# 基于多语言-视觉公共空间学习的多语言文本-视频跨模态检索模型

林俊安[1) ] 包翠竹[1)] 董建锋[1)] 杨 勋[2)] 王 勋[1)]

[1)](浙江工商大学计算机科学与技术学院 杭州 310018)

[2)](中国科学技术大学信息科学技术学院 合肥 230026)

**摘 要** 本文针对具有挑战性的多语言文本-视频跨模态检索问题进行研究. 传统文本-视频跨模态检索模型通常针对单一语言进行设计,比如英语,模型仅支持某一特定语言的文本查询. 如果有不同语言检索需求,则需另收集目标语言的训练数据并重新训练构建新的检索模型,这使得模型很难快速有效地适用于其他语言的检索任务. 近年来,针对多语言问题的研究逐渐深入,这为多语言跨模态检索的实现打下了良好的基石. 为了解决多语言跨模态检索问题,本文提出了一种简单有效的基于多语言-视觉公共空间学习的多语言文本-视频跨模态检索模型,将不同语言与视觉信息映射到同一公共空间. 该空间以视频向量为锚点,分别与不同的语言向量进行对齐,以此实现多语言跨模态的学习,由此建立了统一的多语言学习框架,使用一个模型满足了多语言的检索需求并探究了不平行语料库、平行语料库、伪平行语料库三种训练场景下的模型性能. 同时,在多语言建模中有效地利用了不同语言之间的互通性和互补性,弥补了单语言文本特征表达的不足;并在文本端与视频端引入了基于对比学习的抗噪音鲁棒性学习方法,进一步提升了不同模态特征的表示能力. 在VATEX、MSR-VTT多语言数据集上实验的数据证明,本文模型不仅能够简单快速地适用于多种语言检索任务,模型性能也较为突出,在较为常见的伪平行场景下和最先进的方法相比,中文VATEX和MSR-VTT在总召回率上分别提升了约5.97%和1.37%.

**关键词** 多语言;跨模态检索;跨模态特征表示;对比学习

**中图法分类号** TP391.3 **DOI号** 10.11897/SP.J.1016.2024.02195

# Multilingual Text-Video Cross-Modal Retrieval Model via Multilingual-Visual Common Space Learning

LIN Jun-An[1)] BAO Cui-Zhu[1)] DONG Jian-Feng[1)] YANG Xun[2)] WANG Xun[1)]

[1)](*Department of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou* 310018)

[2)](*Department of Information Science and Technology, University of Science and Technology of China, Hefei* 230026)

**Abstract** This paper focuses on the challenging multilingual cross-modal text-video retrieval. Traditional cross-modal text-video retrieval models are usually designed for a single language, such as English, and only support text queries in a specific language. If different language retrieval requirements are encountered, training data for the target language needs to be collected, and a new model needs to be built and retrained, which makes it difficult to apply the model to multilingual retrieval tasks quickly and effectively. In recent years, research on multilingual problems has gradually deepened, laying a solid foundation for the implementation of multilingual

收稿日期:2023-07-04;在线发布日期:2024-06-12. 本课题得到浙江省"尖兵""领雁"研发攻关计划项目(No. 2023C01212)、浙江省基础公益技术研究计划(No. LGF21F020010)、第八届中国科协青年人才托举工程项目(No. 2022QNRC001)资助.**林俊安**,学士,主要研究领域为跨模态检索、跨模态定位. E-mail:jalinux@163.com. **包翠竹**(共同第一作者),博士,讲师,中国计算机学会(CCF)会员,主要研究领域为计算机视觉、智能交通控制. **董建锋**(通信作者),博士,研究员,中国计算机学会(CCF)会员,主要研究领域为多媒体理解、计算机视觉. **杨 勋**,博士,研究员,中国计算机学会(CCF)会员,主要研究领域为跨媒体分析与推理、多媒体内容结构化理解. **王 勋**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为移动图形计算、计算机视觉.

cross-modal retrieval. In order to solve the problem of multilingual cross-modal retrieval, this paper proposes a simple and effective multilingual text-video cross-modal retrieval model via multilingual-visual common space learning, which maps different languages and visual feature to the same common space, this space uses video vectors as anchors and aligns them with different language vectors to achieve cross-modal learning in multi-languages. Thus, a unified multilingual learning framework was established. This method uses only one model solves the multilingual retrieval problem, and explores the performance of the model in the three training scenarios of non-parallel corpus, parallel corpus and pseudo-parallel corpus. At the same time, the interoperability and complementarity between different languages in multilingual modeling are effectively used to make up for the lack of monolingual text feature representation; and a robust learning method based on contrastive learning is introduced in the text and video ends, which further improves the representation ability of different modal features. The experimental results on the VATEX and MSR-VTT multilingual datasets demonstrate that the proposed model can not only be applied to multilingual retrieval tasks simply and quickly, but also the model performance is outstanding, compared with state-of-the-art methods in pseudo parallel scenes, Chinese VATEX and MSR-VTT have improved sum recall by approximately 5.97% and 1.37%, respectively.

**Keywords** multilingual; cross-modal retrieval; cross-modal feature representation; contrastive learning

## 1 引 言

近年来,随着互联网技术与智能终端设备的快速发展,内容丰富、文化各异的短视频等多媒体数据的数量急剧增长. 海量的多媒体数据极大地考验着视频平台的检索效能. 传统的基于单一模态的检索方式已逐渐无法满足多样化的用户需求,研究者开始探索如何实现多种模态间的信息检索,因此跨模态检索逐渐受到工业界和学术界的广泛关注[1]. 跨模态检索(Cross-Modal Retrieval)在检索时查询样例和候选对象可以为不同模态的数据,不同模态间基于语义相似度可以进行相互检索. 例如文本-视频跨模态检索[2],用户可以基于视频检索到描述视频的文本或基于文本检索到相应内容的视频[3]. 然而,随着全球化背景下的文化交流快速增长,在面对不同国家不同语言的检索需求时,目前基于单一语言的跨模态检索方法很难快速满足多种语言的用户诉求. 越来越多的研究机构和平台开始关注如何解决多语言检索的问题.

目前主流的文本-视频跨模态检索方法是利用多模态信息的语义一致性,将不同模态信息的特征嵌入到公共空间,并建立一个视觉语言通用的模型. 但是目前大部分的模型都是单语言跨模态检索模型,这些模型往往是针对单一语言的跨模态检索任务. 受到大量人工标记数据集的影响,模型的功能往往都局限于英语检索. 在处理其他语言检索任务时,如中文,一种可行方法是使用机器翻译将不同的语言译成训练时使用的语言,如将中文翻译成英文进行检索. 但由于翻译噪音的影响,其性能表现往往不佳. 另一种可行方法是收集另一种语言的人工标注数据集并重新进行训练. 这不仅受到其他语言标记数据不足训练不充分的影响,还增加了额外的(训练)时间成本. 当下,针对多语言的跨模态工作关注度较低. 相较于单语言跨模态模型,多语言跨模态模型解决了语言泛化问题,顺应了互联网多元化的趋势. 因此,如何建立一个统一的多语言模型快速地满足不同语言的跨模态检索需求已经是一个逐渐重要的研究方向.

单语言跨模态视频检索与多语言跨模态视频检索的对比如图 1 所示. 单语言模型训练数据和检索需求均围绕单一语言,由单一语言和视频对构建训练数据,最终模型能对特定语言进行检索. 如需实现多种语言的跨模态检索,需要训练多个针对某一语言的模型. 而多语言跨模态检索意味着使用一种模型实现多种语言的检索,多语言跨模态检索模型往往将不同语言数据以及对应的视频数据构成统一训练数据对,最终单个模型就能满足多种语言的检

索任务.

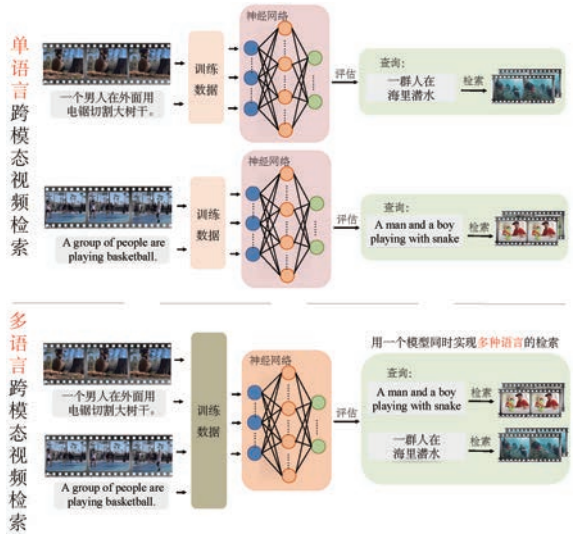



图 1　单语言跨模态视频检索与多语言跨模态视频检索对比

根据训练数据的不同,本文将多语言模型的训练语料库归纳为三种:平行语料库:每个视频同时有多种不同语言的人工标注的文本描述,如图 2(a)所示;伪平行语料库:每个视频既有人工标注文本描述,亦存在由人工标注文本描述通过机器翻译得到的文本描述,如图 2(b)所示;不平行语料库:每个视频仅包含一种语言的文本标注信息,如图 2(c)所示. 不平行语料库可以通过机器翻译转换成伪平行语料库,但受到不具备翻译环境的限制或翻译噪音过大的影响.

