}'$ 新生成的文本 $\boldsymbol{q}_{v'}^f$ 应当更匹配 $\boldsymbol{v}'$ 且无法准确描述 $\boldsymbol{v}$；3）$\boldsymbol{q}_v^f$ 较于 $\boldsymbol{q}_{v'}^f$ 与 $\boldsymbol{v}$ 语义上更匹配，$\boldsymbol{v}'$ 与 $\boldsymbol{q}_{v'}^f$ 的关联程度要大于 $\boldsymbol{q}_v^f$。具体来说，将新生成的数据对 $(\boldsymbol{v}, \boldsymbol{q}_v^f)$ 和 $(\boldsymbol{v}', \boldsymbol{q}_{v'}^f)$ 通过文本和视频编码器进行编码，使其处于同一特征空间。同时，其特征相似度应符合以下语义关联迭代条件，即

$$\begin{cases} S\left(\boldsymbol{v}, \boldsymbol{q}_v^f\right) > S\left(\boldsymbol{v}', \boldsymbol{q}_v^f\right) \\ S\left(\boldsymbol{v}, \boldsymbol{q}_{v'}^f\right) > S\left(\boldsymbol{v}', \boldsymbol{q}_{v'}^f\right) \\ S\left(\boldsymbol{v}, \boldsymbol{q}_v^f\right) > S\left(\boldsymbol{v}, \boldsymbol{q}_{v'}^f\right) \\ S\left(\boldsymbol{v}', \boldsymbol{q}_v^f\right) > S\left(\boldsymbol{v}', \boldsymbol{q}_{v'}^f\right) \end{cases} \tag{4}$$

当相似度不满足该条件时，将大语言模型在上一轮输出的 $\boldsymbol{q}_v^f$，$\boldsymbol{q}_{v'}^f$ 作为启发式提示 Q' 和细粒度信息 $\boldsymbol{c}_v$，$\boldsymbol{c}_{v'}$ 以及原查询文本 $\boldsymbol{q}$ 一起送回到大语言模型中重新优化，即

$$\boldsymbol{q}_v^f, \boldsymbol{q}_{v'}^f = LLM\left(\boldsymbol{q}, \boldsymbol{c}_v, \boldsymbol{c}_{v'}, \boldsymbol{q}_v^f, \boldsymbol{q}_{v'}^f\right) \tag{5}$$

直至优化后的文本符合上述跨模态匹配条件。该模块可以让大语言模型对过去的输出进行自我反思并从错误中总结经验，确保优化后查询文本的高精确匹配。虽然视频检索领域存在大语言模型润色文本的案例，但本文方法与这些方法的主要区别在于：已有方法通常使用大语言模型进行单次优化，而本文针对视频检索和视频片段检索的特点，设计了一种新的查询文本迭代优化方法，可以根据视频和优化后文本的语义关联性对大语言模型形成反馈，进行多轮迭代式的查询文本优化。

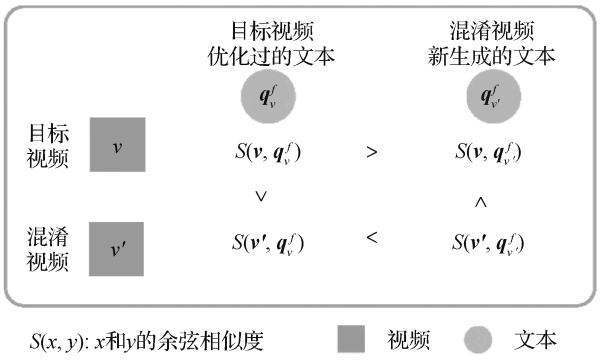

图5　视频文本关联迭代条件图示

Fig. 5　Illustration of iterative conditions for video-text association

2.6　不同视频文本检索任务上的应用策略

2.6.1　视频检索

对于视频检索任务而言，现有方法大多数是利用预先训练的图像文本基础模型，并将其强大的表示能力迁移到视频领域。其中，以图像文本对齐为目标进行训练的CLIP模型特别适合视频文本检索。具体来说，该模型通过两个编码器分别提取视频帧和文本特征向量，并采用对比学习的策略在特征空间内拉近同一批次匹配的视频—文本对，同时将非匹配的视频—文本对推远（张浩宇 等，2022；贺超和魏宏喜，2023）。由于训练时每个批次的大小远小于整体训练集的大小，因此尽管训练数据集中存在一对多的困难样本，但将特定的困难样本（目标视频与混淆视频）置于同一批次中训练概率很小。然而，由于测试阶段是在整个测试集中进行检索，因此在训练和测试之间存在的这种差异会大大限制模型性能。

