Journal of Image and Signal Processing 图像与信号处理, 2025, 14(3), 349-361
Published Online July 2025 in Hans. https://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2025.143032

Hans汉斯

基于视频文本对齐的视频检索模型

张　宇，张天保

合肥工业大学计算机与信息学院，安徽　合肥

收稿日期：2025年6月11日；录用日期：2025年7月2日；发布日期：2025年7月15日

摘　要

针对文本－视频检索遇到的全局对齐方法缺乏细粒度语义匹配以及跨模态语义鸿沟导致特征对齐困难的问题，提出一种高效全局－局部序列对齐方法(ETVA)。模型由文本编码器、视频编码器、文本－视频全局对齐模块和文本－视频细粒度对齐模块构成。其中文本编码器采用ALBERT模型，凭借其双向编码能力精准提取文本特征，能够提升跨模态特征的时序一致性与语义关联性。视频编码器利用多专家模块策略，从多模态、多特征角度全面捕捉视频信息。全局对齐模块通过聚合和变换特征，有效实现全局语义对齐；细粒度对齐模块基于共享聚类中心机制，深入挖掘文本和视频局部细节的语义关联。在实验中采用MSRVTT、ActivityNet Captions和LSMDC数据集，评价指标采用Recall@K和Median Rank，结果表明ETVA在不同数据集上均表现较好，在检索准确性相比其他方法有所提升。

关键词

多模态数据对齐，深度学习，视频理解，视频检索

Video Retrieval Model Based on Video Text Alignment

Yu Zhang, Tianbao Zhang

School of Computer and Information Science, Hefei University of Technology, Hefei Anhui

Received: Jun. 11th, 2025; accepted: Jul. 2nd, 2025; published: Jul. 15th, 2025

Abstract

An efficient global local sequence alignment method (ETVA) is proposed to address the problem of global alignment methods lacking fine-grained semantic matching and cross modal semantic gaps leading to difficulty in feature alignment in text video retrieval. The model consists of a text encoder, a video encoder, a text video global alignment module, and a text video fine-grained alignment module. The text encoder adopts the ALBERT model, which accurately extracts text features with its

文章引用: 张宇, 张天保. 基于视频文本对齐的视频检索模型[J]. 图像与信号处理, 2025, 14(3): 349-361.
DOI: 10.12677/jisp.2025.143032

bidirectional encoding ability, and can improve the temporal consistency and semantic correlation of cross modal features. The video encoder utilizes a multi expert module strategy to comprehensively capture video information from multiple modalities and feature perspectives. The global alignment module effectively achieves global semantic alignment by aggregating and transforming features; The fine-grained alignment module is based on a shared clustering center mechanism to deeply explore the semantic associations between local details in text and video. In the experiment, MSRVTT, ActiveNet Captions, and LSMDC datasets were used, and the evaluation indicators were Recall@K Compared with Median Rank, the results show that ETVA performs well on different datasets and has improved retrieval accuracy compared to other methods.

Keywords

Multimodal Data Alignment, Deep Learning, Video Comprehension, Video Retrieval

1. 引言

在信息传播形式日益多元的当下，视频凭借其丰富的多模态内容(如视觉画面、音频信息等)和动态的时间序列，成为承载海量信息的关键媒介。文本–视频检索作为连接用户自然语言描述与目标视频的桥梁，在众多领域有着广泛的应用前景，如智能视频搜索引擎帮助用户基于复杂文本描述精准定位所需视频，在线教育平台依此实现视频课程的高效推荐等。然而，实现精准的文本–视频检索[1]-[3]并非易事，其核心挑战在于如何在联合嵌入空间中精确衡量文本与视频之间的相似度。回顾现有研究，多数方法侧重于从全局视角学习文本和视频的表征并进行相似性比较，这种方式虽在一定程度上取得了成果，但却忽视了文本和视频中蕴含的丰富局部细节，使得检索结果难以满足复杂场景下的精细化需求。

针对上述问题，本文提出了一种名为 ETVA (Efficient Text-video Aligner)的高效全局–局部序列对齐方法。ETVA 是基于 ALBERT 文本编码器和专家模块的视频编码器相结合的框架，并结合全局和局部对齐策略对其进行优化。首先，文本模块采用了预训练的 ALBERT 模型。ALBERT 模型通过多层次的自注意力机制，能够在不同的语义层次上理解文本，为视频内容与文本描述之间的匹配提供了坚实的基础。在视频编码方面，采用了专家模块的设计策略，充分利用视频中的多模态信息。为了实现更精确的文本–视频匹配，本章引入了全局对齐和局部对齐两个关键模块。在全局对齐模块中，文本与视频的全局特征通过特征聚合与变换操作进行对齐，确保在语义层面上的整体匹配。而在局部对齐模块中，则通过细粒度的特征对齐，进一步优化文本与视频之间的局部细节匹配，从而捕捉更加精细的语义关联。最后，实验结果表明所提出的 ETVA 方法取得了一定效果。

2. 模型概述

如图 1 展示了本方法的全部流程。ETVA 模型主要将文本–视频对编码至联合特征空间，以此衡量二者的相似性，其核心设计理念聚焦于全局–局部对齐，该模型主要由文本编码器、视频编码器、文本–视频全局对齐模块和文本–视频细粒度对齐模块构成。在文本编码器中，选用预训练的 ALBERT 模型，通过对输入文本进行分词、填充并添加特定标记后输入模型，在训练过程中，ALBERT 模型与其他模块以端到端的方式优化，为后续的全局和局部对齐提供准确的文本特征表示，是实现全局–局部对齐的文本信息基础。