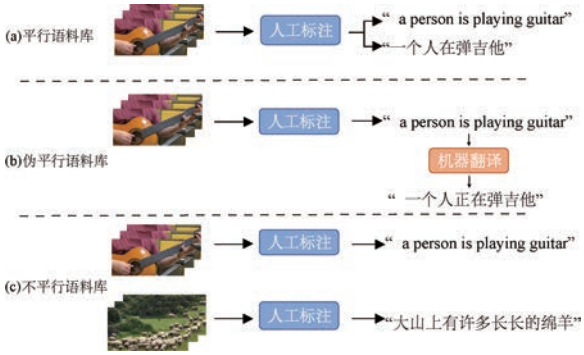



图 2　多语言训练数据构建的三种场景

近年来关于多语言模型的研究受到越来越多的关注,以 mBERT[4] 和 XLM-R[5] 为首的多语言预训练模型的出现进一步促进了这一研究领域的发展. 近期也有部分工作开始研究多语言环境下的跨模态检索,如 Portaz 等人[6] 和 Aggarwal 等人[7] 分别依赖预先训练好的多语言词嵌入和预先训练好的多语言句子编码器进行检索. 而 Huang 等人提出的多语言多模态预训练模型[8](Multilingual Multimodal Pre-training, MMP)则是使用了 XLM-R 进行多语言预训练并建立了基于 Transformer 的多语言视觉文本检索模型. Wang 等人提出的抗噪声跨语言跨模态检索模型[9](Noise-Robust Cross-lingual Cross-modal Retrieval, NRCCR)针对多语言学习中某些语言缺少训练样本的问题,建立了双重翻译机制并使用蒸馏方法解决机器翻译带来的噪音影响. 虽然目前主流的多语言跨模态检索模型能够应用于多语言的检索任务,但由于不同语言在训练过程中会因特征对齐等因素相互影响,众多的多语言跨模态模型不支持不平行语料库这种最为常见的训练场景. 例如, MMP 模型仅支持平行语料库和伪平行语料库两种训练场景,而 NRCCR 则更侧重于解决伪平行语料库的多语言跨模态检索任务. 为此,本文提出了一种基于多语言-视觉公共空间学习的多语言文本-视频跨模态检索模型. 不同于以往计算复杂的数据增强,视频端采用随机抽取视频帧的方法形成缺失一定信息熵的噪声视频;文本端采用随机遮盖方法形成缺失一定信息熵的噪声文本. 提出了一个基于对比学习方法的语言-遮盖语言,视频-抽帧视频的鲁棒学习框架,并进一步建立多语言文本-视觉公共空间,以视频特征为锚点对齐多种语言特征,建立了一种简单有效的多语言的跨模态检索模型,主要贡献如下:

(1)提出了一种简单有效的基于多语言-视觉公共空间学习的多语言文本-视频跨模态检索模型. 建立了一个统一的语言-遮盖语言,视频-抽帧视频的简单的对比学习框架,并通过多语言文本-视觉公共空间学习实现多语言的跨模态检索.

(2)探究了平行多语言语料库、由机器翻译建立的伪平行多语言语料库、不平行多语言语料库三种多语言训练场景下模型的性能,我们的模型最终能支持以上三种语料库的输入.

(3)提出了一种多语言-视觉公共空间,与传统的基于潜在空间的方法不同,该空间以视频向量为锚点,分别与不同的语言向量进行对齐,以此实现多语言跨模态的学习.

## 2　相关工作

### 2.1　文本-视频跨模态检索

跨模态文本-视频检索近年来受到学术界和工业界的关注. 跨模态检索的难点问题在于不同模态之间存在着语义鸿沟. 目前,基于嵌入的方法依旧

是实现文本-视频跨模态检索的主流方案[10-17]，该方法往往是先提取视频和文本的特征表示，然后学习一个文本视频联合特征嵌入空间，在该空间中可以直接测量跨模态相似度. 对于文本特征表示方法，早期有词袋(Bag Of the Word，BoW)、word2vec 等方式. 例如，词到视觉向量(Word2VisualVec，W2VV++[11])采用了 BoW、word2vec、GRU 提取多种特征并进行特征拼接. 而目前较为流行的方式是通过 BERT 等预训练模型进行特征提取. 例如，Gabeur 等人[18]提出了一个用于视频编码的多模态 Transformer，使用 BERT 来获得文本的嵌入. 文本到视频 VLAD(Text to Video VLAD，T2VLAD[19])则采用了端到端的微调 BERT 模型作为文本特征提取器. 对于视频特征的提取，典型的方法是将视频的多个帧级别的特征聚集成视频级别特征. 例如 dong 等人[20]对文本端和视频端分别进行三种特征提取，对于视频端采用了 CNN 提取帧级特征. 而 Liu 等人[21]则采用平均池和 NetVLAD 方法将不同的帧级特征作为视频特征. Wang 等人提出了 TKVTR 模型[22]，使用了图神经网络(Graph Neural Networks，GNNs)，利用结构相似性提取知识并进行知识迁移来辅助检索. 大模型预训练在最近越来越流行，例如 BLIP[23]，通过联合训练视觉和语言模型来提升多模态任务的性能，其衍生的模型 BLIP2[24]和 instructBLIP[25]都展现了较为先进的性能. 通过构建或引入大模型，往往能产生更全面的视频或文本特征. 但上述大部分研究局限于单语言，模型仅具备单语言能力，仅支持针对某种语言的检索.

### 2.2 多语言文本-视频跨模态检索

为了使跨模态检索满足不同语言的检索需求，解决不同语言的泛化性问题，多语言文本-视频跨模态检索工作逐渐受到关注. 例如，Portaz 等人[6]采用了非语境化的多语言词嵌入，这些方法是通过对齐不同语言的词来预先训练的. 随后基于 Transformer 的多语言预训练模型 mBERT[4]的出现使得 NLP 领域在多语言方面有了新的突破. 随后，多语言预训练模型 XML-R[5]被提出，其性能在众多的数据集上优于 mBERT. 以 mBERT 和 XML-R 为首的多语言预训练模型的出现，使得越来越多的多语言研究开始倾向预训练方面. 例如，Huang 等人[8]设计了一个基于 Transformer 的预训练模型 MMP，用于学习语境化的多语言多模态表示，并使用 MultiHowTo100M[14]进行预训练. 但是，上述关于多语言跨模态方面的工作在对多语言问题的挖掘和体现仍旧不足. 对于 MMP 而言，在缺少另一种文本标注语言的情况下，模型性能表现很差. 为了解决这个问题，Wang 等人[9]提出了 NRCCR 模型，NRCCR 注重缺少目标语言的多语言场景，它采用机器翻译建立伪平行数据对，建立了双重翻译模型，并使用蒸馏的方法解决机器翻译带来的噪音影响，其性能相较于其他模型有较大的提升. 而 Madasu 等人则提出了视频检索中的多语言知识转移模型(Multilingual Knowledge Transfer for Video Retrieval，MKTVR[26])，基于 CLIP 进行多语言知识适应与迁移. 目前，基于大模型的方法开始流行，现有的工作也逐渐向更庞大的参数和训练量发展. 例如 Jian 等人提出了跨语言多模态多任务检索模型(MUltimodal，MUltitask Retrieval Across Languages，MURAL[27])，结合图像-文本匹配和文本-文本匹配两种任务实现跨语言任务并有上亿的训练对. Zhang 等人提出的多语言习得框架(Multi-Lingual Acquisition，MLA[28])则是学习一个语言习得编码器，以一种通用的方法将单语言跨模态模型扩展到其他语言上. Wan 等人提出了统一跨语言医学视觉语言预训练(Med-UniC[29])来结合不同语言的医疗多模态数据. 但是先前的模型诸如 NRCCR 与 MKTVR 模型是专门为伪平行多语言场景设计的，应用场景较为狭隘. 针对上述问题，本文致力于通过对比学习的方法构建文本端与视频端的训练模式，并涵盖多种应用场景，构建一个多语言-视觉公共空间，设计一个多语言文本-视频跨模态检索模型.

## 3 方 法

本节将详细介绍所提出的基于多语言-视觉公共空间学习的多语言文本-视频跨模态检索模型(MultiLingual-Visual Common Space Learning model，MLVCSL). 在多语言环境下，给定两种语言-视频数据对，$P_X=(s^x,v)$ 与 $P_Y=(s^y,v)$. 其中 $s^x$ 是用语言 $X$ 描述视频 $v$ 的文本集合，即 $s^x=\{s_1^x,s_2^x,\cdots,s_n^x\}$，$n$ 表示描述文本数量；类似的，$s^y$ 代表用另一种语言 $Y$ 描述视频 $v$ 的文本集合，$s^y$ 可以通过人工标注获得(平行语料)，亦可通过 $s^x$ 进行机器翻译获得(伪平行语料). $P_X$ 与 $P_Y$，可以共同建立一个多语言文本-视频数据集合 $P=\{(s^x,v),(s^y,v)\}$. 当视频样本在不平行语料库的场景下训

练时,也可以通过不同语言-视频平行数据集合 $P=\{(s^x,v^x),(s^y,v^y)\}$ 进行训练.其中,$v^x$ 代表 $X$ 语言描述文本对应的视频,$v^y$ 代表 $Y$ 语言描述文本对应的视频,二者没有交集.下面将详细介绍本文方法.