为了缓解这一问题，本文在每个训练轮次（epoch）中实施一种训练策略，即在每一轮选择部分迭代，将优化后样本的目标视频与混淆视频置于同一批次进行训练，同时保持其余迭代正常训练。如果模型能够有效区分目标视频与混淆视频，那么视频与文本的表征将更为对齐，同时能够减少训练与测试间的差异。每个批次包含困难样本的数量计算为

$$n_{\mathrm{h}}^b = n_{\mathrm{h}} / (n_t \times p / b) \tag{6}$$

式中，n_t 是训练数据的总数，n_{h}^b 表示每个批次中困难样本的数量，n_{h} 是困难样本的总数量，p 为将困难样本的目标视频与混淆视频置于同一批次中的迭代比率，b 代表批次大小。这种方法能够确保训练过程中提高模型从困难样本中学习跨模态匹配关系的效率，进而提高检索性能。

2.6.2　视频片段检索

对于视频片段检索任务，视频片段检索的对象是视频中的某个与查询文本 $\boldsymbol{q}$ 对应的片段。因此，为筛选困难样本，首先需在视频中找出除标注片段外与 $\boldsymbol{q}$ 存在较强语义关联的片段。为实现这一目的，本文将完整/未剪辑视频解码成逐帧序列，并采用聚类算法将视频帧聚为若干个视频片段。具体来

说，给定一个未剪辑视频 $\boldsymbol{U}=\{\boldsymbol{f}_i\}_{i=1}^{T}$，模型首先逐帧提取视频特征，然后计算帧间相似度关系，最后使用K-means聚类算法（Li和Wu，2012）聚集具有高度相似性的视频帧，从而形成候选视频片段集合 $\boldsymbol{H}=\{\boldsymbol{h}_i\}_{i=1}^{N}$。后续处理流程基本相同，区别在于上述处理单元从视频 $\boldsymbol{v}$ 变为片段 $\boldsymbol{h}$。

算法1 大语言模型引导的视频检索数据迭代优化方法

输入：训练数据集 $\boldsymbol{D}=\{\boldsymbol{v}_i,\boldsymbol{q}_i\}_{i=1}^{N}$，困难样本集 $\boldsymbol{D}_{\mathrm{h}}$ 和简单样本集 $\boldsymbol{D}_{\mathrm{e}}$；

辅助信息：$\boldsymbol{v}$、$\boldsymbol{v}'$、$\boldsymbol{q}$ 分别为视频、混淆视频、查询文本；$\boldsymbol{c}_v$、$\boldsymbol{c}_{v'}$、$\boldsymbol{q}_v^f$、$\boldsymbol{q}_{v'}^f$ 分别为视频和混淆视频对应的细粒度信息和优化后的文本；S、C、LLM 分别为余弦相似度、图像描述生成模型、大语言模型；

输出：优化后的数据集 $\tilde{\boldsymbol{D}}=\widetilde{\boldsymbol{D}_{\mathrm{h}}}\cup\boldsymbol{D}_{\mathrm{e}}$；

1）for i to N do；
2）　if $\forall j\neq i, S(\boldsymbol{v}_i,\boldsymbol{q}_i)>S(\boldsymbol{v}_j,\boldsymbol{q}_i)$ then；
3）　　更新简单样本集 $\boldsymbol{D}_{\mathrm{e}}\coloneqq\boldsymbol{D}_{\mathrm{e}}\cup(\boldsymbol{v}_i,\boldsymbol{q}_i)$；
4）　else；
5）　　更新困难样本集 $\boldsymbol{D}_{\mathrm{h}}\coloneqq\boldsymbol{D}_{\mathrm{h}}\cup(\boldsymbol{v}_i,\boldsymbol{q}_i)$；
6）　end if；
7）end for；
8）for $i=1$ to $|\boldsymbol{D}_{\mathrm{h}}|$ do；
9）　$\boldsymbol{v}'=\boldsymbol{v}_j, j=\underset{j\in[1,N]}{\operatorname{argmax}} S(\boldsymbol{v}_j,\boldsymbol{q}_i)$；
10）　$\boldsymbol{c}_{v_i}=C(\boldsymbol{v}_i), \boldsymbol{c}_{v'}=C(\boldsymbol{v}')$；
11）　$(\boldsymbol{q}_{v_i}^f,\boldsymbol{q}_{v'}^f)=LLM(\boldsymbol{q}_i,\boldsymbol{c}_{v_i},\boldsymbol{c}_{v'})$；
12）　while 不满足语义关联迭代条件式(4) do；
13）　　$(\boldsymbol{q}_{v_i}^f,\boldsymbol{q}_{v'}^f)=LLM(\boldsymbol{q}_i,\boldsymbol{c}_{v_i},\boldsymbol{c}_{v'},\boldsymbol{q}_{v_i}^f,\boldsymbol{q}_{v'}^f)$；
14）　end while；
15）　替换 $\boldsymbol{D}_{\mathrm{h}}$ 的查询文本 $\boldsymbol{q}_i\leftarrow\boldsymbol{q}_{v_i}^f$，得到 $\widetilde{\boldsymbol{D}_{\mathrm{h}}}$；
16）end for。

