

基于思维链推理的双向组合式视频检索

孟一平¹⁾ 刘利龙²⁾ 丁智祥³⁾ 杨振宇³⁾ 钱胜胜³⁾

¹⁾(天津理工大学计算机科学与工程学院 天津 300384)

²⁾(郑州大学河南先进技术研究院 郑州 450003)

³⁾(中国科学院自动化研究所多模态人工智能系统全国重点实验室 北京 100190)

摘 要 零样本组合式视频检索(Zero-Shot Composed Video Retrieval, ZS-CVR)旨在根据用户提供的参考视频及其自然语言修改指令(如“找到类似但场景在雪地中的视频”),从未经标注的候选视频库中精准定位目标视频。该任务的核心挑战在于深入理解复杂的用户意图,特别是修改指令引入的语义变化及其与参考视频内容的关联。然而,现有方法通常将参考视频转化为文本描述并与修改文本合并,导致视觉信息丢失和检索意图理解不足,且仅依赖编辑描述进行检索,易因语义偏差而影响精度。针对这些挑战,本文提出了一种基于思维链推理的双向组合式视频检索方法(Bidirectional Chain-of-Thought Reasoning for Composed Video Retrieval, BC4CVR)。该方法利用思维链推理生成目标视频描述以理解用户意图,并引入双向检索机制,通过一致性评估对检索结果重排序。BC4CVR不仅有效解决了现有方法存在丢失关键语义信息和对用户检索意图理解不足的问题,并通过双向检索的一致性,消除生成的目标视频描述带来的语义偏差,进一步优化检索结果。本研究的方法在两个主要基准数据集上进行了广泛实验,结果表明,BC4CVR在WebVid-CoVR-Test数据集上的R@1为53.40%,在EgoCVR数据集上的R@1为12.24%,R@5与R@10指标也均高于现有方法的检索效果,充分证明了本方法的优越性能。

关键词 组合式视频检索;思维链推理;预训练视觉-语言模型;跨模态检索;多模态大模型

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2026.01703

Bidirectional Chain-of-Thought Reasoning for Composed Video Retrieval

MENG Yi-Ping¹⁾ LIU Li-Long²⁾ DING Zhi-Xiang³⁾ YANG Zhen-Yu³⁾ QIAN Sheng-Sheng³⁾

¹⁾(School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384)

²⁾(Henan Institute of Advanced Technology, Zhengzhou University, Zhengzhou 450003)

³⁾(National Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190)

Abstract Zero-Shot Composed Video Retrieval (ZS-CVR) aims to retrieve a target video from an unannotated candidate set according to a reference video and a natural language modification instruction provided by the user. Different from conventional text-video retrieval, this task requires the model not only to understand the original semantic content of the reference video, but also to accurately capture the semantic changes introduced by the modification text. Therefore, the model needs to simultaneously handle the information that should be preserved and the information that should be modified, which makes ZS-CVR a challenging task in compositional semantic understanding. Existing zero-shot methods usually first convert the reference video into a textual description, and then combine it with the modification instruction to generate a target

收稿日期:2025-07-22;在线发布日期:2026-04-20。本课题得到国家重点研发计划(2023YFC3310700)、国家自然科学基金(No. 62276257)资助。孟一平,博士研究生,主要研究领域为多媒体内容分析和跨模态检索。E-mail: mypkitty@gmail.com。刘利龙,硕士研究生,主要研究领域为人工智能与自然语言处理。丁智祥,硕士研究生,主要研究领域为多媒体内容分析、数据挖掘与跨模态检索。杨振宇,博士研究生,主要研究领域为多模态理解与应用。钱胜胜(通信作者),博士,研究员,中国计算机学会(CCF)高级会员,主要研究领域为多媒体内容分析、数据挖掘、跨模态检索与个性化推荐。E-mail: shengsheng.qian@nlpr.ia.ac.cn。

video description for retrieval. Although such methods avoid dependence on large-scale annotated triplet data, they still have two main limitations. On the one hand, important visual information related to the modification text may be lost during the captioning process, which weakens the modeling of the user's actual retrieval intent. On the other hand, retrieval based only on the generated target description is easily affected by semantic bias in the generated text, which may lead to inaccurate retrieval results.

To address these issues, this paper proposes a Bidirectional Chain-of-Thought Reasoning for Composed Video Retrieval method, termed BC4CVR. The proposed method consists of two main components, namely a Chain-of-Thought reasoning module and a bidirectional retrieval module. In the Chain-of-Thought reasoning module, the generation process of the target video description is decomposed into four sub-tasks, including reference video parsing, modification instruction decomposition, cross-modal modification execution, and target description generation. By guiding a multimodal large model to perform step-by-step reasoning based on the reference video and the modification text, the proposed method generates a target video description that is more consistent with the user's retrieval intent. Compared with directly concatenating or simply editing video descriptions, this strategy can improve the fine-grained modeling of both preserved semantics and modified semantics to a certain extent.

On this basis, a bidirectional retrieval module is further designed to alleviate the semantic bias that may be introduced during target description generation. Specifically, the generated target video description is first used to perform forward retrieval and obtain an initial candidate ranking. Then, the Top-N videos from the forward retrieval results are respectively regarded as new reference videos, and the original modification instruction is semantically reversed to generate backward retrieval descriptions. These backward descriptions are used to evaluate whether the candidate videos can be semantically traced back to the original reference video. According to the consistency between forward retrieval and backward retrieval, the initial ranking results are re-ranked, thereby improving the reliability of the final retrieval results. This mechanism helps enhance the discrimination ability of the model against highly similar distractor samples and improves retrieval robustness.

Experiments are conducted on two benchmark datasets, namely WebVid-CoVR-Test and EgoCVR. The experimental results show that the proposed method achieves favorable retrieval performance on both datasets. In particular, BC4CVR obtains an R@1 of 53.40% on WebVid-CoVR-Test and an R@1 of 12.24% on EgoCVR, while also achieving better R@5 and R@10 results than the compared zero-shot methods. These results indicate that the proposed method can effectively improve the modeling of user retrieval intent in zero-shot composed video retrieval, and that the bidirectional consistency verification strategy is beneficial for further optimizing retrieval results. The present study provides a feasible new solution for zero-shot composed video retrieval.