### 3.1　基于对比学习的特征编码

为了提升视频特征的表示能力,引入了图像领域的对比学习方法.基于对比学习的图像特征编码可以通过裁剪、调整大小、重新着色等方式来增加同类图像提升特征学习的鲁棒性.但由于视频包含多帧图像且有较多冗余信息,这种方式不仅极大增加模型的计算量还容易引入噪音.为了在减少计算量的同时提升视频的特征表示能力,提出了一种语言-遮盖语言,视频-抽帧视频的对比学习框架,如图 3 所示.视频端通过抽取视频帧的方式与原视频进行对比学习;文本端也采用类似的遮盖操作来降低文本噪音的影响,框线代表这些结构共享参数.请注意,图 3 描述的是模型在(伪)平行多语言语料库下训练.对于不平行多语言语料库,模型主要架构不变,只是在文本编码部分少了一个语言分支的输入.

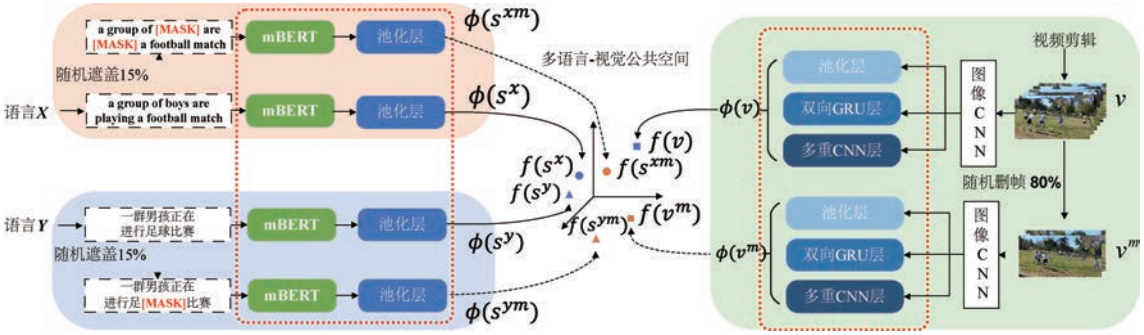



图 3　本文提出的基于多语言-视觉公共空间学习的多语言文本-视频跨模态检索模型框架

#### 3.1.1　基于对比学习的多语言文本特征编码

为了减少标注不准确或翻译质量不佳所引入的文本噪音的影响,同时也为了更好地挖掘不同语言之间的互通性,在文本端对多种语言同时建模,如图 3 中的语言 $X$ 与语言 $Y$,并引入了基于对比学习的多语言文本特征编码.由于多种语言使用统一的模型进行建模时对模型的语义理解能力有更高的要求,因此借鉴了 BERT 模型的遮盖方法,使用遮盖的方式来模拟噪音.通过对比学习的方法在随机遮盖的条件下重构语义信息,进而增加模型对语句的整体语义理解能力.文本端的文本编码器是一种语言无关(language-agnostic)文本编码器,这种编码器事先无需知道输入的是何种语言,提取的是语言无关的语义表征.具体地,文本编码器由多语言 Transformer 预训练模型 mBERT 和平均池组成,所有分支共享 mBERT 预训练模型的参数.mBERT 预先冻结了 5 层及以下的参数,并进一步通过随机遮盖 15% 的词形成了带有噪音的句子 $s^{xm}$ 与 $s^{ym}$.因而文本端细分共有 4 个分支,可以得到 4 个编码结果,分别为语言 $X$ 的原始文本编码结果 $\phi(s^x)$ 和遮盖文本编码结果 $\phi(s^{xm})$,以及语言 $Y$ 的原始文本编码结果 $\phi(s^y)$ 和遮盖文本编码结果 $\phi(s^{ym})$,模型通过遮盖进行对比学习提升文本特征的鲁棒性,使词之间的语义信息不会过分依赖.

文本与遮盖文本之间的损失可以视为一种抗噪音损失,应确保原始文本与在随机遮盖后的噪音文本相接近,具体使用了 InfoNCE 损失[30],使噪音样本和对应的原始样本对视为正对,其余均视为负对.分别以文本和遮盖文本为锚点计算损失并求平均,通过该损失训练,模型的鲁棒性和文本向量特征表示得到进一步增强,在附录的损失消融实验和抗噪训练有效性实验中进一步证明了该损失的有效性.对于给定的原始文本与噪音文本对 $(s,s^m)$,其计算方式如下:

$$L_{infoNCE}(s,s^m)=\sum_{s\in s^x,s^y}\left(-\sum_{i=1}^{N}\log\frac{\exp(sim(f(s_i),f(s_i^m))/\tau)}{\sum_{k=1}^{N}\exp(sim(f(s_i),f(s_k^m))/\tau)}-\sum_{i=1}^{N}\log\frac{\exp(sim(f(s_i),f(s_i^m))/\tau)}{\sum_{k=1}^{N}\exp(sim(f(s_k),f(s_i^m))/\tau)}\right)/2 \tag{1}$$

其中,$\tau$ 为温度系数,$N$ 为一个 mini-batch 的数量,$sim(\cdot)$ 为余弦相似度.$s\in s^x,s^y,s$ 为包含不同语言的多语言句子集合,即,对于每一种语言的文本,分别进行损失计算并求和得到最终的原始文本和噪音文本的损失.

3.1.2 基于对比学习的视频特征编码

视频端以均匀间隔抽取视频帧与原视频进行对比学习，如图 3 所示. $v$ 为原始视频帧，通过抽帧后删除 80%的视频帧形成 $v^m$. 视频编码借鉴 Dual Encoding 的编码方法[20]，每帧使用预先训练好的 ImageNet CNN 提取深层特征，并将此特征经过平均池后通过双向 GRU 进一步提取特征，再通过 biGRU-CNN 增强编码，最后将三种特征拼接，具体的特征编码也可以采用其他方法. 原始视频 $v$ 的编码结果 $\phi(v)$，$v^m$ 的编码结果 $\phi(v^m)$，通过对比学习提升视频特征的鲁棒性，使视频帧之间的语义信息不会过分依赖. 与文本端类似，原始视频与抽帧视频对 $(v,v^m)$ 之间亦采用 InfoNCE 损失.

### 3.2 多语言文本-视觉公共空间学习

为了实现多语言跨模态学习，本文提出了多语言文本-视觉公共空间. 传统的跨模态检索在解决多种语言的检索问题时往往采用单语言模型，建立潜在空间对齐视频特征和一种语言的文本特征. 当有其他语言的检索任务时，需要重新训练对齐视频特征和另一种语言的文本特征. 与传统方法不同，所建立的多语言-视觉公共空间是以视频特征为锚点，同时将多种不同语言文本特征与视频特征进行对齐所构建的，不同语言的文本特征通过视觉特征间接对齐，后续性能实验证明了多语言-视觉公共空间的有效性. 如图 3 所示，在获得文本编码和视频编码后，通过特征投影将不同语言文本和视频投影到一个多语言文本-视觉公共嵌入空间进行学习. 通过全连接(FC)层来实现投影，经过 FC 层后，视频特征通道和文本特征通道由不同维度 $d_v$ 和 $d_s$ 被映射到同一维度 $d$，并在 FC 层之后使用了一个批处理规范化(BN)层，最终得到多语言文本-视觉公共空间中的视频特征向量 $f(v)$ 和 $f(v^m)$ 以及文本特征向量 $f(x)$、$f(x^m)$、$f(y)$、$f(y^m)$，可以通过标准的相似度测量方法直接衡量视频和文本的相似性.

接下来，需要以视频特征向量 $f(v)$ 为锚点，分别和文本特征向量 $f(x)$、$f(y)$ 进行对齐训练，具体实现如下：

为了使相关的文本-视频对在公共空间中接近，无关的文本-视频对在空间中远离，使用了改进的三元排序损失[10]，根据 mini-batch 中最困难的样本，即负样本中与锚点相似度最高的句子和视频，对模型进行惩罚. 与视频或文本语义一致的样本视为正样本，其余均视为负样本. 对给定的 mini-batch 文本-视频对 $(s,v)$ 的损失计算方法如下：

$$L_{rank}(s,v)=\sum_{s\in s^x,s^y}(max(0,m+sim(f(v),f(s^-))-sim(f(v),f(s^+)))+max(0,m+sim(f(v^-),f(s))-sim(f(v^+),f(s)))) \tag{2}$$

其中，$m$ 是边际常数，而 $s^-$ 和 $v^-$ 分别表示 $v$ 和 $s$ 的最困难样本，而 $s^+$ 和 $v^+$ 分别表示 $v$ 和 $s$ 的正样本. $s\in s^x,s^y,s$ 为包含不同语言的多语言句子集合，例如 $s^x$ 代表原始英语文本而 $s^y$ 代表原始中文文本. 带噪音的文本和视频不参与模态间的损失计算，即对于每一种语言的文本，分别进行损失计算并求和得到最终的原始文本和原始视频的损失.