3 实　验

3.1 数据集

本文在两个典型视频文本检索任务上对所提方法进行验证，任务包括视频片段检索和视频检索。

3.1.1 视频片段检索数据集

1）Charades-STA数据集（Sigurdsson等，2016）包含6 672个视频，展示了多种日常室内活动。在训练和测试中，分别有12 408个和3 720个视频片段—查询文本对。查询文本的平均长度为8.6个单词，视频的平均时长为29.8 s，每个视频中的平均活动数量为2.3个。

2）QVHighlights数据集（Lei等，2021）包含10 148个视频，涵盖了18 367个片段和10 310个文本描述。查询文本的平均长度为11.3个单词，视频的平均持续时间为150 s，每个片段平均持续24.6 s。本文使用7 218个样本作为训练集，以及1 550个样本作为验证集。

3.1.2 视频检索数据集

MSR-VTT数据集（Xu等，2016）包含来自YouTube视频网站的10 000个视频，每个视频的长度介于10~30 s之间，并附有20条英文描述。这个数据集涵盖了体育、生活等多种场景，是视频检索领域中最常使用的标准数据集之一。本文采用与已有工作（Gabeur等，2020）同样的设置，9 000个视频作为训练集，1 000个视频作为测试集。

3.2 实验细节

本文采用预先训练的InternVideo模型（Wang等，2022）从视频片段/视频和查询文本中提取特征并计算其相似度，用于划分困难和简单样本集，采样帧数设置为8。对于图像描述生成模型，采用认知视觉语言模型（cognitive visual language model，cogvlm）（Wang等，2024）模型，采样帧数 M 设置为3。大语言模型采用GPT4。视频检索训练时，将 p 设置为0.5。同时采用预训练的模型CLIP（ViT-B/32）初始化主干网络。消融实验中，视频描述生成模型采用VideoChat（Li等，2024）。

3.3 评价指标

对于视频检索任务，本文使用召回率R@k评测指标，即前 k 个检索结果中，能够检索出相关视频的比率。在视频片段检索领域，采用Gao等人（2017）方法来计算评价指标“R@n, IoU = a”。该指标反映了在所有相关样本中，有多少比率的样本在前 n 个检索结果中被正确识别，并且其时间重叠度至少达到 a。IoU代表了预测时间跨度和真实时间跨度的交并比。

3.4 不同模型在多个数据集的性能比较

3.4.1 Charades-STA数据集实验结果

为了公平比较，采用与其他方法相同的实验设

置，并使用与已有方法论文中描述相同的特征，即使用 Visual Geometry Group（VGG）（Simonyan 和 Zisserman，2015）和 Inflated 3D ConvNet（I3D）（Carreira 和 Zisserman，2017）模型抽取的特征。唯一区别在于本文将困难样本替换成本文优化后的样本进行训练。在表 1 中，对所提方法与多种先进方法进行了对比实验，包括 2D temporal adjacent networks（2D-TAN）、moment alignment network（MAN）、fast video moment retrieval（FVMR）、unified multi-modal Transformers（UMT）、query-dependent detection Transformer（QDDETR）、task-reciprocal detection Transformer（TRDETR）、elastic moment bounding（EMB）和 MomentDiff、correlation-guided detection Transformer（CGDETR）。使用优化后的样本进行训练的模型在检索性能上都有显著提升。当 R@1，IoU = 0.5 时，性能最大提升了 3.23%，最少也有 1.55% 的提升。当 IoU = 0.7 时，性能至多有 4.38% 的提升。该实验结果证明了本文方法的有效性。

表 1　训练样本优化前后在 Charades-STA 数据集的性能对比

Table 1　Performance comparison on the Charades-STA dataset before and after training sample optimization

方法	R@1/%	
	IoU = 0.5	IoU = 0.7
2D-TAN（Zhang 等，2020）	39.81	22.85
MAN（Zhang 等，2019）	41.24	20.54
FVMR（Gao 和 Xu，2021）	42.36	24.14
UMT（Liu 等，2022a）	48.31	29.25
QDDETR（Moon 等，2023）	52.77	31.13
TRDETR（Sun 等，2024）	53.47	30.81
EMB（Huang 等，2022）	58.58	38.95
MomentDiff（Li 等 2023）	51.42	27.42
CGDETR（Moon 等，2024）	55.43	33.33
EMB + 优化后数据训练	**60.13(+1.55)**	**40.73(+1.78)**
MomentDiff + 优化后数据训练	**54.65(+3.23)**	**31.80(+4.38)**
CGDETR + 优化后数据训练	**57.90(+2.47)**	**34.49(+1.16)**