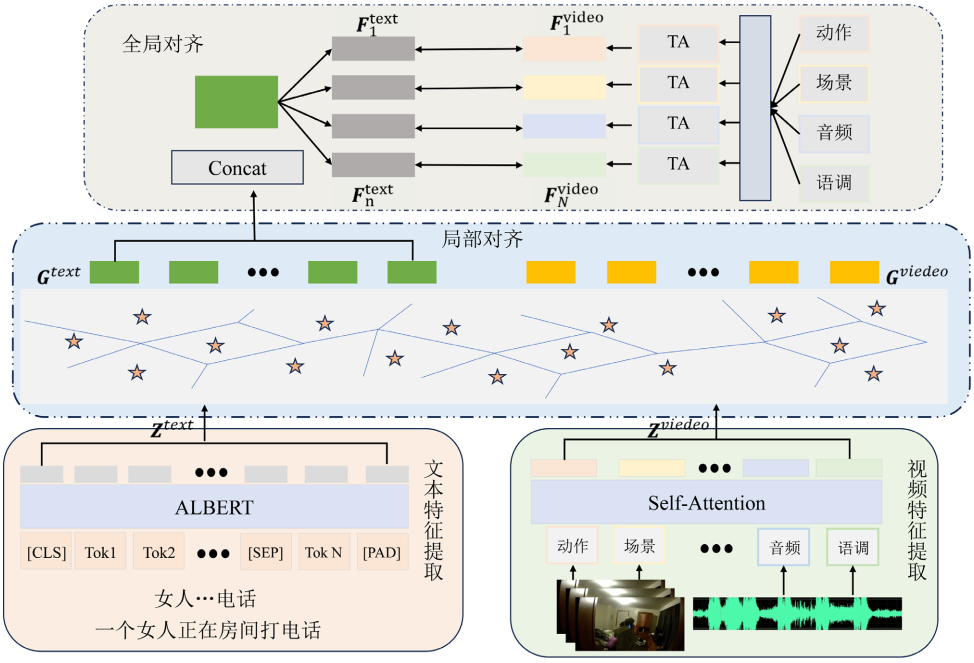

Figure 1. ETVA framework model
图 1. ETVA 框架模型

由于视频包含运动、音频和语音等丰富信息，本模型利用多个专家模块对原始视频编码。从视频每个时间片段提取特征后，通过最大池化操作生成全局特征并经自门控机制增强，同时利用自注意力层融合多专家特征得到局部特征。在全局对齐模块中通过独立聚合和变换每个视频特征，利用局部文本特征拼接生成特定于专家的全局文本表示，计算全局视频特征与相应文本特征之间余弦距离的加权和，实现特征的全局语义对齐。细粒度对齐模块中基于 NetVLAD 操作，模型学习共享聚类中心，通过点积计算局部特征与聚类中心的相似度来分配特征，进而计算聚合残差特征，最终得到文本和视频的局部特征，并利用余弦距离衡量它们之间的局部相似度。这一模块实现了文本和视频在局部细节上的精准对齐，捕捉更细致的语义关联。

通过这些模块的共同协作，ETVA 模型充分贯彻了全局-局部对齐的设计理念，在文本－视频检索任务中展现出强大的性能。

2.1. 文本编码模块

在 ETVA 模型中，文本特征的精准提取对后续实现文本与视频的高效对齐起着关键作用。为此，选用预训练的 ALBERT 模型来提取文本特征，其独特的架构和预训练机制使其在文本特征提取方面展现出卓越的性能。ALBERT 模型基于 Transformer 架构构建，核心在于其双向编码器设计。与传统的单向语言模型不同，ALBERT 能够同时从正向和反向两个方向对文本进行编码，从而全面捕捉文本中词汇的上下文信息。这一特性对于理解文本语义至关重要，因为一个词汇的含义往往受到其前后文的共同影响。

在将文本输入 ALBERT 模型之前，需要进行一系列预处理操作。首先，对输入文本进行分词，将其拆分为一个个词汇单元。经过预处理的文本被输入到 ALBERT 模型中。ALBERT 模型包含多个 Transformer 块，每个块由多头自注意力机制和前馈神经网络组成。在多头自注意力机制中，输入文本的每个词汇会与其他所有词汇进行交互，通过计算不同头的注意力分数，从多个角度捕捉词汇间的关系。例如，在一个句子中，“苹果”这个词汇不仅会关注自身，还会通过注意力机制关注到“水果”“红色”

等相关词汇，从而更好地理解其语义。前馈神经网络则对自注意力机制输出的结果进行进一步处理和特征变换。经过多个 Transformer 块的层层编码，文本中的每个词汇都能获取到丰富的上下文信息，最终得到文本的特征表示。

ALBERT 模型输出的特征表示为一系列词嵌入向量，记为：

$$\boldsymbol{Z}^{\text{text}} = \Phi^{\text{ALBERT}}\left(S\right), \tag{1}$$

其中 Φ^{ALBERT} 为 ALBERT 模型，S 为经过预处理的输入标记序列。$\boldsymbol{Z}^{\text{text}}$ 中的每个向量 $\boldsymbol{z}_i^{\text{text}}$ 对应文本中的一个词汇，这些向量包含了该词汇在整个文本语境中的语义信息，举个例子，对于词汇“读书”，其对应的向量 $\boldsymbol{z}_i^{\text{text}}$ 不仅编码了“读书”本身的含义，还包含了句子中一些其他词汇所传达的相关信息，像是“书本”“翻阅”等词语。

在 ETVA 模型的训练过程中，ALBERT 模型与其他模块以端到端的方式进行优化，这意味着 ALBERT 模型能够根据文本－视频对齐任务的需求，不断调整自身参数，进一步提升文本特征提取的准确性，为后续的全局和细粒度对齐操作提供高质量的文本特征输入。