通过最小化三种损失类型组合在公共空间中学习，对于一个给定的多语言视频间的数据对 $P=(x,y,v)$，提出的最终损失如下：

$$min_\theta L_{rank}(s,v)+L_{infoNCE}(v,v^m)+L_{infoNCE}(s,s^m) \tag{3}$$

其中 $\theta$ 表示整个模型中的所有可训练参数，如果模型在不平行多语言场景下训练，则去掉另一种语言的三元排序损失和文本端的 InfoNCE 损失.

相似性. 最终通过原始的视频与文本编码结果度量文本和视频的相似性. 为了进一步挖掘多语言文本间的互通性，在求解相似度的时候可以使用机器翻译将测试集语言数据集翻译成另一种语言构成翻译数据集，辅助相似度求解. 通过相似度加权和求得最终的相似度. 视频和文本相似度最终为

$$sim(s,v)=\gamma sim(f(s),f(v))+(1-\gamma)sim(f(s^t),f(v)) \tag{4}$$

其中 $\gamma$ 是一个超参数，用于平衡原始语言文本和翻译语言文本的重要性，范围在[0,1]之内. $s^t$ 为机器翻译得到的文本描述.

## 4 实 验

### 4.1 实验设置

(1)数据集. 在通用的文本-视频跨模态检索数据集 VATEX 和 MSR-VTT 上进行实验，包含平行语料库、伪平行语料库和不平行语料库三种场景.

①VATEX[31]：VATEX 为双语文本视频数据集，包含 41250 个视频和 825000 个描述视频的句子. 每个视频片段对应 10 个英语句子和 10 个中文句子. 与文献[12]使用的数据相似，使用 25991 个视频片段训练，1500 个视频片段验证，1500 个视频片段测试. 验证和测试集是通过 3000 个视频片段的验证集随机分割获得.

②MSR-VTT[32]:MSR-VTT 为单语文本视频数据集,实验分为中文查询测试和英文查询测试.中文查询测试与文献[9]相同,使用 9000 和 1000 的视频片段用于训练和测试,验证集和测试集相同.训练集中的中文数据是由数据集原始英文标注通过谷歌翻译获得,而测试集中的中文句子数据则是由人工标注获得.英文查询测试与文献[33]相同,使用 7010 和 1000 个视频片段用于训练和测试.在除训练和测试外的原始数据集中随机选取了 1000 个视频片段用于构建验证集.验证集中的中文句子数据也是由数据集原始英文标注通过谷歌翻译而来.

(2)评价指标.使用基于排序的评价指标 $R@K$($K=1,5,10$),排序中位数(Med $r$)和平均数(mAP)来评估性能.$R@K$ 是前 $K$ 位中正确检索的查询的数.较高的 $R@K$、mAP 和较低的 Med $r$ 为更优性能.为了方便总体比较,给出了召回率的总和(Sum$R$).

(3)实现细节.文本端预处理时将所有单词转换为小写.视频端 VATEX 上采用了数据集开发人员提供的 1024-d I3D[33] 作为预训练视频特征;MSR-VTT 上使用 ResNeXt-ResNet[20] 作为预训练视频特征.根据文献[10]的参数设置将改进的三元组排序损失的参数 $m$ 设置为 0.2.训练的 BatchSize 大小为 64,epoch 总数设定为 50,性能基本在 20 epoch 内达到顶峰,使用的显卡为 Tesla V100-32G,使用 Adam 的随机梯度下降,初始学习率为 0.0001,并采用文献[13]的学习率调整方法,即,每个 epoch 后将学习率乘 0.99 的方式进行衰减,如果模型性能在 3 个 epoch 内没有上升则将学习率乘以 0.5.同时如果模型在 10 个 epoch 内性能没有提升,将会触发 early stopping,结束训练.

### 4.2　性能比较与分析

#### 4.2.1　多语言平行或伪平行语料场景的性能表现

实验中的多语言主要设定为中文和英文两种语言.实验中将本文模型与 MMP[8] 和 NRCCR[9] 进行比较,MMP 专注于多语言视频文本检索工作,而 NRCCR 专注于伪平行多语言视频文本检索工作.同时也对比了 NRCCR 文中所提到的单语视频文本检索方法,包括 W2VV[34]、VSE++[10]、W2VV++[11]、MultiHowTo100M[14]、CE[21]、Dual Encoding[20]、HGR[12]、GPO[15] 和 RIVRL[17],SEA[16].在英语方面,除上述一些模型外,本文还比较了 TKVTR[22]、TCE[35]、MEE[36] 的方法.在模型性能表中,MLVCSL * 表示使用了平行的中英两种人工语言标注训练模型,MLVCSL 表示使用了人工标注和机器翻译得到的伪平行语料库训练最终的模型.

(1)VATEX 实验.表 1 与表 2 对比了同样使用 1024-d I3D 作为预训练视频特征的方法.为了便于与当下的模型进行比较,这里 MLVCSL 的中文训练数据集由原始英文训练数据集通过谷歌翻译获得.在 VATEX 中文测试集上的实验如表 1 所示.实验对比了以 NRCCR 为首的伪平行语料库训练的模型,实验结果表明与 NRCCR 相比,MLVCSL 方法 SumR 性能指标提升了 21.8.而在使用原始人工标记双语资源的情况下,MLVCSL * 相较于普通的 MLVCSL 性能指标 SumR 提升了 20.1.这说明高质量的多语言训练数据有助于提升多语言跨模态检索模型的性能.对于其他的单语或多语模型,本文模型都具有优势.表 2 展示了在 VATEX 英文测试集上的实验结果,MLVCSL 模型依旧能展现出较好的性能,证明了本文方法在多语言视频文本检索中的有效性.

**表 1　在 VATEX 上的中文实验表现**

| 模型 | Text-to-Video Retrieval | | | | | Video-to-Text Retrieval | | | | | Sum$R$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R@1$ | $R@5$ | $R@10$ | Med $r$ | mAP | $R@1$ | $R@5$ | $R@10$ | Med $r$ | mAP | |
| MMP w/o pre-train[8] | 23.9 | 55.1 | 67.8 | — | — | — | — | — | — | — | — |
| MMP[8] | 29.7 | 63.2 | 75.5 | — | — | — | — | — | — | — | — |
| W2VV[34] | 5.46 | 12.3 | 15.4 | 298 | 9.20 | — | — | — | — | — | — |
| VSE++[10] | 21.1 | 48.1 | 59.6 | 6.0 | 33.72 | 34.9 | 67.2 | 77.5 | 3.0 | 21.76 | 307.7 |
| W2VV++[11] | 20.5 | 48.3 | 59.5 | 6.0 | 33.40 | — | — | — | — | — | — |
| MultiHowTo100M [14] | 13.7 | 37.1 | 50.4 | 10.0 | 25.32 | 23.2 | 53.1 | 67.3 | 5.0 | 14.62 | 244.8 |
| CE[21] | 20.2 | 48.5 | 60.8 | 6.0 | 33.48 | 31 | 63.5 | 76.7 | 3.0 | 20.2 | 300.7 |
| Dual Encoding[20] | 23.1 | 52.1 | 62.6 | 5.0 | 36.32 | 35.6 | 67.9 | 79.3 | 3.0 | 23.98 | 320.5 |
| HGR[12] | 14.3 | 37.0 | 47.5 | 12.0 | 25.10 | 27.8 | 61.0 | 72.9 | 4.0 | 15.5 | 261.0 |
| GPO[15] | 19.5 | 46.9 | 57.8 | 6.0 | 32.20 | 33.8 | 65.9 | 77.3 | 3.0 | 20.21 | 301.3 |
| RIVRL[17] | 26.8 | 56.5 | 67.3 | 4.0 | 40.35 | 38.3 | 71.1 | 80.9 | 2.0 | 26.62 | 340.9 |
| NRCCR[9] | 30.4 | 65.0 | 75.1 | 3.0 | 45.64 | 40.6 | 72.7 | 80.9 | 2.0 | 32.40 | 364.7 |
| MLVCSL | 33.1 | 67.1 | 77.1 | 3.0 | 48.18 | 46.7 | 76.6 | 85.9 | 2.0 | 35.40 | 386.5 |
| MLVCSL* | **36.3** | **71.3** | **81.2** | **2.0** | **51.69** | **49.9** | **79.8** | **88.2** | **2.0** | **39.26** | **406.6** |

**表 2 在 VATEX 上的英文实验表现**

| 模型 | Text-to-Video Retrieval | | | Video-to-Text Retrieval | | | SumR |
|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| W2VV[34] | 14.6 | 36.3 | 46.1 | 39.6 | 69.5 | 79.4 | 285.5 |
| VSE++[10] | 31.3 | 65.8 | 76.4 | 42.9 | 73.9 | 83.6 | 373.9 |
| CE[21] | 31.1 | 68.7 | 80.2 | 41.3 | 71.0 | 82.3 | 374.6 |
| W2VV++[11] | 32.0 | 68.2 | 78.8 | 41.8 | 75.1 | 84.3 | 380.2 |
| HGR[12] | 35.1 | 73.5 | 83.5 | — | — | — | — |
| Dual Encoding[20] | 36.8 | 73.6 | 83.7 | 46.8 | 75.7 | 85.1 | 401.7 |
| TKVTR[22] | 37.8 | 74.1 | 83.8 | 49.3 | 78.5 | 86.5 | 410.0 |
| RIVRL[17] | **39.3** | **76.0** | **85.1** | — | — | — | — |
| MLVCSL | 38.3 | 74.0 | 82.9 | 50.2 | **78.5** | **86.7** | 410.7 |
| MLVCSL* | 38.2 | 74.8 | 83.9 | **50.3** | 78.0 | 86.3 | **411.4** |

(2)MSR-VTT 实验. 由于 MSR-VTT 为英语单语数据集,因此仅展示了伪平行语料库训练的 MLVCSL 性能. 表 3 中给出的中文测试实验结果表明,较之于先进的多语言模型 NRCCR,MLVCSL 仍能在性能上和其平分秋色,并超越了其他模型. 由于 MMP 仅在 VATEX 上进行实验且未开源,因此未与其在 MSR-VTT 上进行中文测试的比较,仅在表 4 比较了英文测试的实验结果. 所有开源相关的方法都使用相同的 ResNeXt-ResNet 预训练. 如表 4 所示, MLVCSL 的英文测试性能也取得最优.