注：加粗字体为参考方法使用优化后数据训练的结果。

3.4.2　QVHighlights 数据集实验结果

QVHighlights 数据集提供了手工逐帧标注的显著性分数，在训练时存在两种方式：1）使用手工标注的显著性分数作为额外的监督信息；2）将显著性分数全部置 1。本文采用第 2 种。在表 2 中，本文对 event-aware Transformer（EATR）、modal-enhanced semantic modeling（MESM）、Moment detection Transformer（MomentDETR）、QDDETR 与本文方法进行了性能对比实验。利用本文优化后的样本训练，当 R@1，IoU = 0.5 时，MomentDETR，QDDETR 方法性能分别取得了 5.48% 和 1.35% 的显著提升。当 IoU = 0.7 时，两种方法性能平均提升 3%。实验结果充分证明本文方法为数据驱动改进模型性能的范式提供了新思路。

表 2　训练样本优化前后在 QVHighlights 数据集的性能对比

Table 2　Performance comparison on the QVHighlights dataset before and after training sample optimization

方法	R@1/%	
	IoU = 0.5	IoU = 0.7
EATR（Zhang 等，2020）	61.36	45.79
MESM（Liu 等，2024）	62.78	45.20
MomentDETR（Lei 等，2021）	53.55	33.68
QDDETR（Moon 等，2023）	61.23	43.55
MomentDETR + 优化后数据训练	**59.03 (+5.48)**	**38.65 (+4.97)**
QDDETR + 优化后数据训练	**62.58 (+1.35)**	**44.58 (+1.03)**

注：加粗字体为参考方法使用优化后数据训练的结果。

3.4.3　MSR-VTT 数据集实验结果

MSR-VTT 包含 18 万个视频—查询文本对，其中检测出来的困难样本有 14 万个，如果优化所有困难样本将十分耗时。因此，本文通过优化其中最困难的一部分文本，来提升模型性能。具体而言，困难样本与混淆视频的相似度会大于目标视频，可计算出一个相似度差值，按相似度差值排序选取前 r 个进行优化。如图 6 所示，将所有差值排序后进行了可视化，最后取 r 为 10 000 个进行优化。在表 3 中，对所提方法与多种先进方法进行了对比实验，包括 CLIP for end to end video clip retrieval（CLIP4 Clip）、token shift and selection network（TS2-Net）、X-CLIP、spatial-temporal auxiliary network（STAN）、Prompt Switch 和 uncertainty-adaptive text-video retrieval（UATVR）。当指标为 R@1 时，经过优化样本训练后

的模型性能最高提升了 1.6%，最低提升了 1.2%。说明本文方法在不同的视频文本检索任务上都能获得显著提升，并不局限于特定的任务和数据集。

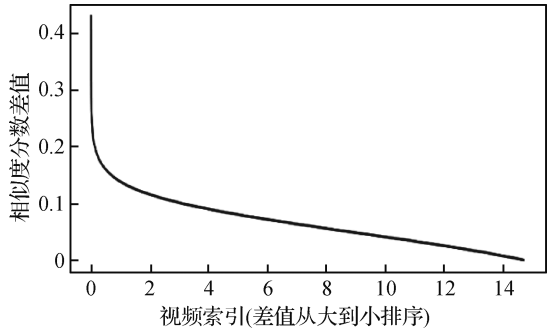

图 6　MSR-VTT 困难样本中查询文本与混淆视频/目标视频的相似度分数差值图示

Fig. 6　Illustration of the difference in similarity scores between the query text and the obfuscated video/target video in the MSR-VTT difficult sample

表 3　训练样本优化前后在 MSR-VTT 数据集的性能对比

Table 3　Performance comparison on the MSR-VTT dataset before and after training sample optimization

方法	R@1/%
CLIP4Clip（Luo 等，2022）	44.5
TS2-Net（Liu 等，2022b）	47.0
X-CLIP（Ma 等，2022）	46.1
STAN（Liu 等，2023）	46.9
Prompt Switch（Deng 等，2023）	47.8
UATVR（Fang 等，2023）	46.9
Xpool（Gorti 等，2022）	46.6
UATVR + 优化后数据训练	**48.1 (+1.2)**
Xpool + 优化后数据训练	**48.2 (+1.6)**

注：加粗字体为参考方法使用优化后数据训练的结果。

3.5　消融实验

3.5.1　迭代优化模块的有效性分析

在利用大语言模型对困难样本进行优化时，本文设计了一个基于视频文本语义关联的迭代优化模块，为了验证其有效性，设计如下实验：用/不用迭代优化模块去总结所获得的细粒度信息。在不使用迭代优化模块的情况下，仅要求大语言模型输出第 1 轮的优化结果，未将不符合条件的文本反馈给大语言模型重新优化。如表 4 所示，在 Charades-STA 数据集中，当 R@1，IoU = 0.5 时，MomentDiff 和 CGDETR 方法使用迭代优化比不使用时性能分别提高了 1.29% 和 0.78%。结果表明，迭代优化模块通过不断反馈和优化，能够充分利用大语言模型的上下文理解能力，使大语言模型能够更准确地捕捉和整合细粒度信息，提高优化文本的质量，从而有效提高模型的检索性能。