2.2. 视频编码模块

视频编码模块是 ETVA 模型的核心组件，负责从原始视频中提取多模态特征，以支持文本－视频对齐任务。考虑到视频数据包含视觉、音频和语音等复杂信息，本模型采用多专家模块策略，针对不同模态进行特征提取。多专家模块的引入旨在充分挖掘视频数据中不同模态的信息。每个专家专注于特定的模态或特征类型，通过在相关任务上的预训练，积累了对该模态的专业知识。这种分工协作的方式类似于一个由不同领域专家组成的团队，能够从多个角度对视频进行分析，从而更全面地捕捉视频中的关键信息。例如，某些专家擅长处理视频中的视觉运动信息，而另一些专家则在音频特征提取方面表现出色。通过整合这些专家的输出，视频编码器能够获得更丰富、更具代表性的视频特征表示。

对于给定的视频，首先将其划分为多个时间片段。针对每个时间片段，利用 N 个专家 $\boldsymbol{E}^1, \boldsymbol{E}^2, \cdots, \boldsymbol{E}^N$ 分别提取特征。例如，专家 E^1 可能专注于提取视频帧中的视觉外观特征，通过卷积神经网络对每个时间片段的视频帧进行处理，得到片段级的视觉特征表示 $\boldsymbol{E}^1\left(x_1\right), \boldsymbol{E}^1\left(x_2\right), \cdots, \boldsymbol{E}^1\left(x_T\right)$，其中 x_t 表示第 t 个时间片段，T 为视频划分的总片段数。同理，其他专家也会针对各自擅长的模态或特征类型，从每个时间片段中提取相应的特征。

为了获取视频的全局特征，对每个专家提取的片段级特征进行时间聚合。具体而言，采用最大池化操作对每个专家的特征进行处理。最大池化操作能够在时间维度上选取每个特征通道中的最大值，从而突出视频在不同时刻的关键特征。例如，对于专家 $\boldsymbol{E}^N$ 提取的特征 $\boldsymbol{E}^n\left(x_1\right), \boldsymbol{E}^n\left(x_2\right), \cdots, \boldsymbol{E}^n\left(x_T\right)$ 经过最大池化后，得到一个全局特征向量。这个全局特征向量代表了该专家在整个视频中提取到的最重要信息。随后，通过自门控机制对这些全局特征进行增强。自门控机制可以自适应地调整每个特征的权重，突出对视频理解更为关键的特征，抑制噪声或无关信息。经过自门控机制处理后，得到一组全局专家特征 $\boldsymbol{F}_1^{\text{video}}, \boldsymbol{F}_2^{\text{video}}, \cdots, \boldsymbol{F}_N^{\text{video}}$。这些全局专家特征从不同专家的角度反映了视频的整体特征，为后续的全局对齐提供了重要的数据支持。

除了生成全局特征，视频编码模块还需要获取视频的局部特征，以支持细粒度对齐任务。为此，利用一个自注意力层来融合多专家的特征。首先，通过全连接层将不同专家的特征投影到 M 维嵌入空间，使得不同专家的特征在同一维度空间中具有可比性。然后，将所有专家投影后的特征进行拼接，形成一个包含多模态信息的特征序列。接下来，通过自注意力机制对这个特征序列进行处理。自注意力机制能够让每个位置的特征与其他所有位置的特征进行交互，从而探索多模态特征之间的关系。例如，在这个

特征序列中，视觉特征和音频特征可以通过自注意力机制相互影响，捕捉到它们在时间和语义上的关联。经过自注意力层处理后，得到局部特征：

$$\boldsymbol{Z}^{\text{video}}=\left\{\boldsymbol{z}_1^{\text{video}},\boldsymbol{z}_2^{\text{video}},\cdots,\boldsymbol{z}_P^{\text{video}}\right\}, \tag{2}$$

其中 P 为局部特征的数量。该局部特征融合方式不仅能够整合多专家提取的信息，还能够通过自注意力机制挖掘多模态特征之间的潜在关系，为视频的细粒度对齐提供了丰富且准确的局部特征表示。

通过上述多专家模块设计和特征提取流程，视频编码器能够从原始视频数据中提取出全面、丰富的全局和局部特征，为 ETVA 模型在文本－视频对齐任务中的高效运行奠定了坚实基础。

2.3. 文本－视频全局对齐模块

在 ETVA 模型里，文本－视频全局对齐模块对于实现文本与视频的全局语义对齐起着关键作用，它通过一系列对特征的聚合与变换操作达成这一目标。

视频编码器利用多个专家模块对原始视频进行编码。从每个时间片段提取特征后，得到片段级视频表示 $\boldsymbol{E}^n\left(x_1\right),\boldsymbol{E}^n\left(x_2\right),\cdots,\boldsymbol{E}^n\left(x_T\right)$，这里的 n 代表不同的专家，x_T 是第 T 个时间片段，T 为总的时间片段数。每个专家专注于不同模态或特征类型，如有的专家擅长提取视觉运动特征，有的对音频特征敏感。

对每个专家的片段级特征进行时间聚合，采用最大池化操作。这一过程突出了视频在不同时刻的关键特征，得到初步的全局专家特征。为进一步优化全局专家特征，引入自门控机制。自门控机制能够自适应地调整每个特征的权重。在实际视频中，并非所有特征对理解视频内容都同等重要。例如在一个体育赛事视频中，运动员的关键动作特征权重应高于背景观众的一些微小动作特征权重。自门控机制通过学习，能够突出对视频理解更为关键的特征，抑制噪声或无关信息，最终得到增强后的全局专家特征 $\boldsymbol{F}_1^{\text{video}},\boldsymbol{F}_2^{\text{video}},\cdots,\boldsymbol{F}_N^{\text{video}}$。这些特征从不同专家的视角反映了视频的整体特性，为后续与文本特征的匹配提供了丰富的视频全局信息。

文本编码器采用预训练的 ALBERT 模型。将输入文本进行分词、填充并添加特殊标记“[CLS]”和“[SEP]”后输入 BERT 模型，ALBERT 模型能够有效提取上下文词嵌入，得到文本特征：

$$\boldsymbol{Z}^{\text{text}}=\left\{\boldsymbol{z}_1^{\text{text}},\boldsymbol{z}_2^{\text{text}},\cdots,\boldsymbol{z}_B^{\text{text}}\right\}, \tag{3}$$