Keywords composed video retrieval; chain-of-thought reasoning; pre-trained vision-language models; cross-modal retrieval; multimodal large language models

1 引 言

组合式视频检索(Composed Video Retrieval,

CVR)是跨模态检索任务的一个重要扩展,旨在通过引入视频与文本的双模态信息,来检索符合目标的视频^[1-3]。与传统的跨模态检索任务不同,CVR不仅关注参考视频和目标视频之间的相似性,还考

虑了由修改文本描述引入的语义差异。具体而言，给定一个参考视频和一段修改文本，目标是检索出一个与参考视频整体相似，但在修改文本所描述的方面存在特定差异的视频。

为推动组合式视频检索的研究，Ventura 等人^[1]构建了一个大规模的组合式视频检索数据集 WebVid-CoVR。该数据集包含了一个训练集和较小规模的评估集，数据集中的每个样本均由一个参考视频、修改文本和目标视频构成三元组，总数据规模达 160 万条。WebVid-CoVR 的提出为 CVR 任务的进一步研究奠定了基础。此外，Hummel 等人^[2]也提出了 EgoCVR 基准数据集，专注于动作变化的检索任务，进一步提升了组合式视频检索方法的挑战性。

现有的 CVR 方法^[1,3,4]通常依赖于大量带标注的三元组(参考视频,修改文本,目标视频)进行有监督训练,通过多样化的训练策略来优化模型性能。这些方法通过训练过程不断提高检索效果,但由于它们在很大程度上依赖于标注数据的质量和规模,

模型的泛化能力通常较弱。

为了解决现有方法对大量标注数据的依赖,零样本组合式视频检索(Zero-Shot Composed Video Retrieval, ZS-CVR)逐渐成为一个重要的研究方向。ZS-CVR 方法^[2]旨在通过预训练视觉-语言模型,自动生成参考视频的视觉文本描述,并利用大语言模型(Large Language Models, LLMs)将该描述与修改文本进行语义组合,生成编辑后的描述文本。通过这种方法,系统能够在不依赖大量标注数据的情况下,进行有效的视频检索。然而,如图 1 所示现有的 ZS-CVR 方法存在两个主要局限:一方面,生成参考视频描述再结合修改文本得到目标视频描述的两阶段方法,会导致丢失参考视频中相关的视觉信息,并且视频描述与修改文本的简单结合未能充分挖掘用户的真实检索意图。另一方面,现有方法通常只通过生成的目标视频描述与候选视频的相似度进行检索,忽略了目标视频描述中存在的语义偏差,从而限制了检索结果的准确性。

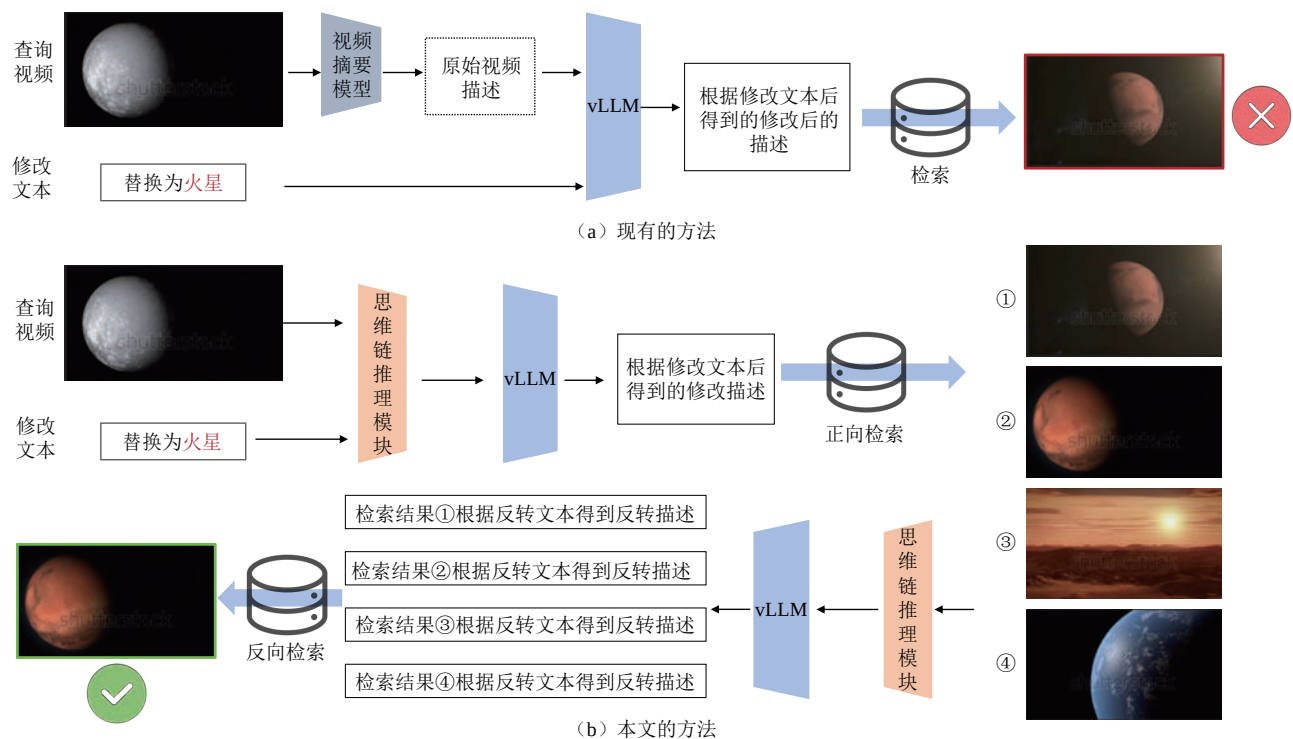


图 1 现有零样本组合式视频检索方法与本文方法的对比示意图

针对这些挑战,本文提出了一种新颖的零样本组合式视频检索方法——基于思维链推理的双向组合式视频检索方法(Bidirectional Chain-of-Thought Reasoning for Composed Video Retrieval, BC4CVR)。该方法通过引入思维链推理和双向检索机制,显

著提升了 ZS-CVR 任务的检索性能。具体而言,BC4CVR 首先利用思维链推理,按照预定义的子任务,引导模型逐步推理得到目标视频描述,通过计算目标视频描述和候选视频的相似度检索得到 Top N 视频集,分别将 Top N 结果作为参考视频,并对修

改文本进行反转,生成反向检索的视频描述并计算其与原始视频的相似度,对初始的检索结果重排序以提高满足双向检索一致性视频的排名,从而消除生成的目标视频描述带来的语义偏差,有效提高了检索结果的准确性。

在 WebVid-CoVR-Test^[1]和 EgoCVR^[2]数据集上的大量实验结果表明,BC4CVR方法在性能上显著优于现有的 ZS-CVR 方法,验证了其在零样本组合式视频检索任务中的有效性和先进性。具体来说,本文的方法通过思维链推理生成目标视频描述,并通过双向检索优化了相似度排序,显著提高了检索的准确性。本文的主要贡献如下:

(1) 提出了一种面向零样本组合式视频检索 (Zero-Shot Composed Video Retrieval, ZS-CVR) 任务的方法——基于思维链推理的双向组合式视频检索方法 (BC4CVR)。该方法无需依赖任务特定训练数据,即可仅基于参考视频与修改文本实现对目标视频的准确检索。——基于思维链推理的双向组合式视频检索方法 (BC4CVR),该方法无需训练数据即可基于参考视频和修改文本准确检索目标视频。

(2) 提出了一个思维链推理模块,通过设计精细的子任务,引导多模态模型根据参考视频和修改文本逐步推理,最终生成符合用户检索意图的目标视频描述。

(3) 设计了一个双向检索模块,通过对修改的文本进行反转,生成反向检索的视频描述以评估双向检索的一致性,根据评估结果对正向检索的结果进行调整,进一步提升检索准确性。

(4) 在 WebVid-CoVR-Test 和 EgoCVR 数据集上的大量实验证明了本文提出的 BC4CVR 方法的有效性与先进性,验证了该方法在零样本组合式视频检索中的优越性。