表 4　迭代优化模块在 Charades-STA 数据集的性能评估

Table 4　Performance evaluation of iterative optimization module on the Charades-STA dataset

方法	实验设置	R@1/%, IoU = 0.5
MomentDiff	不使用迭代优化	53.36
	使用迭代优化	54.65
CGDETR	不使用迭代优化	57.12
	使用迭代优化	57.90

3.5.2　图像和视频描述生成模型性能对比

在生成视频描述时，可以选择使用图像描述生成模型和视频描述生成模型来获取视频的细节信息。本文对比了两种类型的描述生成模型的性能。如表 5 所示，在 Charades-STA 数据集中，当评估指标为 R@1，IoU = 0.5 时，两种方法使用图像描述生成模型的性能较视频描述生成模型平均提升了 2.19%。可能原因是视频描述生成模型更加关注目标的动作，而困难样本更需要补充细粒度的场景和物体信息。使用的图像描述生成模型生成的文本能够更准确地描述物体外观和场景等信息，为大语言模型的总结和优化奠定了基础。

表 5　Charades-STA 数据集上采用不同描述生成模型的性能对比

Table 5　Performance comparison of different text generation models on the Charades-STA dataset

方法	实验设置	R@1/%, IoU = 0.5
MomentDiff	视频—文本	50.97
	图像—文本	54.65
CGDETR	视频—文本	57.20
	图像—文本	57.90

3.5.3　大语言模型总结的有效性分析

事实上，仅用描述生成模型就可以得到困难样本中目标视频的细粒度信息。为了评估大语言模型总结细粒度信息的有效性，本文设计如下实验：

用/不用大语言模型总结细粒度信息，在不使用大语言模型的情况下，仅提取视频中间帧输入到描述生成模型，将生成的对应描述作为查询文本，后续不进行总结和优化，直接用于模型训练。如表6所示，在Charades-STA数据集中，当评估指标为R@1，IoU = 0.5时，使用大语言模型总结细粒度信息的性能较不使用大语言模型至多提升了2.47%，至少提升了1.58%。结果表明，大语言模型通过其强大的推理和总结能力，能够有效过滤冗余信息，保留重要的细粒度描述，并将这些信息整合到原始查询文本中，生成语义更匹配的描述，从而提高模型的检索性能。

表6　Charades-STA数据集上大语言模型总结模块的性能评估

Table 6　Performance evaluation of the summarization module of a large language model on the Charades-STA dataset

方法	大语言模型文本优化	R@1/%, IoU = 0.5
MomentDiff	×	52.18
	√	54.65
CGDETR	×	56.32
	√	57.90

注："√"和"×"分别表示已采用和未采用。

3.5.4　不同迭代比率对性能的影响

对于视频检索任务，由于不适定问题仅存在于测试过程，导致训练与测试存在差异。故本文在训练中每一轮次的部分迭代，将优化后的查询文本所对应的目标视频和混淆视频置于同一批次训练以弥补这一差异。此过程占每一轮次的比率是可调节的。因此，本文研究了不同迭代比率对模型性能的影响。如图7所示，对于MSR-VTT数据集，模型性能在比率0.5达到峰值，随后，性能随着比率的增加而下降。可能原因是迭代比率过小不足以弥补训练和测试之间的差异。迭代比率过大，将目标视频和混淆视频放在一个批次里训练增加了任务难度。故本文将不同迭代比率设置为0.5。

3.5.5　优化时间分析

如表7中所示，在单张RTX4090 GPU上，本文方法平均5.92 s就可以优化一个困难样本。且由于整个过程不需要人工干预，因此避免了大量人力投入。更重要的是，优化文本只需执行一次，所产生的数据

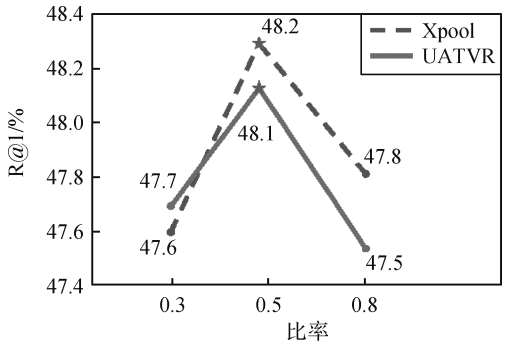

图7　在MSR-VTT数据集上不同迭代比率对性能的影响

Fig. 7　The impact of different iteration ratios on performance on the MSR-VTT dataset