**表 3 在 MSR-VTT 上的中文实验表现**

| 模型 | Text-to-Video Retrieval | | | | | Video-to-Text Retrieval | | | | | SumR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | Med r | mAP | R@1 | R@5 | R@10 | Med r | mAP | |
| W2VV[34] | 10.4 | 18.7 | 22.9 | 153 | 1.50 | — | — | — | — | — | — |
| VSE++[10] | 17.1 | 43.9 | 54.0 | 9.0 | 29.57 | 17.5 | 43.3 | 55.0 | 7.0 | 29.86 | 230.8 |
| W2VV++[11] | 23.8 | 50.3 | 61.0 | 5.0 | 36.40 | — | — | — | — | — | — |
| MultiHowTo100M [14] | 13.2 | 38.3 | 52.9 | 9.0 | 25.95 | 13.5 | 36.2 | 48.9 | 11.0 | 25.31 | 203.0 |
| Dual Encoding[20] | 19.5 | 45.9 | 56.6 | 7.0 | 31.75 | 20.7 | 44.3 | 57.1 | 7.0 | 32.25 | 244.1 |
| CE[21] | 21.0 | 49.7 | 63.4 | 7.0 | — | 19.6 | 49.0 | 62.7 | 6.0 | — | 265.4 |
| SEA[16] | 21.0 | 48.3 | 61.1 | 6.0 | 33.80 | 11.8 | 31.2 | 42.4 | 17.0 | 21.90 | 215.8 |
| HGR[12] | 14.4 | 41.4 | 53.3 | 9.0 | 27.18 | 16.2 | 40.9 | 53.3 | 9.0 | 27.79 | 219.5 |
| GPO[15] | 18.2 | 42.3 | 53.2 | 8.0 | 29.86 | 16.8 | 43.1 | 52.9 | 9.0 | 29.15 | 226.5 |
| RIVRL[17] | 24.3 | 51.4 | 63.0 | 5.0 | 37.03 | 21.8 | 49.7 | 63.5 | 6.0 | 35.46 | 273.7 |
| NRCCR[9] | **29.6** | **55.8** | 67.4 | **4.0** | **41.93** | **31.3** | 56.0 | 67.2 | **4.0** | 43.00 | 307.3 |
| MLVCSL | 28.9 | 55.5 | **69.2** | **4.0** | 41.54 | 30.4 | **57.8** | **69.7** | **4.0** | **43.07** | **311.5** |

**表 4 在 MSR-VTT 上的英文实验表现**

| 模型 | Text-to-Video Retrieval | | | | | Video-to-Text Retrieval | | | | | SumR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | Med r | mAP | R@1 | R@5 | R@10 | Med r | mAP | |
| W2VV[34] | 1.9 | 9.9 | 15.2 | 79.0 | 6.8 | 17.3 | 39.3 | 50.2 | 10.0 | 27.8 | 133.8 |
| VSE++[10] | 16.0 | 38.5 | 50.9 | 10.0 | 27.4 | 16.2 | 39.3 | 51.2 | 10.0 | 27.4 | 212.1 |
| MEE[36] | 14.6 | 38.4 | 52.4 | 9.0 | 26.1 | 15.2 | 40.9 | 53.8 | 9.0 | 27.9 | 215.3 |
| W2VV++[11] | 19.0 | 45.0 | 58.7 | 7.0 | 31.8 | 16.9 | 42.7 | 54.6 | 8.0 | 29.0 | 236.9 |
| CE[21] | 17.2 | 46.2 | 58.5 | 7.0 | 30.3 | 15.8 | 44.9 | 59.2 | 7.0 | 30.4 | 241.8 |
| TCE[35] | 17.8 | 46.0 | 58.3 | 7.0 | 31.1 | 18.9 | 43.5 | 58.8 | 7.0 | 31.4 | 243.3 |
| HGR[12] | 21.7 | 47.4 | 61.1 | 6.0 | 34.0 | 20.4 | 47.9 | 60.6 | 6.0 | 33.4 | 259.1 |
| Dual Encoding[20] | 21.1 | 48.7 | 60.2 | 6.0 | 33.6 | 21.7 | 49.4 | 61.6 | 6.0 | 34.7 | 262.7 |
| MLVCSL | **26.9** | **55.3** | **67.7** | **4.0** | **40.20** | **29.8** | **55.7** | **67.7** | **5.0** | **42.16** | **303.1** |

4.2.2 多语言不平行语料场景的性能表现

为了进行多语言不平行语料训练场景的实验对比,使用 VATEX 构造了不平行语料库. 将 VATEX 的训练集与验证集的视频和描述句子等分成两部分,前一部分用仅使用英文描述,后一部分仅使用中文描述,测试集不变. 在实验中,将不使用对比结构且使用单语训练的模型作为英文基线(EN baseline)与中文基线(ZH baseline);将不使用对比结构且在语言方面使用不平行多语言语料库训练的模型作为多语言基线(Mutilingual baseline);而

本文提出的方法通过对比学习方法学习多语言公共空间,使用与多语言基线相同的不平行多语言语料库进行训练. 表 5 实验结果充分证明了本文模型在不平行语料训练场景下的有效性. 也就是说,对于互联网上各式各样不同语言的视频,都能作为本文模型的训练数据进而服务于各种语言的检索任务,这也充分证明了模型在实际应用场景中的意义.

**表 5　在 VATEX 上不平行多语言情况下的实验表现**

| 模型 | 双语训练 | 对比结构 | Text-to-Video Retrieval | | | Video-to-Text Retrieval | | | SumR |
|---|---|---|---|---|---|---|---|---|---|
| | | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| EN baseline | | | 25.2 | 54.6 | 65.4 | 30.2 | 58.9 | 69.4 | 303.6 |
| Mutilingual baseline | ✔ | | 30.3 | 63.0 | 73.7 | 40.5 | 68.9 | 79.3 | 355.7 |
| MLVCSL | ✔ | ✔ | **32.8** | **66.7** | **77.4** | **44.5** | **74.4** | **84.3** | **380.1** |
| ZH baseline | | | 24.0 | 54.3 | 65.9 | 33.1 | 60.5 | 70.5 | 308.4 |
| Mutilingual baseline | ✔ | | 28.0 | 60.8 | 71.9 | 39.5 | 68.6 | 79.0 | 347.8 |
| MLVCSL | ✔ | ✔ | **31.6** | **64.7** | **76.0** | **44.9** | **75.5** | **84.5** | **377.1** |

4.2.3　基于大模型视觉编码器的实验

大模型当前在视频特征表示方面展现了强大的能力. 为了进一步验证我们的模型,我们引入了当下较为热门的大模型包括 CLIP、BLIP、BLIP2 的视频预训练编码器替代原先的基于 ResNet 或 CNN 的视频预训练编码器. 我们在 MSRVTT 上进行了我们的实验,结果如表 6 所示,模型更换视频预训练编码器后在性能上展现了强大的进步,这也证明所提出的模型能有效地接入大模型.