2 相关工作

2.1 组合式图像检索

组合式图像检索 (Composed Image Retrieval, CIR)^[5-6]旨在根据参考图像和修改文本检索目标图像。近年来,深度学习的快速发展极大地推动了 CIR 的研究进展,其应用场景包括商品搜索、人脸识别和场景修改检索等^[7-9]。早期方法主要关注将图像-文本对投影至公共嵌入空间,或利用跨模态注意力机制,通常依赖于人工标注的图像-文本-图像三

元组进行训练^[10-14]。近期研究开始探索零样本组合图像检索 (Zero-Shot CIR, ZS-CIR),旨在消除对标注三元组的依赖,并提升模型的泛化能力。ZS-CIR 方法通常利用图像描述生成模型或文本反演技术将图像数据转化为文本^[15-16],从而能够灵活地与文本查询进行组合。Karthik 等人^[17]提出了一种无需训练的方法,使用预训练的视觉-语言模型 (如 CLIP^[18]) 进行图像描述,并结合 LLMs 根据修改文本对描述进行重构。然而,大多数方法忽视了 ZS-CIR 中语义模糊性的特征,即目标图像的语义并不由输入严格限定。为应对这一问题,近期研究提出了 LDRE 方法^[19],该方法无需训练,利用密集描述与发散式组合推理,通过语义相关性加权集成策略整合多个描述,以增强在语义模糊的零样本场景下的鲁棒性与准确性。为进一步提升零样本鲁棒性,SEIZE 方法^[20]以“语义编辑增量”建模参考图像的多描述与 LLM 发散推理,度量编辑前后图文相似度变化并重排候选,兼顾最终相似度与语义增益。

2.2 组合式视频检索

组合式视频检索 (Composed Video Retrieval, CVR) 将组合式检索的概念扩展至视频领域。其在实际应用中具有广泛潜力,例如在电商平台的视频检索场景中,用户可以通过上传参考视频 (如某款服装的展示视频) 并输入修改需求 (如“找到类似但颜色为蓝色的款式”),系统能够准确检索出符合要求的目标商品视频,极大提升了购物体验 and 检索效率。当前文本到视频的检索任务在多媒体信息检索领域中取得了显著突破^[21-23],早期方法多采用基于内容的检索技术,例如关键帧分析、颜色直方图以及局部特征匹配。随着视觉-语言模型 (Vision-Language Model, VLM) 的出现,如 BLIP^[24],大量方法通过引入多模态嵌入和注意力机制,革新了文本到视频的检索方式^[25-28]。

近年来,CVR 任务受到关注,其目标是在给定参考视频和修改文本的情况下,检索符合修改语义的目标视频。最新研究提出了一种框架,通过详细的语言描述显式编码与查询相关的上下文信息,从而提升视频检索性能。该方法在视觉、文本及视觉-文本三种模态中学习判别性嵌入表示,并取得了当前最优的性能表现^[3]。另一个研究方向提出了特征解耦与跨模态对齐方法 (Feature Disentanglement and Cross-modal Alignment, FDCA),该方法在句子级与词元级进行特征解耦,增强视频与文本数据的融合与对齐,从而提升检索准确性^[4]。

2.3 思维链推理

思维链(Chain-of-Thought, CoT)技术通过将复杂问题分解为序列化子任务^[29],在零样本^[30]和少样本^[31-32]场景中显著提升模型推理能力。该技术已拓展至多模态领域:MM-CoT^[33]构建两阶段推理框架,DD-CoT^[34]实现文本问题的递进式求解,CCoT^[35]则专注于视觉场景理解。然而,现有方法存在领域局限性^[36-37],为此零样本组合检索领域提出创新方案—OSrCIR^[38]采用反思式CoT增强用户意图理解,CIRCOT^[39]则首创预分解范式驱动结构化推理。在此基础上,本研究提出BC4CVR框架,将任务预先分解为语义解构、链式推理、多样性生成三个核心子任务,通过多步推理机制实现目标视频的渐进式语义建模。

3 方法

3.1 问题描述

组合式视频检索(CVR)是一类复杂的跨模态检索任务,其核心在于同时处理视觉和文本模态的语义交互。给定参考视频 V_r 及其对应的修改文本指令 T_q ,该任务的目标是从视频数据库 $\mathcal{D}=\{V_i\}_{i=1}^N$ 中检索出最符合用户意图的目标视频 V_t 。理想的目标视频需同时满足两个关键条件:首先要严格遵循修改文本 T_q 指定的内容变更(如场景转换、对象

属性修改等),同时保留参考视频 V_r 中未被修改文本提及的潜在语义信息(如背景风格、对象关系等)。

零样本组合式视频检索(ZS-CVR)进一步限定检索过程不依赖任何任务特定标注数据,要求模型具备跨模态理解和组合推理能力。该任务可形式化为学习评分函数:

$$\Phi:(V_r, T_q, D) \mapsto V_t \quad (1)$$

该函数用于衡量参考视频 V_r , 修改文本 T_q 和数据库中每个候选视频 v_i 之间的相似度,根据函数 Φ 得分最高的视频 V_t 被认为是目标视频。

3.2 整体架构

当前零样本组合式视频检索方法缺乏对用户真实检索意图的细粒度分析,且直接通过转化得到的目标视频描述来检索目标视频会引入不可避免的语义偏差。针对现有方法的不足,本研究提出基于思维链推理的双向组合式视频检索方法(BC4CVR)。如图2所示,该方法包括思维链推理模块和双向检索模块,思维链推理模块通过将生成目标视频的任务分解为多个子任务,引导模型根据参考视频和修改文本逐步推理得到最终的目标视频描述,并与候选视频计算相似度得到检索结果。双向检索模块通过将正向检索的Top N结果分别作为参考视频,然后对修改文本进行反转,提示模型生成反向检索的视频描述,通过计算其与原始参考视频的相似度,对正向检索的排名结果进行重排,进一步提升检索准确性。