表7　优化时间分析

Table 7　Optimization time analysis

生成模型/s	大语言模型/s	总计/s
5.66	0.26	5.92

可以增强各种跨模态检索模型的性能，进一步展现了本文方法生成数据的可复用性。

3.5.6　目标视频原查询文本和改进文本示例

本文在图8中展示了优化后查询文本的一些示例。可以发现，优化前的查询文本除了与目标视频匹配，也适用于描述混淆视频。例如"乐队在舞台上表演"，既没有描述这个舞台的风格，也没有描述这个乐队的具体情况。这种粗粒度的文本描述显然过于宽泛，无法准确区分相似的视频。而本文方法能够提取视频的各个方面的细节以改进粗粒度文本。例如，通过提取视频中的详细物品（如蔬菜和葡萄酒）、人物的外观（如服装）以及舞台布置识别出的音乐风格（如古典交响乐）等细节信息，使得查询文本更加具体和精确。通过补充这些细粒度信息，改进后的查询文本不仅能够更加准确地描述目标视频，还能有效区分混淆视频。

4　结　论

本文针对现有视频文本跨模态数据集中存在一对多的查询文本，进而导致训练过程中存在不适定问题的情况，提出了一种大语言模型引导的视频检索数据迭代优化方法。该方法通过图像描述生成模型提取困难样本视频的细粒度信息，并利用大语言模型对这些信息进行总结推理，将困难样本迭代优

图 8　目标视频原查询文本和改进文本的示例

Fig. 8　Example of original query text and improved text for the target video

化为更匹配的细粒度查询文本，以缓解困难样本训练时使模型混淆的问题，进而提升模型性能。实验结果表明，本文方法对现有视频片段检索算法在 Charades-STA 和 QVHighlights 数据集上 R@1，IoU = 0.5 分别最高有 3.23% 与 5.48% 的提升，对现有视频检索算法在 MSR-VTT 上 R@1 最高有 1.6% 的提升，所提方法为跨模态视频文本检索的研究提供了一种新的思路。

在使用本文方法对数据集进行一次优化后，便可在多个方法提升模型性能，但是仍存在一些局限。具体而言，本文提出的方法使用了大语言模型进行细粒度信息推理总结，故整体运行时间受到大语言模型的运行速度的影响。在今后的研究中，可以改进优化流程，避免一些冗余的接口调用。同时，本文方法的框架不局限于特定的描述生成模型和大语言模型，未来可与效率更高效果更好的模型结合，进一步提升优化效率和跨模态检索模型性能。

参考文献(References)

Carreira J and Zisserman A. 2017. Quo vadis, action recognition? A new model and the kinetics dataset//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 4724-4733 [DOI: 10.1109/CVPR.2017.502]

Chen L, Xi Y M and Liu L B. 2024. Survey on video-text cross-modal retrieval. Computer Engineering and Applications, 60(4): 1-20 (陈磊, 习怡萌, 刘立波. 2024. 视频文本跨模态检索研究综述. 计算机工程与应用, 60(4): 1-20) [DOI: 10.3778/j.issn.1002-8331.2306-0382]

Chen S Z, Zhao Y D, Jin Q and Wu Q. 2020. Fine-grained video-text retrieval with hierarchical graph reasoning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 10635-10644 [DOI: 10.1109/CVPR42600.2020.01065]

Deng C R, Chen Q, Qin P D, Chen D and Wu Q. 2023. Prompt switch: efficient CLIP adaptation for text-video retrieval//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 15602-15612 [DOI: 10.1109/ICCV51070.2023.01434]

Fang B, Wu W H, Liu C, Zhou Y, Song Y X, Wang W P, Shu X B, Ji X Y and Wang J D. 2023. UATVR: uncertainty-adaptive text-video retrieval//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 13677-13687 [DOI: 10.1109/ICCV51070.2023.01262]

Gabeur V, Sun C, Alahari K and Schmid C. 2020. Multi-modal Transformer for video retrieval//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 214-229 [DOI: 10.1007/978-3-030-58548-8_13]

Gao J Y, Sun C, Yang Z H and Nevatia R. 2017. TALL: temporal activity localization via language query//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 5277-5285 [DOI: 10.1109/ICCV.2017.563]

Gao J Y and Xu C S. 2021. Fast video moment retrieval//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 1503-1512 [DOI: 10.1109/ICCV48922.2021.00155]

Gorti S K, Vouitsis N, Ma J W, Golestan K, Volkovs M, Garg A and

Yu G W. 2022. X-pool: cross-modal language-video attention for text-video retrieval//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 4996-5005 [DOI: 10.1109/CVPR52688.2022.00495]