其中 B 是序列长度。这些文本特征包含了词汇在整个文本语境中的语义信息。基于 ALBERT 提取的文本特征，利用局部文本特征 $\boldsymbol{G}^{text}$ 进行拼接，生成特定于专家的全局文本表示：

$$\boldsymbol{F}^{\text{text}}=\left\{\boldsymbol{F}_1^{\text{text}},\boldsymbol{F}_2^{\text{text}},\cdots,\boldsymbol{F}_N^{\text{text}}\right\}. \tag{4}$$

对于关注视频视觉场景的专家，在生成其对应的全局文本表示时，会更侧重于拼接与视觉描述相关的局部文本特征，像描述颜色、形状、物体等的词汇特征；对于关注音频的专家，对应的全局文本表示则会重点拼接与声音、音效、语音等相关的局部文本特征。通过这种方式，使得文本特征能够与视频的不同专家全局特征在语义上更好地对应。

在得到全局视频专家特征 $\boldsymbol{F}_1^{\text{video}},\boldsymbol{F}_2^{\text{video}},\cdots,\boldsymbol{F}_N^{\text{video}}$ 和特定于专家的全局文本表示 $\boldsymbol{F}_1^{\text{text}},\boldsymbol{F}_2^{\text{text}},\cdots,\boldsymbol{F}_N^{\text{text}}$ 后，计算每个全局视频专家特征与相应文本特征之间的余弦距离，即 $\text{dist}\left(\boldsymbol{F}_1^{\text{video}},\boldsymbol{F}_1^{\text{text}}\right)$。余弦距离能够衡量两个向量在方向上的相似程度，在文本－视频全局对齐中，通过计算余弦距离可以判断对应文本和视频特征在语义方向上的接近程度。

为了综合考虑不同专家特征匹配的重要性，对每个余弦距离结果进行加权。权重 w_i 通过对文本表示 $\boldsymbol{G}^{\text{text}}$ 进行线性投影和 softmax 归一化得到。线性投影可以将文本特征映射到合适的权重空间，softmax 归一化则将权重值转化为概率分布形式，使得所有权重之和为 1。例如，如果文本中对视频某一特定模态

(如视觉)的描述更为详细和关键，那么对应视觉专家的权重 w_i 会相对较大。最终通过加权和得到：

$$S_{\text{global}} = \sum_{i=1}^{N} w_i * \text{dist}\left(\boldsymbol{F}_i^{\text{text}}, \boldsymbol{F}_i^{\text{video}}\right), \tag{5}$$

S_{global} 即为全局相似度。当这个全局相似度越高时，表明文本和视频在全局语义上越接近，从而实现了文本－视频的全局语义对齐。全局对齐流程如图 2 所示。

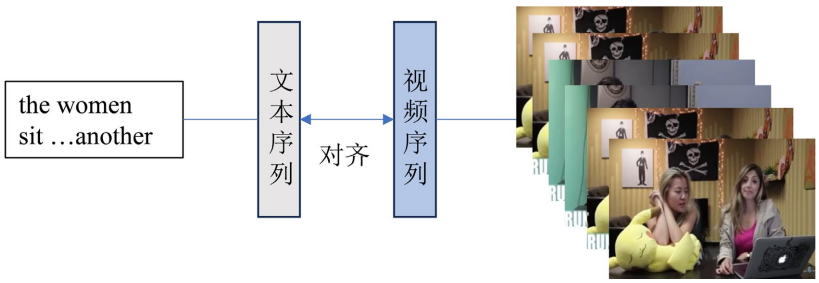

Figure 2. Global alignment process
图 2. 全局对齐流程

通过上述对视频和文本特征的聚合、变换以及相似度计算等一系列操作，文本－视频全局对齐模块成功实现了文本与视频在全局层面的语义对齐，为 ETVA 模型在文本-视频检索等任务中提供了重要的全局语义匹配依据。

2.4. 文本－视频细粒度对齐模块

在细粒度对齐模块对于精准捕捉文本和视频在局部细节上的语义关联至关重要。该模块基于共享聚类中心的机制，实现了文本与视频的细粒度对齐，下面将详细阐述其工作方式。图 3 展示了细粒度对齐方式。图中展示了跨模态数据在局部细节层面的匹配过程，右侧将视频序列通过卷积神经网络等模型分解为多个短时序片段并提取局部视觉特征，同时左侧将文本通过文本编码器拆分为词级嵌入并保留语序信息。

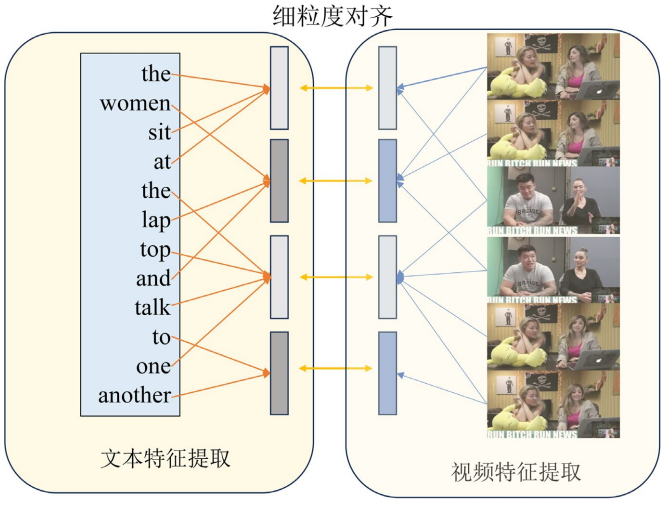

Figure 3. Fine-grained alignment process
图 3. 细粒度对齐流程

模型学习 $K+1$ 个 C 维共享聚类中心 $\{c_1, c_2, \cdots, c_K, c_{K+1}\}$。其中 K 个中心专门用于局部对齐，负责捕捉文本和视频局部特征中的相似模式；额外的一个中心 c_{K+1} 具有特殊作用，用于去除背景信息。在实际