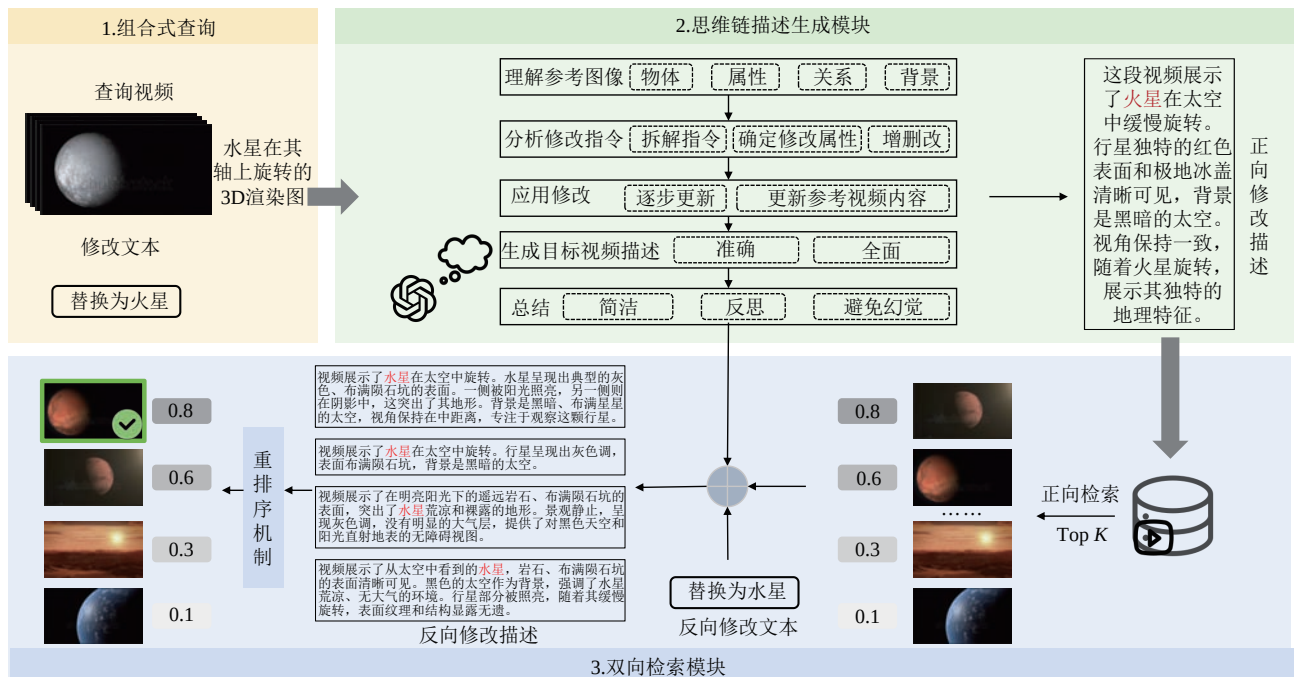


图2 本文的主要框架(包括思维链描述生成模块的设计及其与双向检索模块的整合)

3.3 思维链推理模块

组合式视频检索(CVR)面临的核心挑战是准确理解用户的检索意图,现有无训练方法一般先通过预训练视觉-语言模型(VLM)生成参考视频的描述,然后提示大语言模型(LLMs)根据修改文本生成编辑后的视频描述以检索目标视频。这种两阶段生成方法在生成参考视频描述时没有考虑修改文本的内容,导致丢失相关的视觉信息;另外仅仅通过简单提示大模型将视频描述和修改文本结合得到目标视频描述,没有充分利用大模型的推理能力。为解决这一问题,本研究提出了思维链描述模块,通过精心设计多个子任务,提示多模态大模型同时考虑参考视频和修改文本,逐步推理得到目标视频描述。整个推理过程被结构化为四个顺序执行的子任务,形成完整的思维链条。

参考视频解析:对输入视频 V_r 进行深度理解,提取关键视觉元素并构建结构化表示。该过程可形式化为 $\Phi(V_r) = (\mathcal{O}, \mathcal{A}, \mathcal{R})$, 其中 $\mathcal{O}$ 表示视频中的对象集合, $\mathcal{A}$ 记录各对象的属性特征, $\mathcal{R}$ 则表示对象间的空间与语义关系。例如,对于包含“红色汽车在蓝色房子左侧”的视频场景,模型需要准确识别对象集合 $\mathcal{O} = \{\text{汽车}, \text{房子}\}$, 属性映射 $\mathcal{A} = \{(\text{汽车}, \text{红色}), (\text{房子}, \text{蓝色})\}$ 以及空间关系 $\mathcal{R} = \{(\text{汽车}, \text{房子}, \text{左侧})\}$ 。

修改指令解构:该阶段专注于对修改文本的分析,要求模型理解修改文本,将修改文本分解为多个子步骤,并且推测每一步需要修改和保留的视觉元素。具体来说,给定修改文本 T_r , 大模型将其分解为原子操作序列 $\Psi(T_r) = \langle \tau_1, \tau_2, \dots, \tau_K \rangle$ 。每个原子操作 τ_k 包括操作类型(如添加 ADD、删除 DELETE、修改 MODIFY)和操作对象。例如,对于指令“将汽车颜色改为绿色并添加树木”,大模型将其分解为两个原子操作:修改汽车颜色(τ_1)和添加树木对象(τ_2)。

跨模态修改实施:该阶段要求大模型根据分解之后的操作类型和操作对象,对提取的参考视频视觉元素进行逐步修改,实现视觉表示的渐进式更新。该过程可以形式化表示为 $\Phi_{target} = \Gamma(\Phi(V_r), \Psi(T_r))$, 其中 Γ 为修改操作,逐步应用分解之后的修改步骤,确保每次操作仅修改对应元素,并保留其他不相关的内容,最终得到更新后的视觉描述 Φ_{target} 。

目标描述生成:该阶段基于更新后的视觉表示 Φ_{target} 生成最终目标视频描述 T_c 。要求模型生成覆盖所有修改后元素的视频描述,该过程通过三重约

束机制确保输出质量:语义完整性控制要求生成的描述覆盖多种语义维度(如人物、动作、环境、物体等);跨模态一致性约束防止描述中出现未在视觉表示中出现的实体(幻觉抑制);表达最优化则要求模型生成的描述简洁流畅,不要出现重复冗余的内容。

通过提示大模型逐步完成以上四个子任务,最终推理得到符合用户检索意图的目标视频描述,为进一步标准化推理的流程以及促进多模态大模型的推理能力,该模块通过提供具体示例进行上下文学习,从而辅助模型的理解与推理。

3.4 双向检索模块

在组合式视频检索中,高视觉相似性负样本引发的误检问题是影响检索精度的关键挑战。现有的组合式视频检索方法直接使用生成的目标视频描述进行文本-视频检索,但是描述中存在的有偏差语义内容会导致检索到的视频偏离目标视频,为解决这一问题,本研究提出双向检索一致性验证模块,通过将正向检索到的视频作为新的参考视频,反过来检索用户给定的视频,若能够成功反向检索,则说明该视频更有可能为目标视频,则对其原有的检索排名进行调整。

3.4.1 正向检索

为了进行跨模态检索,需要从所有模态中提取特征并将它们对齐到一个共同的特征空间。具体来说,对于视频数据库 $\mathcal{D} = \{V_i\}_{i=1}^N$ 中的每个视频 V_i , 数据库的候选视频数量为 N , 使用 LanguageBind 视频编码器 Φ_v 提取视频特征并对其进行归一化。形式化地,提取视频特征 H_v 如下:

$$H_v = \left\{ h_{vi} = \frac{\Phi_v(V_i)}{\|\Phi_v(V_i)\|_2} \mid 1 \leq i \leq N \right\} \quad (2)$$

使用 LanguageBind 文本编码器 Φ_t 提取目标视频描述 T_c 的文本特征 h_c 如下:

$$h_c = \frac{\Phi_t(T_c)}{\|\Phi_t(T_c)\|_2} \quad (3)$$

然后计算目标视频描述文本特征 h_c 和每个候选视频特征 h_{vi} 的相似度。形式化地,计算正向检索相似度 S 为

$$S = \{s_1, s_2, \dots, s_n\} = h_c(H_v)^T \quad (4)$$

基于计算得到的相似度,从高到低对视频进行排序得到正向检索的 Top N 视频排名结果为

$$\text{Rank} = \{V_1, V_2, \dots, V_n\} \quad (5)$$

3.4.2 反向检索

在得到正向检索的视频排名结果之后,为了实

现反向组合式视频检索,首先需要对修改文本进行反转操作。具体来说,给定原始参考视频 V_r 、原始修改文本 T_q 和正向检索得到的候选视频 V_k ,通过提示多模态大模型生成将候选视频 V_k 还原至原始参考视频 V_r 的反向修改文本 T_q' 。反向修改文本的生成基于原始参考视频描述、原始修改指令和候选视频描述,确保语义的准确反转。