Hadamard J. 1902. Sur les problèmes aux dérivées partielles et leur signification physique. Princeton University Bulletin, 13: 49-52

He C and Wei H X. 2023. Image retrieval based on Transformer and asymmetric learning strategy. Journal of Image and Graphics, 28(2): 535-544 (贺超, 魏宏喜. 2023. 结合Transformer与非对称学习策略的图像检索. 中国图象图形学报, 28(2): 535-544) [DOI: 10.11834/jig.210842]

Hendricks L A, Wang O, Shechtman E, Sivic J, Darrell T and Russell B. 2017. Localizing moments in video with natural language//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 5804-5813 [DOI: 10.1109/ICCV.2017.618]

Huang J B, Jin H L, Gong S G and Liu Y. 2022. Video activity localisation with uncertainties in temporal boundary//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 724-740 [DOI: 10.1007/978-3-031-19830-4_41]

Jin P, Li H, Cheng Z S, Li K H, Ji X Y, Liu C, Yuan L and Chen J. 2023. DiffusionRet: generative text-video retrieval with diffusion model//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 2470-2481 [DOI: 10.1109/ICCV51070.2023.00234]

Lei J, Berg T L and Bansal M. 2021. QVHIGHLIGHTS: detecting moments and highlights in videos via natural language queries//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: Curran Associates: 11846-11858

Li K C, He Y N, Wang Y, Li Y Z, Wang W H, Luo P, Wang Y L, Wang L M and Qiao Y. 2024. VideoChat: chat-centric video understanding [EB/OL]. [2024-09-04]. https://arxiv.org/pdf/2305.06355.pdf

Li P D, Xie C W, Xie H T, Zhao L M, Zhang L, Zheng Y, Zhao D L and Zhang Y D. 2023. MomentDiff: generative video moment retrieval from random to real//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 65948-65966

Li Y G and Wu H Y. 2012. A clustering method based on K-means algorithm. Physics Procedia, 25: 1104-1109 [DOI: 10.1016/j.phpro.2012.03.206]

Liu H F, Chen J J, Li L, Bao B K, Li Z C, Liu J Y and Nie L Q. 2023. Cross-modal representation learning and generation. Journal of Image and Graphics, 28(6): 1608-1629 (刘华峰, 陈静静, 李亮, 鲍秉坤, 李泽超, 刘家瑛, 聂礼强. 2023. 跨模态表征与生成技术. 中国图象图形学报, 28(6): 1608-1629) [DOI: 10.11834/jig.230035]

Liu R Y, Huang J J, Li G, Feng J S, Wu X L and Li T H. 2023. Revisiting temporal modeling for CLIP-based image-to-video knowledge transferring//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 6555-6564 [DOI: 10.1109/CVPR52729.2023.00634]

Liu Y, Cheng M, Wang F P, Li D X, Liu W and Fan J L. 2020. Deep Hashing image retrieval methods. Journal of Image and Graphics, 25(7): 1296-1317 (刘颖, 程美, 王富平, 李大湘, 刘伟, 范九伦. 2020. 深度哈希图像检索方法综述. 中国图象图形学报, 25(7): 1296-1317) [DOI: 10.11834/jig.190518]

Liu Y, Li S Y, Wu Y, Chen C W, Shan Y and Qie X H. 2022a. UMT: unified multi-modal Transformers for joint video moment retrieval and highlight detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 3032-3041 [DOI: 10.1109/CVPR52688.2022.00305]

Liu Y Q, Xiong P F, Xu L H, Cao S M and Jin Q. 2022b. TS2-net: token shift and selection Transformer for text-video retrieval//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 319-335 [DOI: 10.1007/978-3-031-19781-9_19]

Liu Z H, Li J, Xie H T, Li P D, Ge J N, Liu S A and Jin G Q. 2024. Towards balanced alignment: modal-enhanced semantic modeling for video moment retrieval//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 3855-3863 [DOI: 10.1609/aaai.v38i4.28177]

Luo H S, Ji L, Zhong M, Chen Y, Lei W, Duan N and Li T R. 2022. CLIP4Clip: an empirical study of CLIP for end to end video clip retrieval and captioning. Neurocomputing, 508: 293-304 [DOI: 10.1016/j.neucom.2022.07.028]

Ma Y W, Xu G H, Sun X S, Yan M, Zhang J and Ji R R. 2022. X-CLIP: end-to-end multi-grained contrastive learning for video-text retrieval//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portugal: ACM: 638-647 [DOI: 10.1145/3503161.3547910]

Moon W, Hyun S, Lee S and Heo J P. 2024. Correlation-guided query-dependency calibration for video temporal grounding [EB/OL]. [2024-09-04]. https://arxiv.org/pdf/2311.08835.pdf

Moon W, Hyun S, Park S, Park D and Heo J P. 2023. Query-dependent video representation for moment retrieval and highlight detection//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 23023-23033 [DOI: 10.1109/CVPR52729.2023.02205]

Pennington J, Socher R and Manning C D. 2014. Glove: global vectors for word representation//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: ACL: 1532-1543 [DOI: 10.3115/v1/D14-1162]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I.