的文本－视频场景中，文本描述和视频内容往往包含一些无关紧要的背景元素，这个额外的中心能够帮助模型识别并排除这些干扰，使对齐更加聚焦于关键信息。例如，在一段描述“公园里人们在喂鸽子”的文本和对应的视频中，公园的背景环境属于相对次要的信息，通过这个特殊的聚类中心，可以减少背景对关键对齐信息(如人物动作、鸽子等)的影响。对于局部视频特征 Z_i^{video}，通过点积计算其与每个聚类中心的相似度，以此确定该特征分配到各个聚类的概率。具体计算公式为：

$$a_{i,j}=\frac{\exp\left(\boldsymbol{z}_i^{\text{video}}\boldsymbol{c}_j^{\top}+b_j\right)}{\sum_{k=1}^{K+1}\exp\left(\boldsymbol{z}_i^{\text{video}}\boldsymbol{c}_k^{\top}+b_k\right)}, \tag{6}$$

其中 $\boldsymbol{b}_j$ 为可学习偏置项。这个公式利用了指数函数和归一化操作，使得概率值在 0 到 1 之间，且所有聚类的概率之和为 1。若局部视频特征 Z_i^{video} 在语义上更接近聚类中心 c_3，那么经过计算，分配到 c_3 的概率 $a_{i,3}$ 会相对较大。

以同样的方式对文本的局部特征进行处理。对于局部文本特征，通过与共享聚类中心进行点积计算，并结合可学习偏置项，得到其分配到各个聚类的概率。在确定了局部视频特征的分配概率后，计算每个聚类中心的聚合残差特征。对于第 j 个聚类中心，其聚合残差特征 g_j^{video} 的计算公式为：

$$\boldsymbol{g}_j^{\text{video}}=\text{normalize}\left(\sum_{i=1}^{M}a_{i,j}\left(\boldsymbol{z}_i^{\text{video}}-\boldsymbol{c}_j'\right)\right) \tag{7}$$

其中 $\boldsymbol{c}_j'$ 为与 $\boldsymbol{c}_j$ 大小相同的可训练权重，“normalize”表示 L2 归一化操作。该公式的含义是，将分配到第 J 个聚类中心的所有局部视频特征，减去该聚类中心对应的可训练权重 $\boldsymbol{c}_j'$，然后根据分配概率 $a_{i,j}$ 进行加权求和，最后进行 L2 归一化。通过这种方式，得到的聚合残差特征能够突出每个聚类中心所代表的局部特征与该中心本身的差异，从而更准确地反映视频局部特征的特点。同样地，对于文本也按照上述方式计算聚合残差特征 g_j^{text}。通过对文本局部特征的类似处理，得到与视频相对应的文本聚合残差特征。

最终，利用余弦距离来衡量文本和视频的局部特征之间的相似度，即：

$$s_{\text{local}}=\text{dist}\left(\boldsymbol{G}^{\text{video}},\boldsymbol{G}^{\text{text}}\right) \tag{8}$$

其中 $\boldsymbol{G}^{\text{video}}$ 和 $\boldsymbol{G}^{\text{text}}$ 分别是由各个聚类中心的聚合残差特征组成的文本和视频的局部特征表示。余弦距离能够有效衡量两个向量在方向上的相似程度，在这里用于判断文本和视频在细粒度上的语义相似性。当 s_{local} 的值越小时，说明文本和视频在局部细节上的语义越接近，从而实现了文本－视频的细粒度对齐。

通过基于共享聚类中心的一系列操作，ETVA 模型的细粒度对齐模块能够深入挖掘文本和视频在局部细节上的语义关联，为模型在文本－视频检索等任务中提供了更为准确的局部匹配能力，能够提升模型的整体性能。

3. 实验结果和分析

3.1. 实验环境设置

主要要使用 Ubuntu 18.04，CPU 使用了 Intel Xeon e5-2620 六核处理器，主频 2.4GHz，并基于 CUDA 10.2 环境在 Nvidia GTX 3060ti 显卡上进行训练。在本实验中，初始学习率设定为 0.0001。每经过 10 个 epoch，学习率衰减为原来的 0.9 倍，并设置了 50 个训练轮次 epoch，批处理大小为 32，模型使用 Adam 优化器。主要采用了 MSRVTT 数据集，ActivityNet Captions 数据集 LSMDC 数据集。评价指标采用 Recall@K(R@K)和 Median Rank(MdR)两个指标。

1) 数据集

a) MSRVTT (Microsoft Research Video to Text)数据集包含了大约 10,000 个视频，这些视频来源多样，

涵盖了日常生活、新闻、电影片段等多个场景。视频时长从几秒到几分钟不等，平均时长约为 9 秒。

b) ActivityNet Captions 数据集包含约 20,000 个未修剪的长视频，视频内容主要围绕人类的各种日常活动，如体育赛事、社交聚会、工作场景等。

c) LSMDC (Large-Scale Movie Description Challenge)数据集来源于电影片段。它包含了从大量电影中剪辑出的视频片段，以及对应的详细文本描述。