例如,当原始修改文本为“将白色花朵变成粉色”时,反向修改文本应为“将粉色花朵变成白色”。这种反转操作能够有效验证候选视频与原始参考视频之间的语义一致性。

形式化地,反向修改文本 T_q' 的生成过程为:

$$T_q' = \Theta(T_r, T_q, T_k) \quad (6)$$

其中, Θ 为生成反向修改指令的大模型, T_r 为原始参考视频描述, T_k 为候选视频描述。

接着选择正向检索的 Top N 个视频分别作为候选视频,根据反向修改指令生成对应的反向检索视频描述。形式化地,对于每一个正向检索排名中的视频 V_k ,生成反向检索视频描述 T_c^k :

$$T_c^k = \Gamma(V_k, T_q') \quad (7)$$

其中, Γ 为基于反向修改文本生成视频描述的多模态大模型。

然后,使用 LanguageBind 文本编码器 Φ_t 提取反向检索视频描述 T_c^k 的文本特征 h_c^k 如下:

$$h_c^k = \frac{\Phi_t(T_c^k)}{\|\Phi_t(T_c^k)\|_2} \quad (8)$$

接着,计算每一个反向检索视频描述 T_c^k 与原始参考视频 V_r 的相似度。形式化地,计算相似度 S' 为:

$$s_k' = h_c^k \cdot (h_r)^T \quad (9)$$

$$S' = \{s_1', s_2', \dots, s_N'\} \quad (10)$$

其中, h_r 为原始参考视频 V_r 的特征表示。

通过反向检索机制,我们能够验证正向检索结果的语义一致性:如果候选视频能够通过反向修改文本准确还原到原始参考视频,说明该候选视频在语义上更符合用户的检索意图,应当获得更高的排名权重。

3.4.3 基于双向检索一致性的重排

基于双向检索一致性原理,如果正向检索得到的视频能够通过反向修改文本检索到参考视频,那么该视频为真实的目标视频的可能性越高,相应地需要将其原始的排名上调。为了平衡正向检索和反向检索对最终检索排名的影响,通过设置权重因子

λ , 对正向检索的 Top N 视频的相似度进行调整,具体如下:

$$S'' = S + \lambda S' \quad (11)$$

基于调整之后的相似度 S'' 对正向检索得到的 Top N 视频进行重新排序,得到调整之后的视频排名结果为:

$$Rank' = \{V_1', V_2', \dots, V_n'\} \quad (12)$$

其中, V_1' 为最终检索的目标视频。

4 实验

4.1 数据集

WebVid-CoVR-Test^[1]是一个用于组合式视频检索的评估基准,包含 2556 个测试样本。每个样本由一个参考视频、一个修改文本以及一个目标视频三元组组成,均来源于大规模视频-文本数据集 Webvid10M^[40]。该基准通过挖掘具有相似描述的视频对,并利用大型语言模型生成相应的修改文本,研究人员对这些文本进行人工筛选和标注,以确保测试样本的高质量与高可靠性。

EgoCVR^[2]是一个新的组合式视频检索基准,其视频与相关标注来源于 Ego4D^[41]中的手部与物体预测任务。该基准包含 2295 个测试样本,每个样本的参考视频与目标视频均采自同一段长视频。对于每个查询样本,平均包含 1.2 个目标视频。为提升检索难度,研究人员为每个已标注的目标视频收集了具有高视觉相似性的干扰视频片段,从而避免模型仅依据视觉相似性完成检索任务。

4.2 实现细节

在思维链描述生成模块中,本文使用 GPT-4o^[42] 结合上下文示例提示,通过逐步完成预定义的子任务,推理得到最终的目标视频描述。在双向检索模块,选择正向检索的 Top K 的视频作为参考视频,使用 GPT-4o^[42] 来生成反向修改文本以及反向检索的视频描述。在视觉-语言模型方面,本文选择了 LanguageBind^[43] 模型,使用其视频编码器提取视频特征,文本编码器提取文本特征。特征空间维度设为 768,并采用余弦相似度进行视频检索。对于 WebVid-CoVR-Test 数据集,双向检索权重因子 λ 设置为 0.5。对于 EgoCVR 数据集, λ 设置为 0.2。所有模型均基于 PyTorch 框架实现,并在单张 A800 GPU 上进行实验。

4.3 结果和分析

4.3.1 定量分析

为全面评估所提方法的性能,本文选取了多种先进的组合式视频检索方法进行对比实验。具体而言,本文评估了两种基于训练的组合式视频检索框架:BLIP-CoVR^[1]和BLIP-CoVR-ECDE^[3],这两种方法均基于BLIP模型,通过视觉与文本编码器之间的交叉注意力机制实现模态对齐。以及FDCA^[4]在句子和词汇层面上解构修改文本特征,执行跨模态特征对齐与融合,以更精确地捕捉用户需求。此外,本文还比较了几种零样本组合式视频检索方法:Visual-only模型仅使用参考视频的视觉特征进行相似度计算;Text-only模型则仅基于修改文本的特征进行相似度计算。CIReVL^[44]是一种无需训练的方法,通过预训练的生成式视觉语言模型并结合大型语言模型,根据修改文本重新组合描述来完成检索任务。TFR-

CVR^[2]则采用了两阶段的零样本组合式视频检索框架:首先基于参考视频与数据库视频的视觉特征相似度检索出Top N候选集,随后根据修改文本进行文本引导的描述编辑,并对候选集进行再排序。

在实验结果分析中,本文提出的BC4CVR方法在两个具有代表性的数据集上均展现出优异的性能。如表1和表2所示,在WebVid-CoVR-Test数据集上,BC4CVR在R@1、R@5和R@10指标上分别达到了53.40%、76.52%和82.39%,相比于现有的零样本方法如TFR-CVR均有不同程度的提升。而在更具挑战性的EgoCVR数据集上,BC4CVR的优势更加明显,在R@1指标上提升至12.24%,相比TFR-CVR的8.38%(TFR-CVR的指标为LanguageBind作为视觉语言模型复现的结果)提升了约46%,同时在R@5和R@10上也保持领先,分别达到38.03%和57.25%。

表 1 WebVid-CoVR-Test 数据集实验结果

方法	零样本	输入	融合方法	VLM	召回率@K(%)		
					R@1	R@5	R@10
BLIP-CoVR ^[1]	×	V+T	CA	BLIP	55.95	81.22	89.05
BLIP-CoVR-ECDE ^[3]	×	V+T	CA	BLIP	60.12	84.32	91.27
FDCA ^[4]	×	V+T	CA	BLIP	54.80	82.27	89.94
Random	✓	-	-	BLIP	0.08	0.23	0.35
Text-only	✓	T	-	BLIP	19.68	37.09	45.85
Visual-only	✓	V	-	BLIP	34.90	59.23	68.04
CIReVL ^[44]	✓	V+T	Caption	LanguageBind	44.09	67.18	74.57
OSrCIR ^[38]	✓	V+T	Caption	LanguageBind	48.55	70.34	78.48
TFR-CVR ^[2]	✓	V+T	Caption	LanguageBind	51.70	75.30	80.70
BC4CVR(Ours)	✓	V+T	Caption	LanguageBind	53.40	76.52	82.39