**表 6　在 MSRVTT 上的大模型视觉编码器的实验表现**

| 模型 | 测试语言 | T2V | | | | V2T | | | | SumR |
|---|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | mAP | R@1 | R@5 | R@10 | mAP | |
| MLVCSL+ResNet | EN | 26.9 | 55.3 | 67.7 | 40.20 | 29.8 | 55.7 | 67.7 | 42.16 | 303.1 |
| MLVCSL+CLIP | EN | 37.1 | 68.4 | 78.3 | 50.98 | 38.2 | 69.1 | 79.3 | 51.83 | 370.4 |
| MLVCSL+BLIP | EN | 34.2 | 66.7 | 77.0 | 48.95 | 36.0 | 66.2 | 76.3 | 49.77 | 356.4 |
| MLVCSL+BLIP2 | EN | **40.3** | **70.2** | **80.6** | **53.60** | **39.8** | **70.6** | **80.4** | **53.90** | **381.9** |
| MLVCSL+ResNet | ZH | 28.9 | 55.5 | 69.2 | 41.54 | 30.4 | 57.8 | 69.7 | 43.07 | 311.5 |
| MLVCSL+CLIP | ZH | 36.4 | 68.0 | 79.3 | 51.52 | 39.1 | 69.4 | 79.8 | 52.80 | 372.0 |
| MLVCSL+BLIP | ZH | 35.9 | 66.6 | 77.5 | 50.06 | 36.7 | 66.4 | 76.9 | 50.28 | 360.0 |
| MLVCSL+BLIP2 | ZH | **40.9** | **70.9** | **80.3** | **54.28** | **40.5** | **72.0** | **79.8** | **54.12** | **384.4** |

4.2.4　模型消融

本节探讨 MLVCSL 每个模块的贡献. 为了展示模型多语言特性,分别对中文和德语的性能进行了消融测试,在 VATEX 上的消融实验结果如表 7 所示,对于中文测试的性能(1～4 行),与去除对比学习模块的性能相比(第 1 行),基于文本遮盖和基于视频抽帧的模型在 SumR 上分别提高了 5.7 和 4.2,德语(GR,5～8 行)则分别提升了 8.4 和 13.8,综合使用这两个模块则可以进一步将性能指标 SumR 从 373.8 提升至 386.5,德语从 384.5 提升至 405.0,这表明两个模块之间在不同语言之间均互补有效. 其中,德语数据均由原始英语数据经机器翻译得到.

**表 7　在 VATEX 上的模型消融实验表现**

| 训练语言 | 文本遮盖 | 视频抽帧 | T2V | | | | V2T | | | | SumR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | R@1 | R@5 | R@10 | mAP | R@1 | R@5 | R@10 | mAP | |
| EN+ZH | | | 31.7 | 66.0 | 75.7 | 46.66 | 44.9 | 73.9 | 81.9 | 34.42 | 373.8 |
| EN+ZH | ✔ | | 31.7 | 66.1 | 76.6 | 47.07 | 44.7 | 75.4 | 84.9 | 34.68 | 379.5 |
| EN+ZH | | ✔ | 32.4 | 66.7 | 76.8 | 47.36 | 44.5 | 74.0 | 83.5 | 34.94 | 378.0 |
| EN+ZH | ✔ | ✔ | **33.1** | **67.1** | **77.1** | **48.18** | **46.7** | **76.6** | **85.9** | **35.40** | **386.5** |
| EN+GR | | | 35.6 | 71.5 | 80.6 | 51.32 | 44.4 | 71.6 | 80.9 | 40.66 | 384.5 |
| EN+GR | ✔ | | 36.2 | 71.3 | 80.7 | 51.66 | 45.7 | 75.3 | 83.7 | 38.78 | 392.9 |
| EN+GR | | ✔ | 36.8 | **73.6** | **82.2** | 52.80 | 46.9 | 75.5, | 83.4 | 40.44 | 398.3 |
| EN+GR | ✔ | ✔ | **38.2** | 73.4 | **82.2** | **53.65** | **48.3** | **76.8** | **86.2** | **40.66** | **405.0** |

#### 4.2.5 模型 t-SNE 可视化分析

为了进一步证明本文模型结构的有效性，以完全消融后的基础模型，即去除文本端和视频端的对比学习框架后的模型作为比较对象，并在 VATEX 上进行了基于 t-SNE 的可视化分析. 具体地，随机选择了 20 个视频以及对应的中英文本并进行可视化分析. 如图 4 所示，可以看出，MLVCSL 模型的文本视频表示相互之间较消融后的模型更加紧凑，即模型的编码效果更好，以此证明所提出的 MLVCSL 结构的有效性.

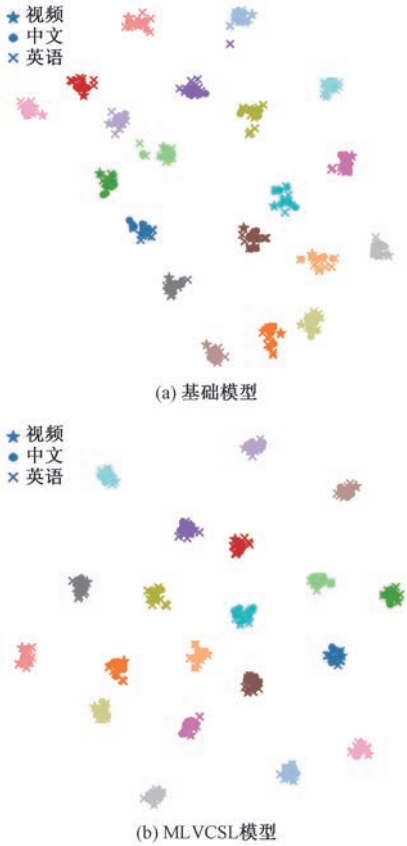



图 4 基础模型和 MLVCSL 模型的 t-SNE 可视化(相同颜色的点代表它们始于同一类)

#### 4.2.6 模型抗噪音鲁棒性分析

本实验探究模型对翻译噪声的鲁棒性. 我们选择 NRCCR[9] 模型以及基础模型(BASIC)作为对比模型. NRCCR 为近期提出的跨语言跨模态检索模型，已被证明具有较强的抗翻译噪音能力；基础模型为我们提出的模型中去除文本端和视频端的对比学习结构后得到的模型. 我们在 VATEX 上进行了不同程度的训练噪声实验，分别通过两个方式得到噪音程度不同的训练数据：使用机器翻译(英语→中文)获得的数据训练模型；使用两次机器翻译(英语→中文→英语→中文)获得的数据训练模型. 为了便于描述，将前者称为谷歌翻译，将后者称为谷歌翻译++. 因后者使用更多次的机器翻译，所以后者得到的训练数据噪声更大. 如图 5 所示，本文所提出的 MLVCSL 模型在两种噪音下都表现出更好的性能. 在噪音更大的谷歌翻译++的数据上，NRCCR 模型相比于谷歌翻译数据的基础上略有下降但依旧保持良好的性能，抗噪音干扰能力都较为突出，但基础模型受到翻译噪音加强的影响后性能退化较为明显. 这显示了本文提出 MLVCSL 模型对于翻译产生的噪音具有更好的鲁棒性.

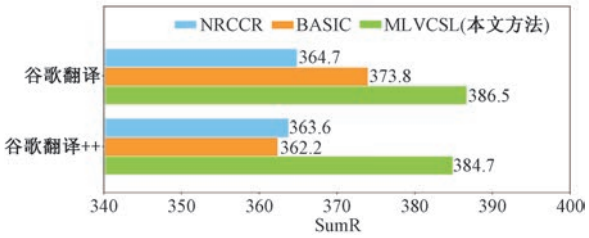



图 5 训练期间不同程度噪声的性能比较

#### 4.2.7 模型检索效率分析

对模型的检索效率进行分析. NRCCR 是当下性能最好且代码开源的具有代表性的多语言跨模态模型，因此我们在实验中与 NRCCR 进行对比. 给定一个中英文本和对应的视频，计算量和参数量在编码这两个文本样本和对应视频样本时统计的. 注意，由于两个模型使用相同的视频特征预提取方法，所以未将该步骤的计算量和模型参数包括在内. 计算量和参数量对比如表 8 所示，MLVCSL 在模型性能提升的同时，模型的计算量和参数量和 NRCCR 模型相差不大.

**表 8 推理阶段模型大小和计算开销的比较**

| 模型 | 模型计算量 FLOPs/G | 模型参数 Parameters/M |
|---|---|---|
| NRCCR | 3.10 | 187.57 |
| MLVCSL | 3.27 | 197.40 |

#### 4.2.8 模型检索能力定性展现

为了定性地展示模型检索能力，图 6 列出了从 VATEX 中随机选择的两个中英文本对，并展示了视频检索的前三位，本文模型在中英文上都展现了令人满意的检索效果，除正确的结果外，前三位的检索结果都基本和文本在语义上相互关联. 由此可见，本文模型在一定程度上跨越了语言之间的语义鸿沟，实现了一种简单高效的多语言文本-视频跨模态检索模型.


图 6　MLVCSL 多语言跨模态检索示例

## 5　总结语

本文提出了一种统一的基于多语言-视觉公共空间的多语言文本-视频跨模态学习框架 MLVCSL,能简单有效地适用于多种人工标注语言或机器翻译语言所构建的不同场景的多语言环境. 模型通过在文本端与视频端引入对比学习机制,降低特征噪音增强表示能力; 同时能够以视频特征为锚点在公共空间对多种语言特征进行对齐,可以有效地挖掘多语言之间文本互通性与抗鲁棒性. 在 VATEX 和 MSR-VTT 两个多语言视频文本数据集上的实验结果表明,本文方法无论是在(伪)平行语料场景还是在不平行语料场景,均展现了较好的性能并能显著地提高多语言跨模态的学习质量. 在未来的工作中将继续引入更多的语言实验来验证完善本文模型.