2) 评价指标

Recall@K 是文本－视频检索任务中常用的评价指标之一。其含义是在检索结果的前 K 个位置中，能够正确匹配到相关视频(即与查询文本语义相符的视频)的比例。计算公式为：

$$\text{Recall@K} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{9}$$

其中 TP 表示在 K 个检索视频中正确的数量，FN 表示在前 K 个结果中未被检索的数量。

MdR (Median Rank)衡量的是相关视频在整个检索结果列表中的中位排名。

3.2. 实验结果对比和分析

为了证明方法的有效性，通过对比 ETVA 与其他先进方法的实验结果，能直观展现出 ETVA 模型的具体效果。

在 MSRVTT 数据集上，ETVA 在文本到视频检索和视频到文本检索任务中，都取得了较好的结果。如表 1 所示。以 MPT 为例，在 1k-B 分割子集的文本到视频检索任务中，ETVA 的 R@1 指标比 MPT 高出 3.5% (52.7%对比 49.2%)；在视频到文本检索任务中，同样在 1k-B 分割子集上，R@1 指标提升了 4.3% (51.7%对比 47.4%)。即使与在大规模数据集上预训练的 T-MASS 相比，ETVA 在 1k-A 分割子集的所有指标上也有明显优势，如在文本到视频检索任务中，R@1 指标高出 1.1%。在推理时间上，ETVA 的视频编码模块(不包含专家编码)处理 1k 个视频仅需 0.4 秒，而 T-MASS 则需要 1.1 秒，体现了 ETVA 的高效性。

Table 1. Experimental results on the MSRVTT dataset
表 1. 在 MSRVTT 数据集上的实验结果

方法	分类	文本－视频				视频－文本			
		R@1	R@5	R@10	Mdr	R@1	R@5	R@10	MdR
CLIP4Clip [4]	1K-A	44.5	71.4	81.6	2.0	42.7	70.9	80.6	2.0
X-CLIP [5]	1K-A	45.1	72.3	82.3	2.0	46.8	72.5	84.2	2.0
X-Pool [6]	1K-A	46.9	72.8	83.2	2.0	48.2	73.3	84.3	2.0
MPT [7]	1K-A	46.3	70.9	80.7	-	45.0	70.9	80.6	-
T-MASS [8]	1K-A	50.2	75.1	85.2	1.0	47.7	78.2	85.3	2.0
ETVA	1K-A	**51.3**	**76.2**	**85.7**	**1.0**	**48.8**	**80.3**	**86.4**	**1.0**
UATVR [9]	1K-B	50.8	76.3	85.5	1.0	48.1	76.3	85.4	2.0
TEFAL [10]	1K-B	49.9	76.2	84.4	2.0	-	-	-	-
X-Pool [6]	1K-B	48.2	77.3	85.1	2.0	-	-	-	-
MPT [7]	1K-B	49.2	72.9	82.4	-	47.4	73.9	83.4	-

续表

T-MASS [8]	1K-B	51.3	76.5	85.6	1.0	50.9	80.2	88.0	1.0
ETVA	1K-B	**52.7**	**77.3**	**86.1**	**1.0**	**51.7**	**81.1**	**89.4**	**1.0**
UATVR [9]	Full	22.0	35.4	42.3	6.0	18.7	35.3	41.2	6.0
TEFAL [10]	Full	23.7	37.1	44.8	6.0	20.2	35.4	45.1	6.0
PAU [11]	Full	22.2	36.2	46.5	4.0	20.7	36.7	58.8	4.0
Cap4Video [12]	Full	22.9	34.0	47.4	3.0	-	-	-	-
T-MASS [8]	Full	23.0	39.0	56.2	3.0	21.6	46.9	61.2	3.0
ETVA	Full	**24.7**	**40.8**	**57.1**	**2.0**	**22.7**	**48.9**	**62.1**	**2.0**

对于 ActivityNet Captions 数据集，实验结果如表 2 所示。因其长视频与复杂人类活动场景的特性，对模型要求更高。ETVA 在该数据集的 R@10 指标上达到 93.5%，在视频－文本 R@10 指标上也达到 94.1%。超越对比方法 MMT 和 Cap4Video。在 LSMDC 数据集上，ETVA 同样取得了较好的结果。如表 3 所示。在视频到文本检索任务中，ETVA 在 R@1 指标上相较于 MMT 实现了 1.4%的提升，这说明 ETVA 能够有效处理从电影中提取的视频数据，对不同领域的视频都有较好的适应性电影片段的视觉元素丰富且情节复杂，ETVA 的全局－局部对齐设计使其能更好地处理这种复杂场景。

Table 2. Experimental results on the ActivityNet Captions dataset
表 2. ActivityNet Captions 数据集上的实验结果

方法	文本－视频				视频－文本			
	R@1	R@5	R@10	Mdr	R@1	R@5	R@10	MdR
VIP	18.2	44.8	89.1	6	16.7	43.1	88.4	7
X-Pool	18.2	47.7	91.4	6	35.0	64.9	77.6	3
Cap4Video	33.9	62.1	76.4	-	-	-	-	-
MMT	42.7	74.2	93.2	2	42.9	74.8	93.1	4
ETVA	**45.3**	**75.5**	**93.5**	**2**	**45.0**	**76.6**	**94.1**	**2**

Table 3. Experimental results on the LSMDC dataset
表 3. 在 LSMDC 数据集上的实验结果

方法	文本－视频				视频－文本			
	R@1	R@5	R@10	Mdr	R@1	R@5	R@10	MdR
UCOFIA	13.6	27.9	33.5	32.0	-	-	-	-
CLIP4Clip	15.1	28.5	36.1	28.0	-	-	-	-
Frozen	12.9	28.7	35.1	35.5	-	-	-	-
UATVR	15.8	37.6	40.1	-	-	-	-	-

续表

TEFAL	10.1	25.6	34.6	27.0	-	-	-	-
X-Pool	22.3	38.0	-	-	18.5	36.4	-	-
MMT	21.2	40.9	48.8	11.0	20.9	39.4	48.3	11.0
ETVA	**22.6**	**41.8**	**49.2**	**9.0**	**21.8**	**40.3**	**49.8**	**10.0**

综合多个数据集的实验结果，ETVA 模型在文本－视频检索任务中相较于其他方法，在准确性和检索效率上具有一定的优势，通过全局－局部对齐能够有效提升了模型对文本与视频语义匹配的能力。

3.3. 消融实验

为深入剖析 ETVA 模型各组件的作用，开展了消融实验。如表 4 所示。通过逐步去除模型的部分组件，观察其对实验结果的影响，进而明确各模块在模型整体性能中的重要性。