表 2 EgoCVR 数据集实验结果

方法	零样本	输入	融合方法	VLM	召回率@K(%)		
					R@1	R@5	R@10
BLIP-CoVR ^[1]	×	V+T	CA	BLIP	5.40	15.20	24.30
BLIP-CoVR-ECDE ^[3]	×	V+T	CA	BLIP	6.00	16.30	24.80
FDCA ^[4]	×	V+T	CA	BLIP	8.70	19.20	27.10
Random	✓	-	-	BLIP	0.01	0.05	0.10
Text-only	✓	T	-	BLIP	0.40	1.40	2.70
Visual-only	✓	V	-	BLIP	6.50	32.60	55.30
CIReVL ^[44]	✓	V+T	Caption	LanguageBind	8.19	31.33	53.03
OSrCIR ^[38]	✓	V+T	Caption	LanguageBind	7.10	31.81	54.86
TFR-CVR ^[2]	✓	V+T	Caption	LanguageBind	8.38	34.76	55.59
BC4CVR(Ours)	✓	V+T	Caption	LanguageBind	12.24	38.03	57.25

此外,BC4CVR在EgoCVR数据集上的表现均超越了基于监督学习的方法(如BLIP-CoVR和BLIP-CoVR-ECDE),这一结果表明,BC4CVR能够更好地适应高相似干扰样本带来的复杂语义分辨任务,具备更强的泛化能力和语义辨析能力。整体来看,BC4CVR所融合的思维链推理机制增强了对用户检索意图的细粒度建模能力,而双向一致性检索策略则有效缓解了语义偏差对最终排序的干扰。这些创新模块协同提升了检索的准确性与鲁棒性,展示了本方法在零样本组合式视频检索任务中的先进性与实用性。

4.3.2 定性分析

如图3,我们展示了正向检索过程中可能出现的一些问题及其对检索结果的影响。在图3(a)中,正向检索时忽视了视频中花朵数量以及环境与参考图像的不一致,导致检索结果中的视频虽然在某些视觉特征上符合修改指令(如颜色变更),但在其他关键特征上存在偏差(例如花朵数量不一致)。这一

问题表明,正向检索过于关注整体的颜色或外观特征,而未能细致考虑视频中其他细节的语义一致性,导致最终的检索结果可能不符合用户的期望。图3(b)则展示了另外一种情况:正向检索时虽然有效地将场景调整为黄昏,但忽略了视频的颗粒度和其他重要细节。尽管场景色调发生了变化,但视频整体的语义结构和背景元素却未能与修改文本完全对接,导致正向检索时,生成的描述与原始参考视频之间的相似度降低,从而影响了检索的准确度。基于双向检索一致性原理,我们的方法通过引入反向检索机制解决了上述问题。如果正向检索得到的视频能够通过反向修改文本成功检索回参考视频,说明该视频在语义上更为一致,因此应当上调其排名。这一双向检索一致性原理有效地弥补了正向检索过程中对视频细节不足的考虑,提高了检索结果的准确度和鲁棒性,确保了检索过程中没有因语义偏差而产生不符合用户意图的错误结果。

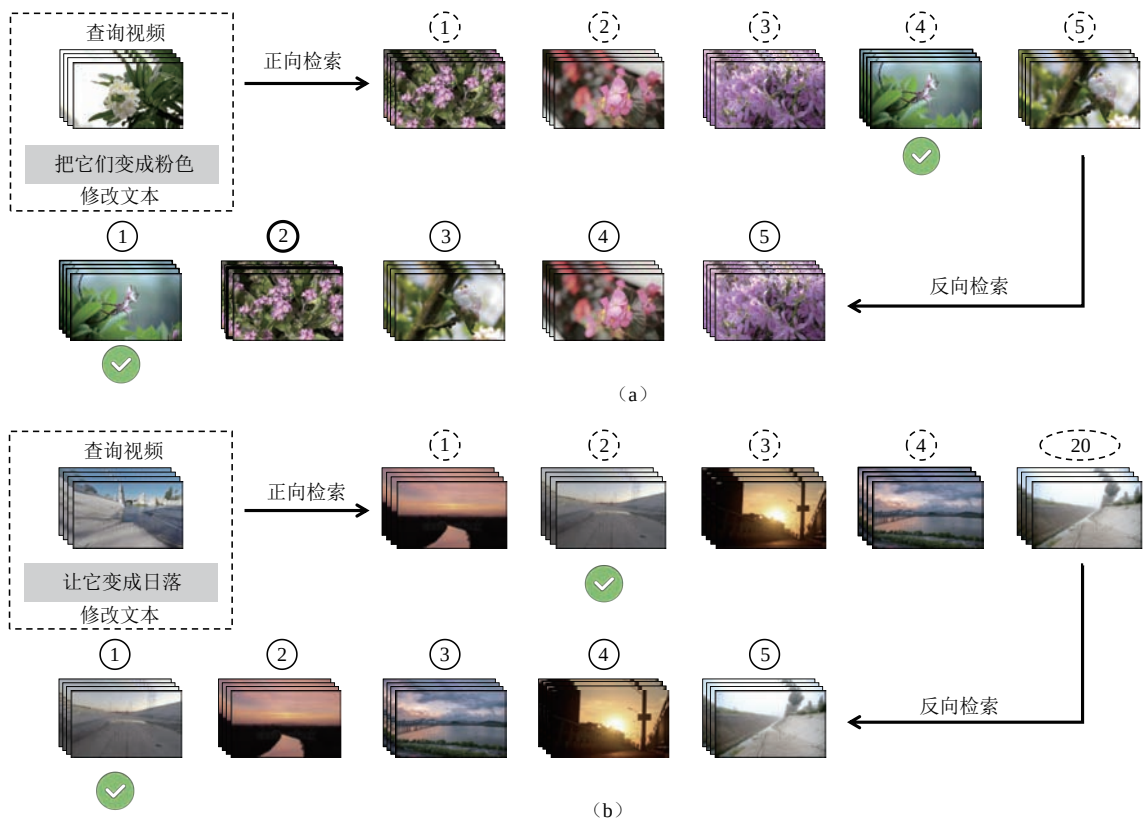


图3 本文方法在组合式视频检索任务中的定性分析示例

4.4 消融实验

4.4.1 不同模块的影响

为验证思维链推理模块与双向检索模块的有效性,本文设计了以下消融实验配置:(1)Base:使用现

有的两阶段方法生成目标视频描述进行文本-视频检索;(2)Chain-of-Thought(CoT):使用本文提出的思维链推理模块生成目标视频描述进行正向相似度检索;(3)Full Method(CoT+Rev):使用思维链

推理模块生成目标视频描述,首先进行正向检索,然后通过双向检索模块对视频排名进行调整,得到最终的检索结果。

如表3所示,在WebVid-CoVR-Test数据集上的实验结果分析如下:CoT配置在R@1指标上达到51.87%。这表明通过本文构建的思维链推理模块生成的目标视频描述,与用户检索意图更加对齐,从而有效提升检索的准确性。CoT+Rev配置在R@1指标上达到53.40%。这证明了通过双向验证机制能够识别和抑制硬负样本、目标视频描述中的语义偏差对检索结果的影响,从而进一步提升检索精度。这一结果强有力地说明:(1)通过引入思维链生成的目标视频描述与用户的检索意图高度对齐;(2)反向验证模块能有效利用双向检索的一致性,消除目标视频描述中有偏差语义内容对检索结果的影响。