## 参　考　文　献

[1] Wu Fei, Zhuang Yue Ting. Internet cross media analysis and retrieval: Theory and algorithms. Journal of Computer Aided Design and Graphics, 2010, 22(1): 1-9 (in Chinese)
(吴飞, 庄越挺. 互联网跨媒体分析与检索: 理论与算法. 计算机辅助设计与图形学学报, 2010, 22(1): 1-9)

[2] Chen Zhuo, Du Hao, Wu Yu Fei, et al. Cross-modal video clip retrieval based on visual-text relationship alignment. SCIENTIA SINICA Informationis, 2020, 50(6): 862-876. (in Chinese)
(陈卓, 杜昊, 吴雨菲等. 基于视觉-文本关系对齐的跨模态视频片段检索. 中国科学: 信息科学, 2020, 50(6): 862-876)

[3] Huang Qing Ming, Wang Shu Hui, Xu Qian Qian, et al. Cross media analysis and inference centered on image and video. Journal of Intelligent Systems, 2021, 16 (5): 835-848 (in Chinese)
(黄庆明, 王树徽, 许倩倩等. 以图像视频为中心的跨媒体分析与推理. 智能系统学报, 2021, 16(5): 835-848)

[4] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018

[5] Lample G, Conneau A. Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291, 2019

[6] Portaz M, Randrianarivo H, Nivaggioli A, et al. Image search using multilingual texts: A cross-modal learning approach between image and text. arXiv preprint arXiv:1903.11299, 2019

[7] Aggarwal P, Kale A. Towards zero-shot cross-lingual image retrieval. arXiv preprint arXiv:2012.05107, 2020

[8] Huang P Y, Patrick M, Hu J, et al. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. arXiv preprint arXiv:2103.08849, 2021

[9] Wang Y, Dong J, Liang T, et al. Cross-lingual cross-modal retrieval with noise-robust learning//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portuguese, 2022: 422-433

[10] Faghri F, Fleet D J, Kiros J R, et al. Vse++: Improving visual-semantic embeddings with hard negatives. arXiv pre-

print arXiv:1707.05612, 2017

[11] Li X, Xu C, Yang G, et al. W2vv++ fully deep learning for ad-hoc video search//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France, 2019: 1786-1794

[12] Chen S, Zhao Y, Jin Q, et al. Fine-grained video-text retrieval with hierarchical graph reasoning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2020: 10638-10647

[13] Dong J, Li X, Snoek C G. Predicting visual features from text for image and video caption retrieval. IEEE Transactions on Multimedia. Seoul, Republic of Korea, 2018, 20(12): 3377-3388

[14] Miech A, Zhukov D, Alayrac J B, et al. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 2630-2640

[15] Chen J, Hu H, Wu H, et al. Learning the best pooling strategy for visual semantic embedding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 15789-15798

[16] Li X, Zhou F, Xu C, et al. Sea: Sentence encoder assembly for video retrieval by textual queries. IEEE Transactions on Multimedia. Seattle, USA, 2020, 23: 4351-4362

[17] Dong J, Wang Y, Chen X, et al. Reading-strategy inspired visual representation learning for text-to-video retrieval. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(8): 5680-5694

[18] Gabeur V, Sun C, Alahari K, et al. Multi-modal transformer for video retrieval//Computer Vision-ECCV 2020: 16th European Conference, Part IV 16. Glasgow, UK, 2020: 214-229

[19] Wang X, Zhu L, Yang Y. T2vlad: Global-local sequence alignment for text-video retrieval//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 5079-5088

[20] Dong J, Li X, Xu C, et al. Dual encoding for video retrieval by text. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(8): 4065-4080

[21] Liu Y, Albanie S, Nagrani A, et al. Use what you have: Video retrieval using representations from collaborative experts. arXiv preprint arXiv:1907.13487, 2019

[22] Wang W, Gao J, Yang X, et al. Many hands make light work: Transferring knowledge from auxiliary tasks for video-text retrieval. IEEE Transactions on Multimedia, 2022, 25: 2661-2674

[23] Li J, Li D, Xiong C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation//International Conference on Machine Learning. Maryland, USA, 2022: 12888-12900

[24] Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597, 2023

[25] Dai W, Li J, Li D, et al. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv: 2305.06500, 2023

[26] Madasu A, Aflalo E, Stan G B M, et al. Improving video retrieval using multilingual knowledge transfer. arXiv preprint arXiv:2208.11553, 2022

[27] Jain A, Guo M, Srinivasan K, et al. Mural: Multimodal, Multitask retrieval across languages. arXiv preprint arXiv: 2109.05125, 2021

[28] Zhang L, Hu A, Jin Q. Multi-lingual acquisition on multimodal pre-training for cross-modal retrieval. Advances in Neural Information Processing Systems, 2022, 35: 29691-29704

[29] Wan Z, Liu C, Zhang M, et al. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. Advances in Neural Information Processing Systems, 2024, 36: 56186-56197

[30] Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018

[31] Wang X, Wu J, Chen J, et al. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 4581-4591

[32] Xu J, Mei T, Yao T, et al. Msr-vtt: A large video description dataset for bridging video and language//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seoul, Republic of Korea, 2016: 5288-5296

[33] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 6299-6308

[34] Dong J, Li X, Snoek C G. Word2visualvec: Cross-media retrieval by visual feature prediction. arXiv preprint arXiv: 1604.06838, 2016, 2

[35] Yang X, Dong J, Cao Y, et al. Tree-augmented cross-modal encoding for complex-query video retrieval//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Xi'an, China , 2020: 1339-1348

[36] Miech A, Laptev I, Sivic J. Learning a text-video embedding from incomplete and heterogeneous data. arXiv preprint arXiv:1804.02516, 2018

[37] Wang Y, Wang H, Shen Y, et al. Semi-supervised semantic segmentation using unreliable pseudo-labels//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 4248-4257

[38] Han J, Shoeiby M, Petersson L, et al. Dual contrastive learning for unsupervised image-to-image translation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 746-755

[39] Li Z, Zhang X, Zhang Y, et al. Towards general text embeddings with multi-stage contrastive learning. arXiv preprint arXiv:2308.03281, 2023

## 附　录

(1)损失函数的选择

本文的 Triplet 损失和 InfoNCE 损失在形式上较为相近,都是拉近相似的样本,拉远不相似的样本. 区别在于,本文采用的 Triplet 损失选择了最困难的样本计算损失,而 InfoNCE 损失则计算了正样本与所有负样本的损失. 本文的模型共涉及两类损失,同模态损失和跨模态损失,同模态损失即文本和遮盖文本之间的损失,以及视频和抽帧视频之间的损失. 跨模态损失即文本和视频之间的损失. 此外,InfoNCE 常用于模态内对比学习的损失[30,37-39],基于最困难样本的 Triplet 损失则常用于跨模态计算[10,16,17,20]. 因此,模型的同模态损失选用了 InfoNCE 损失,跨模态损失选择了 Triplet 损失.

表 9 给出了同模态损失两者的性能对比. 实验结果表明,无论是文本端还是视频端,使用 InfoNCE 损失的对比学习性能均更优,证明了使用该损失的有效性. 表 10 给出了跨模态的损失结果,可以发现,跨模态损失采用 Triplet 损失比 InfoNCE 损失在性能上更优. 上述实验证明了我们模型的损失函数选用在跨语言跨模态任务上的合理性.

**表 9　在 VATEX 上探究模型同模态损失对结果的影响**

| 文本遮盖 | 视频抽帧 | T2V | | | | V2T | | | | SumR |
|---|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | mAP | R@1 | R@5 | R@10 | mAP | |
| Triplet | Triplet | 31.5 | 65.9 | 76.1 | 46.91 | 45.2 | 73.9 | 84.2 | 34.15 | 376.8 |
| InfoNCE | Triplet | 31.9 | 66.4 | 76.4 | 47.11 | 46.0 | 75.9 | 84.5 | 34.74 | 381.2 |
| Triplet | InfoNCE | 32.3 | 66.9 | 77.2 | 47.63 | 45.9 | 74.8 | 85.1 | 34.60 | 382.3 |
| InfoNCE | InfoNCE | **33.1** | **67.1** | **77.1** | **48.18** | **46.7** | **76.6** | **85.9** | **35.40** | **386.5** |

**表 10　在 VATEX 上探究模型跨模态损失对结果的影响**

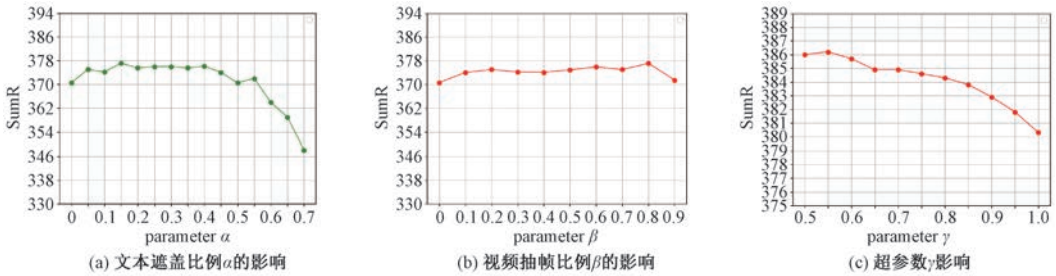
| 跨模态损失 | T2V | | | | V2T | | | | SumR |
|---|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | mAP | R@1 | R@5 | R@10 | mAP | |
| InfoNCE | 32.9 | 67.0 | **78.7** | **48.32** | 45.1 | 73.6 | 82.9 | 34.85 | 380.2 |
| Triplet | **33.1** | **67.1** | 77.1 | 48.18 | **46.7** | **76.6** | **85.9** | **35.40** | **386.5** |

(2)各个模块超参数的确定

文本遮盖百分比 $\alpha$ 对检索性能的影响. 超参数 $\alpha$ 代表遮盖原始文本词的比例. 如图 7(a)所示,取 0.15 时性能最佳,故本文 $\alpha$ 取值 0.15,[0.15,0.4]是一个较为合适的超参数取值区间,其中横坐标为 0 则代表去除遮盖结构. 适当地遮盖一些文本可以提升文本端的语义理解能力进而提升检索性能,因此随着的 $\alpha$ 增加,模型性能较之于未增加遮盖有一定的提升. 但随着遮盖的比例的上升,句子本身的语义已经不完整了,因此 $\alpha$ 超过 0.55 时,性能会逐渐下降.