Table 4. Ablation experiment results on the MSRVTT dataset
表 4. MSRVTT 数据集上的消融实验结果

方法	文本－视频				视频－文本			
	R@1	R@5	R@10	MdR	R@5	R@5	R@10	MdR
去除全局对齐	44.3	71.5	79.4	4	46.6	74.9	81.6	4
去除细粒度对齐	42.2	69.9	81.6	3	44.0	75.7	83.6	4
全模型	**51.3**	**76.2**	**85.7**	**1**	**48.8**	**80.3**	**86.4**	**1**

当从 ETVA 模型中去除全局对齐模块后，在 MSRVTT 数据集上，Recall@1 指标从原本的 51.3%大幅下降至 44.3%，Recall@5 指标也从 76.2%降至 71.5%，Median Rank 从 1 升至 4。这表明缺少全局对齐模块，模型难以从整体上把握文本与视频的语义关系，导致在检索任务中，相关视频在检索结果中的排名显著靠后，命中准确率大幅降低。例如，对于描述“一群人在公园里举行欢乐聚会”的文本，没有全局对齐模块的模型，可能无法有效整合视频中关于公园场景、人物活动等整体信息与文本的对应关系，容易被局部细节干扰，从而难以准确筛选出相关视频。

当去除全局对齐分支，仅对局部对齐进行训练时，模型出现了损失无法收敛的现象。这一现象充分说明了全局对齐对于局部对齐的优化过程起着不可或缺的辅助监督作用。进一步地，在测试阶段仅去除全局对齐的实验结果显示，与完整模型相比，文本到视频检索任务中的 R@1 指标下降了 7.0%。这一数据直观地证明了全局特征与局部信息之间具有显著的互补性。全局对齐不仅能够提供更为全面的语义信息，还能帮助局部对齐更好地优化，二者协同工作，才能使模型达到更优的性能表现。

去除细粒度对齐模块后，模型在 MSRVTT 数据集上表现恶化明显。该数据集包含复杂的长视频和详细的人类活动描述，细粒度对齐模块对于捕捉其中的局部语义细节至关重要。去除该模块后，Recall@10 指标从 85.7%骤降至 81.6%。在处理如“运动员先进行拉伸热身，然后开始跑步训练，中途调整呼吸，最后冲刺”这类包含多个连续具体动作的文本－视频对时，模型无法精准匹配文本中的每个动作细节与视频中的对应片段，因为缺少了细粒度对齐模块对局部特征的深入挖掘和匹配能力，使得模型在处理复杂活动序列时效果较差。

在细粒度对齐模块中，共享聚类中心策略是该模型的核心机制。若调整该策略(如减少聚类中心数量)，

MSRVTT 数据集上的模型性能将受到影响。当只采用最大池化操作时，用“text”表示用最大池化操作替换用于局部视频特征编码的共享 NetVLAD 层，然后将该特征投影到与文本局部特征相同的维度。R@1 指标从 51.3%下降至 49.4%，性能大幅下降。MSRVTT 数据集包含视频片段及丰富复杂的情节描述，共享聚类中心负责将文本和视频的局部特征进行有效聚类和匹配。减少聚类中心数量后，模型无法充分捕捉到文本和视频中多样的局部语义模式，导致在处理视频中复杂的场景变化、人物关系等细节时，难以准确对齐文本与视频的局部特征，进而降低了模型的检索准确性。

Table 5. Ablation experimental results of VLAD encoding on the MSRVTT dataset
表 5. 在 MSRVTT 数据集上关于 VLAD 编码的消融实验结果

方法	文本 - 视频				视频 - 文本			
	R@1	R@5	R@10	MdR	R@5	R@5	R@10	MdR
text	49.4	73.3	78.2	4	43.5	72.4	67.7	4
separate	50.6	75.2	79.4	4	41.4	71.7	70.1	3
ETVA	**51.3**	**76.2**	**85.7**	**1**	**48.8**	**80.3**	**86.4**	**1**

多专家模块在视频编码器中起着提取多模态视频特征的重要作用。如表 5 所示。当去除多专家模块后，模型性能也出现大幅下滑。“separate”表示不执行文本特征编码和视频特征编码之间的中心共享。在 MSRVTT 数据集上，Recall@5 指标下降了一个百分点。多专家模块能够从不同模态(如视觉、音频等)和特征类型(如运动、外观等)对视频进行编码，去除该模块后，模型只能获取单一或有限类型的视频特征，无法全面捕捉视频中的丰富信息，使得文本与视频特征之间的匹配度降低，检索性能随之变差。

通过上述消融实验结果可知，ETVA 模型的各个组件，包括全局对齐模块、细粒度对齐模块、共享聚类中心策略以及多专家模块，对于模型在文本 - 视频检索任务中的性能表现均具有不可或缺的作用。任何一个组件的缺失或改变，都会显著影响模型对文本与视频语义关系的理解和匹配能力，进而降低模型的检索准确性和效率，这充分证明了 ETVA 模型整体架构设计的合理性和各组件之间的协同重要性。

3.4. 可视化分析

文本特征与本地视频特征被分配至提出的 ETVA 模型中的一组共享中心。从图 4 中可以看到，在中心 1 上，分配值最高的文本特征与“man”相关。而所有分配到中心 1 的视频帧，同样包含“man”相关的外观信息。在中心 2 上，分配值最高的文本内容为“contraption”，且分配到该中心的唯一视频帧，内容正是视频中的“contraption”。在中心 6 上，文本“flying”“in”和“field”均具有较高分配值，而分配到该中心的视频帧，也包含这些内容。值得注意的是，“crashes”这个词，在所有中心的分配值始终较低。这是由于训练数据有限，模型难以理解这类低频词。