表3 WebVid-CoVR-Test数据集上的模块消融实验

配置	召回率@K(%)		
	R@1	R@5	R@10
Base	47.81	71.09	79.03
CoT	51.87	73.51	80.86
CoT + Rev	53.40	76.52	82.39

4.4.2 权重因子 λ 的影响

在WebVid-CoVR-Test数据集上,本文系统分析了双向检索机制中权重因子 λ 对检索性能的影响,实验结果如图4所示。该权重因子用于调控正向相似度得分 $S_{\text{正向}}$ 与反向验证得分 $S_{\text{反向}}$ 在最终匹配评分中的相对权重,其整体得分形式如下: $S=S_{\text{正向}}+\lambda S_{\text{反向}}$ 。从图中可以观察到,随着 λ 从0.1增大至1.0,召回率在不同指标(R@1, R@5, R@10)上呈现出一定的变化趋势,表明 λ 的设定对检索性能具有显著影响。具体而言, λ 设置过大时,模型在决策过程中对反向验证信号的依赖增强,可能导致对语义上高度相关但在反向方向存在轻微偏差的样本施加过高惩罚,进而损害整体召回效果。此外,反向验证模块本身存在一定噪声,当其影响被放大时,容易掩盖正向检索中的强匹配信号,导致性能下降。

相反,当 λ 设置过小时,反向验证信号的作用被弱化,模型难以有效识别和抑制与查询存在潜在语义冲突的硬负样本,从而削弱对高相似度干扰项的判别能力,影响整体检索表现。因此,合理设置 λ 以实现正向与反向信号的有效协同,是提升模型性能的关键因素。综合实验结果显示,对于WebVid-

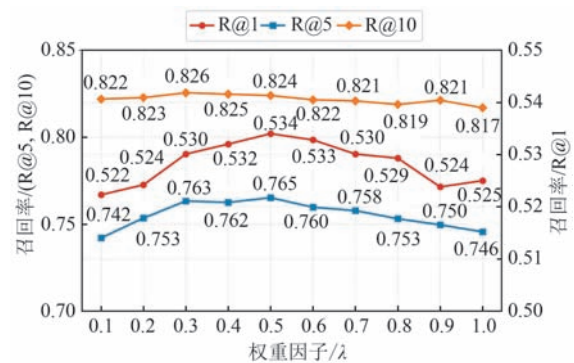


图4 权重因子 λ 对检索性能的影响

CoVR-Test数据集 $\lambda=0.5$ 附近可获得较优的检索性能平衡。

5 总 结

本文提出了基于思维链推理的双向组合式视频检索方法(BC4CVR),旨在解决现有方法中对参考视频视觉信息的丢失和对用户检索意图理解不足的问题。BC4CVR通过引入思维链推理,将生成目标视频描述的任务分解为多个子任务,逐步推理生成更精准的视频描述,确保检索结果更好地反映用户意图。同时,双向检索机制通过反向检索验证正向检索结果,并对检索排名进行重排序,有效消除语义偏差,提高检索精度。实验结果表明,BC4CVR在WebVid-CoVR-Test和EgoCVR数据集上显著优于现有方法,证明了其在零样本组合式视频检索任务中的有效性。本文的贡献在于提出了无训练的BC4CVR方法,突破了对大量标注数据的依赖,通过思维链推理和双向检索机制优化了检索精度和鲁棒性,为未来的组合式视频检索研究提供了新的思路。尽管BC4CVR在零样本组合式视频检索中表现优秀,但其依赖高计算成本的大模型、处理复杂视频内容时的描述不精确以及双向检索机制带来的额外开销仍是挑战,未来将探索更高效的描述生成方法、细粒度视频理解技术及扩展至更多应用场景。

致 谢 在此我们向对本文的工作给予支持和宝贵建议的评审老师表示衷心的感谢!

参 考 文 献

- [1] Ventura L, Yang A, Schmid C, et al. CoVR: Learning composed video retrieval from web video captions//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver,

- Canada, 2024: 5270-5279
- [2] Hummel T, Karthik S, Georgescu M I, et al. EgoCVR: An egocentric benchmark for fine-grained composed video retrieval// Proceedings of the Computer Vision-European Conference on Computer Vision. Milan, Italy, 2024: 1-17
- [3] Thawakar O, Naseer M, Anwer R M, et al. Composed video retrieval via enriched context and discriminative embeddings// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 26896-26906
- [4] Yue W, Qi Z, Wu Y, et al. Learning fine-grained representations through textual token disentanglement in composed video retrieval// Proceedings of the Thirteenth International Conference on Learning Representations. Singapore, 2025
- [5] Song X, Lin H, Wen H, et al. A comprehensive survey on composed image retrieval. *ACM Transactions on Information Systems*, 2025, 44(1): 1-54
- [6] Wan Y, Zou G, Zhang B. Composed image retrieval: A survey on recent research and development. *Applied Intelligence*, 2025, 55: 482
- [7] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 815-823
- [8] Vo N, Jiang L, Sun C, et al. Composing text and image for image retrieval: An empirical odyssey// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 6439-6448
- [9] Zhang S, Gong Y H, Wang J J. The development of deep convolution neural network and its applications on computer vision. *Chinese Journal of Computers*, 2019, 42(3): 453-482 (in Chinese)
(张顺, 龚怡宏, 王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用. *计算机学报*, 2019, 42(3): 453-482)
- [10] Liu Z, Rodriguez-Opazo C, Teney D, et al. Image retrieval on real-life images with pre-trained vision-and-language models// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 2125-2134
- [11] Baldrati A, Bertini M, Uricchio T, et al. Effective conditioned and composed image retrieval combining clip-based features// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 21466-21474
- [12] Delmas G, de Rezende R S, Csürka G, et al. ARTEMIS: Attention-based retrieval with text-explicit matching and implicit similarity// Proceedings of the 10th International Conference on Learning Representations. Virtual, 2022
- [13] Lin J A, Bao C Z, Dong J F, et al. Multilingual text-video cross-modal retrieval model via multilingual-visual common space learning. *Chinese Journal of Computers*, 2024, 47(9): 2195-2210 (in Chinese)
(林俊安, 包翠竹, 董建峰, 等. 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. *计算机学报*, 2024, 47(9): 2195-2210)
- [14] Chen X T, Wei S K, Zhang G J, et al. Combination image retrieval based on adaptive and cascade structures. *Journal of Beijing Jiaotong University*, 2022, 46(5): 42-49 (in Chinese)
(陈小彤, 韦世奎, 张港鉴, 等. 基于自适应和级联结构的组合图像检索. *北京交通大学学报*, 2022, 46(5): 42-49)
- [15] Baldrati A, Agnolucci L, Bertini M, et al. Zero-shot composed image retrieval with textual inversion// Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 15338-15347
- [16] Saito K, Sohn K, Zhang X, et al. Pic2Word: Mapping pictures to words for zero-shot composed image retrieval// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 19305-19314
- [17] Karthik S, Mancini M, Akata Z. KG-SP: Knowledge guided simple primitives for open world compositional zero-shot learning// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 9336-9345
- [18] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision// Proceedings of the International Conference on Machine Learning. Virtual, 2021: 8748-8763
- [19] Yang Z, Xue D, Qian S, et al. LDRE: LLM-based divergent reasoning and ensemble for zero-shot composed image retrieval// Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. Washington, USA, 2024: 80-90
- [20] Yang Z, Qian S, Xue D, et al. Semantic editing increment benefits zero-shot composed image retrieval// Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia, 2024: 1245-1254
- [21] Liu Y, Xiong P, Xu L, et al. TS2-Net: Token shift and selection transformer for text-video retrieval// Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 319-335
- [22] Luo H, Ji L, Zhong M, et al. CLIP4Clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 2022, 508: 293-304
- [23] Ma Y, Xu G, Sun X, et al. X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval// Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal, 2022: 638-647
- [24] Li J, Li D, Savarese S, et al. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models// Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA, 2023: 19730-19742
- [25] Rasheed H, Khattak M U, Maaz M, et al. Fine-tuned clip models are efficient video learners// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 6545-6554
- [26] Xu H, Ghosh G, Huang P Y, et al. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic, 2021: 6787-6800
- [27] Xue H, Sun Y, Liu B, et al. CLIP-ViP: Adapting pre-trained