2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR: 8748-8763

Sigurdsson G A, Varol G, Wang X L, Farhadi A, Laptev I and Gupta A. 2016. Hollywood in homes: crowdsourcing data collection for activity understanding//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 510-526 [DOI: 10.1007/978-3-030-58568-6_39]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2024-09-04]. https://arxiv.org/pdf/1409.1556.pdf

Sun H, Zhou M Y, Chen W J and Xie W. 2024. TR-DETR: task-reciprocal Transformer for joint moment retrieval and highlight detection//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 4998-5007 [DOI: 10.1609/aaai.v38i5.28304]

Wang W H, Lv Q S, Yu W M, Hong W Y, Qi J, Wang Y, Ji J H, Yang Z Y, Zhao L, Song X X, Xu J Z, Chen K Q, Xu B, Li J Z, Dong Y X, Ding M and Tang J. 2024. CogVLM: visual expert for pretrained language models [EB/OL]. [2024-09-04]. https://arxiv.org/pdf/2311.03079.pdf

Wang Y, Li K C, Li Y Z, He Y N, Huang B K, Zhao Z Y, Zhang H J, Xu J L, Liu Y, Wang Z, Xing S, Chen G, Pan J T, Yu J S, Wang Y L, Wang L M and Qiao Y. 2022. InternVideo: general video foundation models via generative and discriminative learning [EB/OL]. [2024-09-04]. https://arxiv.org/pdf/2212.03191.pdf

Wang Y G, Kang X D, Guo J, Li B, Zhang H L and Liu H Q. 2020. Image Hash retrieval with DenseNet. Journal of Image and Graphics, 25(5): 900-912 (王亚鸽, 康晓东, 郭军, 李博, 张华丽, 刘汉卿. 2020. 密集网络图像哈希检索. 中国图象图形学报, 25(5): 900-912) [DOI: 10.11834/jig.190416]

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

Xu Z, Chen D, Wei K, Deng C and Xue H. 2022. HiSA: hierarchically semantic associating for video temporal grounding. IEEE Transactions on Image Processing, 31: 5178-5188 [DOI: 10.1109/TIP.2022.3191841]

Yin Q Y, Huang Y, Zhang J G, Wu S and Wang L. 2021. Survey on deep learning based cross-modal retrieval. Journal of Image and Graphics, 26(6): 1368-1388 (尹奇跃, 黄岩, 张俊格, 吴书, 王亮. 2021. 基于深度学习的跨模态检索综述. 中国图象图形学报, 26(6): 1368-1388) [DOI: 10.11834/jig.200862]

Yu Y, Kim J and Kim G. 2018. A joint sequence fusion model for video question answering and retrieval//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 487-503 [DOI: 10.1007/978-3-030-01234-2_29]

Zhang D, Dai X Y, Wang X, Wang Y F and Davis L S. 2019. Man: moment alignment network for natural language moment retrieval via iterative graph adjustment//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 1247-1257 [DOI: 10.1109/CVPR.2019.00134]

Zhang H Y, Wang T B, Li M Z, Zhao Z, Pu S L and Wu F. 2022. Comprehensive review of visual-language-oriented multimodal pre-training methods. Journal of Image and Graphics, 27(9): 2652-2682 (张浩宇, 王天保, 李孟择, 赵洲, 浦世亮, 吴飞. 2022. 视觉语言多模态预训练综述. 中国图象图形学报, 27(9): 2652-2682) [DOI: 10.11834/jig.220173]

Zhang S Y, Peng H W, Fu J L and Luo J B. 2020. Learning 2D temporal adjacent networks for moment localization with natural language//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 12870-12877 [DOI: 10.1609/aaai.v34i07.6984]

作者简介

曾润浩，男，副教授，主要研究方向为视频分析和多模态信息处理。E-mail: zengrh@smbu.edu.cn

陈奇，通信作者，男，博士后研究员，主要研究方向为计算机视觉、生成模型。E-mail: qi.chen04@adelaide.edu.au

李嘉梁，男，硕士研究生，主要研究方向为视频片段检索及视频检索。E-mail: lijialiang2022 @email.szu.edu.cn

卓奕深，男，硕士研究生，主要研究方向为视频片段检索及视频检索。E-mail: zhuoyishen2022@email.szu.edu.cn

段海涵，男，副教授，主要研究方向为区块链与 Web3、元宇宙。E-mail: duanhaihan@smbu.edu.cn

胡希平，男，教授，主要研究方向为情感智能计算。E-mail: huxp@smbu.edu.cn