视频抽帧百分比 $\beta$ 对检索性能的影响. 超参数 $\beta$ 代表对原始视频帧的抽出比例. 如图 7(b)所示,取 0.8 时性能最佳,故本文 $\beta$ 取值 0.8,[0.5,0.8]是一个较为合适的超参数取值区间,其中横坐标为 0 则代表去除抽帧结构. 请注意,视频抽帧比例要远大于文本的遮盖比例. 这是因为,视频帧之间有较大的冗余性,所以删除帧对语义损耗较文本要少,模型性能随 $\beta$ 的变化波动较小.

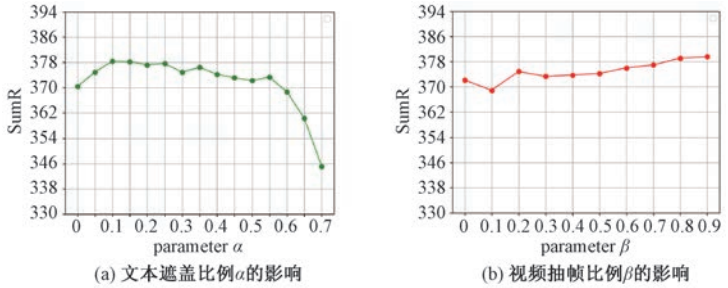
超参数 $\gamma$ 对检索性能的影响. 超参数 $\gamma$ 用于平衡原始语言文本和翻译语言文本的重要性. 公式 4 中,需要讨论 $\gamma$ 具体取值对模型性能的影响. 由于翻译语言带来了一定的噪音,因此可以认为原语言重要程度应大于翻译语言. 因此将 $\gamma$ 取值范围设为[0.5,1.0],间隔为 0.05,在伪平行多语言语料场景下训练. 图 7(c)展示了实验结果. $\gamma$ 取 0.55 时实验性能最佳,故本文 $\gamma$ 取值 0.55,而[0.5,0.7]都是一个较为合适的取值区间.




(a) 文本遮盖比例α的影响　(b) 视频抽帧比例β的影响　(c) 超参数γ影响

图 7　不同参数对模型性能的影响

不平行语料库下关于遮盖参数和抽帧比例的实验. 图 8 展示了在不平行语料库上,即没有天然语料对齐的情况下的随机遮盖和抽帧实验. 图 8(a)显示了不同文本遮盖比例影响,其中横坐标为 0 代表不使用文本遮盖模块. 图 8(b)显示了视频抽帧比例的影响,其中横坐标为 0 则代表不使用视频抽帧模块. 我们发现文本遮盖比例在区间[0.1,0.25]性能表现较好;视频抽帧比例在区间[0.7,0.9]性能表现较好,且都优于不使用文本遮盖模块或视频抽帧模块的性能. 实验结果进一步证明了我们提出的文本遮盖结构和视频抽帧的有效性.




(a) 文本遮盖比例$\alpha$的影响 (b) 视频抽帧比例$\beta$的影响

图 8 不平行语料库下,遮盖和抽帧对模型性能的影响

(3)模拟噪音训练的有效性

理论上,通过随机抹去少量 token 能让原始文本的语义信息产生轻微的不完整,从而带来噪音. 为了进一步说明通过随机遮盖少量文本来模拟噪音是有效的,将没有进行遮盖抗噪训练和进行遮盖抗噪训练的模型进行检索实验对比. 如图 9 所示,对于 VATEX 中的一个文本示例“一个女人在教另一个人歌曲的手语动作”,将其中的“个”“一”“的”删除后变成“一女人在教另个人歌曲手语动作”,前后语义几乎没变. 经过抗噪训练的模型几乎不受影响地检索出对应的视频(正确样本检索排名从第 1 位下降到了第 2 位),而未遮盖抗噪训练的模型的性能受到了较大影响(正确样本检索排名从第 1 位下降到了第 12 位). 上述例子说明了模拟噪音训练使得模型的鲁棒性得到了提升.


图 9 噪音训练鲁棒性实验检索对比

另一种更普遍的情况是,不同的人对同一场景的描述很可能是不同的,本文采用重写来模拟这种表述不同. 在 VATEX 中选取了一个文本示例,并使用预训练大语言模型 GPT3.5-turbo 进行重写,语义不发生变化. 如图 10 所示,对于重写后的文本查询,经过抗噪训练的模型依旧能很好的检索出对应的视频,而未遮盖抗噪训练的模型则受噪音变化干扰导致检索性能受到了影响,对正确样本检索的排名有了一定的下降。实验展示了模拟噪音训练有助于提升模型对于不同文本查询的泛化性. 由此进一步证明少量遮盖文本模拟噪音训练的有效性.


图 10　噪音训练泛化性实验检索对比

(4)多语言直接对齐有损性能

在一开始的设想中,为了充分挖掘(伪)平行语料库的天然对齐信息,在实验时引入了多语言直接对齐训练,同样采用了改进的三元排序损失. 但实验结果证明,添加多语言直接对齐的损失训练对模型性能有害. 如表 11 所示,在原有 VATEX 平行语料库训练 MLVCSL 的基础上加入了语言直接对齐损失训练,结果证明无论中文和英文,模型性能都受到了不利影响. 我们推测这可能是因为任务的目标依旧是语言到视频两种模态的跨越,并不是语言到语言的任务,添加平行语料之间的对齐可能会干扰目标任务的训练,分散模型的注意力.

**表 11　在 VATEX 上探究多语言直接对齐对结果的影响**

| 语言直接对齐 | 测试语言 | T2V | | | V2T | | | Sum*R* |
|---|---|---|---|---|---|---|---|---|
| | | *R*@1 | *R*@5 | *R*@10 | *R*@1 | *R*@5 | *R*@10 | |
| 有 | EN | **38.6** | 74.2 | 82.8 | 47.9 | 76.6 | 85.1 | 405.3 |
| 无 | EN | 38.2 | **74.8** | **83.9** | **50.3** | **78.0** | **86.3** | **411.4** |
| 有 | ZH | 35.6 | 70.7 | 80.3 | 47.1 | 76.5 | 86.3 | 396.5 |
| 无 | ZH | **36.3** | **71.3** | **81.2** | **49.9** | **79.8** | **88.2** | **406.6** |

**LIN Jun-An**, B. S.. His research interests are cross-modal retrieval and localization.

**BAO Cui-Zhu**, Ph. D., lecturer. Her research interests are computer vision and intelligent traffic control.

**DONG Jian-Feng**, Ph. D., researcher. His research interests are multimedia understanding and computer vision.

**YANG Xun**, Ph. D., professor. His research interests are cross-media analysis and inference, structured understanding of multimedia content.

**WANG Xun**, Ph. D., professor. His research interests are mobile graphic computing and computer vision.

**Background**

In recent years, related research in the field of video-text cross-modal retrieval has developed rapidly. However, existing models are often limited to a single language, i. e. English, which has abundant annotation data. Faced with the retrieval tasks of another language, these models are often retrieved through model retraining or machine translation of the language into English, but the former will greatly increase the time cost, while the latter will introduce translation noise and reduce the performance. Facing the development trend of Internet diversity and convergence, the demand for video-text retrieval models that can handle multiple languages has been increasing.

With the deepening study, many multilingual video-text cross-modal retrieval models are coming forth at present. For example, MMP builds a multilingual visual text conversion model based on Transformer, NRCCR solves the problem caused by machine translation noise through distillation, and MKTVR realizes multilingual knowledge transfer based

on CLIP. Different from the above models, the MLVCSL proposed in this paper is a unified multilingual text-video cross-modal retrieval model, covering different multilingual application scenarios, non-parallel corpus, parallel corpus and pseudo-parallel corpus. Solve multilingual retrieval problems using one model. MLVCSL constitutes its main model structure based on mBERT pre-training on the text side and three feature extraction on the video side, and establishes a comparative anti-noise robustness learning method on both side. Based on the above methods, this paper builds a simple and unified multilingual text-video cross-modal retrieval model via multilingual-visual common space learning.

We have also conducted extensive experiments on VATEX and MSR-VTT, which demonstrate that the performance of MLVCSL is also at a relatively advanced level at present.

This work is supported by the Zhejiang Province 'Jianbing' and 'Lingyan' Research and Development Plan Project (No. 2023C01212), Zhejiang Province Basic Public Welfare Technology Research Program (No. LGF21F020010), and Young Elite Scientists Sponsorship Program by China Association for Science and Technology (No. 2022QNRC001).