- image-text model to video-language representation alignment// Proceedings of the Eleventh International Conference on Learning Representations. Kigali, Rwanda, 2023
- [28] Yao L, Huang R, Hou L, et al. FILIP: Fine-grained interactive language-image pre-training//Proceedings of the Tenth International Conference on Learning Representations. Virtual, 2022
- [29] Chu Z, Chen J, Chen Q, et al. Navigate through enigmatic labyrinth: A survey of chain of thought reasoning: Advances, frontiers and future//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand, 2024: 1173-1203
- [30] Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 2022, 35: 22199-22213
- [31] Rubin O, Herzig J, Berant J. Learning to retrieve prompts for in-context learning//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, USA, 2022: 2655-2671
- [32] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 2022, 35: 24824-24837
- [33] Zhang Z, Zhang A, Li M, et al. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023
- [34] Zheng G, Yang B, Tang J, et al. DDCOT: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 2023, 36: 5168-5191
- [35] Mitra C, Huang B, Darrell T, et al. Compositional chain-of-thought prompting for large multimodal models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 14420-14431
- [36] Sprague Z, Yin F, Rodriguez J D, et al. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning//Proceedings of the Thirteenth International Conference on Learning Representations. Singapore, 2025
- [37] Yu Z, He L, Wu Z, et al. Towards better chain-of-thought prompting strategies: A survey. *arXiv*, 2023, arXiv:2310.04959
- [38] Tang Y, Zhang J, Qin X, et al. Reason-before-retrieve: One-stage reflective chain-of-thoughts for training-free zero-shot composed image retrieval//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2025: 14400-14410
- [39] Sun Z, Jing D, Lu Z. CoTMR: Chain-of-thought multi-scale reasoning for training-free zero-shot composed image retrieval// Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu, USA, 2025: 22675-22684
- [40] Bain M, Nagrani A, Varol G, et al. Frozen in time: A joint video and image encoder for end-to-end retrieval//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 1728-1738
- [41] Grauman K, Westbury A, Byrne E, et al. Ego4D: Around the world in 3,000 hours of egocentric video//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 18995-19012
- [42] Hurst A, Lerer A, Goucher A P, et al. GPT-4o system card [EB/OL]. *arXiv*, 2024, arXiv:2410.21276
- [43] Zhu B, Lin B, Ning M, et al. LanguageBind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023
- [44] Karthik S, Roth K, Mancini M, et al. Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291*, 2023



MENG Yi-Ping, Ph. D. candidate. Her research interests include multimedia content analysis and cross-modal retrieval.

LIU Li-Long, M. S. candidate. His research interests include artificial intelligence and natural language processing.

DING Zhi-Xiang, M. S. candidate. His research interests include multimedia content analysis, data mining, and cross-modal retrieval.

YANG Zhen-Yu, Ph. D. candidate. His research interests include multimodal understanding and application.

QIAN Sheng-Sheng, Ph. D. professor. His research interests include multimedia content analysis, data mining, cross-modal retrieval, and personalized recommendation.

Background

The rapid advancement of pre-trained Vision-Language Models (VLMs) has significantly propelled the field of artificial intelligence, particularly at the intersection of computer vision

and natural language processing (NLP). These models, such as CLIP, excel in aligning image and text features, enabling them to perform effectively in a variety of cross-modal tasks,

including image-text retrieval, zero-shot classification, and visual captioning. By leveraging large-scale pretraining datasets, VLMs have achieved impressive results in tasks that require deep understanding and generation of multimodal content.

Despite their remarkable capabilities, VLMs face significant challenges in real-world applications, especially in tasks that require fine-grained understanding and manipulation of visual and textual information. One such task is Zero-Shot Composed Video Retrieval (ZS-CVR), which aims to precisely locate target videos from an unannotated candidate video library based on a reference video provided by the user, along with natural language modification instructions. For example, a user might request to find a video similar to a given reference but with specific modifications, such as “a similar video but with a snowy scene.” The core challenge of ZS-CVR lies in deeply understanding the complex user intent, particularly the semantic changes introduced by the modification instruction and their relationship with the content of the reference video. Traditional methods often convert the reference video into a single textual description, which is then combined with the modification text to generate an edited video description for retrieving the target video. However, this approach often leads to a loss of relevant visual information from the reference video and an inadequate understanding of the user’s retrieval intent. Moreover, the retrieval process relies solely on the similarity between the generated edited description and the candidate videos, which may result in retrieving negative samples with high visual similarity due to biased content in the edited description, thus affecting retrieval accuracy.

To address these limitations, recent research has explored

more sophisticated techniques to enhance the understanding and generation of video descriptions. These methods include leveraging pre-trained VLMs to generate more accurate video descriptions and using large language models (LLMs) to combine these descriptions with modification instructions more effectively. However, existing methods still struggle with accurately capturing the user’s intent and handling the semantic complexity of the task.

In this context, our research proposes a novel approach to ZS-CVR by introducing a Bidirectional Chain-of-Thought Reasoning for Composed Video Retrieval method (BC4CVR). This method leverages Chain-of-Thought (CoT) reasoning to guide a multimodal large model in generating a description of the target video by incrementally reasoning through predefined sub-tasks based on the query video and modification text. This facilitates a fine-grained understanding of the user’s retrieval intent. Additionally, a bidirectional retrieval mechanism is introduced to validate the consistency of the retrieval results, thereby improving retrieval accuracy and mitigating the semantic bias introduced by the generated target video descriptions. Our proposed method aims to address the limitations of existing ZS-CVR techniques by providing a more accurate and robust solution that better aligns with user intent and reduces the impact of semantic biases. Through extensive experiments on major benchmark datasets, we demonstrate that BC4CVR achieves state-of-the-art performance in the ZS-CVR task, highlighting its potential for practical applications in video retrieval and related fields.

This work was supported by the National Key Research and Development Program of China (2023YFC3310700), and the National Natural Science Foundation of China (No. 62276257).