此外，为了展示模型的检索效果，在 MSR-VTT 数据集的测试集上对两个文本检索视频的示例进行了可视化处理，具体内容如图 5 所示。从图 5(a)可以看到，当输入检索内容“A girl is singing”时，模型成功检索出了女人正在唱歌的场景。虽然这些检索结果都符合预期要求，但仅有一个结果被标记为绿色框。这是因为模型默认一个查询仅对应一个最优结果。再看图 5(b)，当检索“girl is giving a speech”时，正确的检索结果位于首位。该检索语句同时包含了动作和场景信息，而其他检索结果虽然与最佳结果有一定相似性，但并不符合检索语句中设定的场景。通过这两个示例可以发现，本模型能够实现检索语句中的动作和场景与视频内容的细粒度对齐，从而有效提高了检索效率。

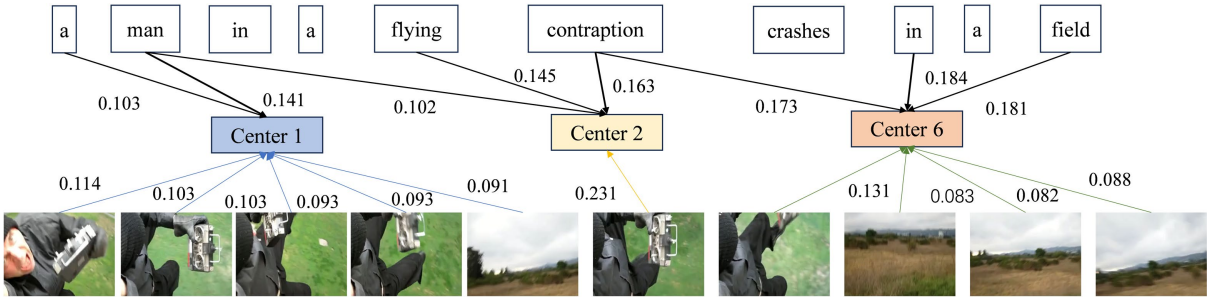

Figure 4. Visualization of weight allocation
图 4. 分配权重可视化

(a) a girl is singsing

(b) girl is giving a speech

Figure 5. Text-video retrieval results
图 5. 文本－视频检索结果

4. 总结

本章提出了一种端到端的文本－视频序列对齐方法，旨在优化文本与视频之间的局部语义对齐，提升检索系统的性能。通过引入基于 NetVLAD 的局部对齐机制，并提出一种 ETVA (Efficient Text-video Aligner)模型，用于实现文本－视频的协同编码。实验结果在三个标准文本－视频检索基准上展现了 ETVA 方法的有效性，同时可视化分析进一步验证了联合语义主题学习策略的有效性。

参考文献

[1] Yang, X., Dong, J., Cao, Y., Wang, X., Wang, M. and Chua, T. (2020) Tree-Augmented Cross-Modal Encoding for Complex-Query Video Retrieval. *Proceedings of the* 43*rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 25-30 July 2020, 1339-1348. https://doi.org/10.1145/3397271.3401151

[2] Wang, Z., Zhong, Y., Miao, Y., *et al.* (2022) Contrastive Video-Language Learning with Fine-Grained Frame Sampling. arXiv: 2210.05039.

[3] Chen, S., Zhao, Y., Jin, Q. and Wu, Q. (2020) Fine-Grained Video-Text Retrieval with Hierarchical Graph Reasoning. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 10635-10644. https://doi.org/10.1109/cvpr42600.2020.01065

[4] Bar-Shalom, G., Leifman, G. and Elad, M. (2024) Weakly-Supervised Representation Learning for Video Alignment and Analysis. 2024 *IEEE*/*CVF Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 3-8 January 2024, 6895-6904. https://doi.org/10.1109/wacv57701.2024.00676

[5] Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., *et al.* (2022) Clip4clip: An Empirical Study of CLIP for End to End Video Clip Retrieval and Captioning. *Neurocomputing*, **508**, 293-304. https://doi.org/10.1016/j.neucom.2022.07.028

[6] Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J. and Ji, R. (2022) X-CLIP: End-To-End Multi-Grained Contrastive Learning for Video-Text Retrieval. *Proceedings of the* 30*th ACM International Conference on Multimedia*, Lisboa, 10-14 October 2022, 638-647. https://doi.org/10.1145/3503161.3547910

[7] Gorti, S.K., Vouitsis, N., Ma, J., Golestan, K., Volkovs, M., Garg, A., *et al.* (2022) X-Pool: Cross-Modal Language-Video Attention for Text-Video Retrieval. 2022 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 4996-5005. https://doi.org/10.1109/cvpr52688.2022.00495

[8] Zhang, H., Zeng, P., Gao, L., Song, J. and Shen, H.T. (2024) MPT: Multi-Grained Prompt Tuning for Text-Video Retrieval. *Proceedings of the* 32*nd ACM International Conference on Multimedia*, Melbourne, 28 October-1 November 2024, 1206-1214. https://doi.org/10.1145/3664647.3680839

[9] Wang, Z., Sung, Y., Cheng, F., Bertasius, G. and Bansal, M. (2023) Unified Coarse-To-Fine Alignment for Video-Text Retrieval. 2023 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 2804-2815. https://doi.org/10.1109/iccv51070.2023.00264

[10] Bain, M., Nagrani, A., Varol, G. and Zisserman, A. (2021) Frozen in Time: A Joint Video and Image Encoder for End-To-End Retrieval. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 1708-1718. https://doi.org/10.1109/iccv48922.2021.00175

[11] Wang, J., Wang, P., Sun, G., Liu, D., Dianat, S., Rao, R., *et al.* (2024) Text Is MASS: Modeling as Stochastic Embedding for Text-Video Retrieval. 2024 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 16551-16560. https://doi.org/10.1109/cvpr52733.2024.01566

[12] Li, H., Song, J., Gao, L., *et al.* (2023) Prototype-Based Aleatoric Uncertainty Quantification for Cross-Modal Retrieval. *Advances in Neural Information Processing Systems*, **36**, 24564-24585.