Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(11): 20073–20092.
DOI: 10.3934/mbe.2023889
Received: 14 August 2023
Revised: 10 October 2023
Accepted: 17 October 2023
Published: 03 November 2023

Research article

A cross-modal conditional mechanism based on attention for text-video retrieval

Wanru Du[1,2,*], Xiaochuan Jing[2], Quan Zhu[1,2], Xiaoyin Wang[2] and Xuan Liu[1,2]

[1] China Aerospace Academy of Systems Science and Engineering, Beijing 100048, China
[2] Aerospace Hongka Intelligent Technology (Beijing) CO., LTD., Beijing 100048, China

* **Correspondence:** Email: wanru_du@163.com; Tel: +8618810937231.

Abstract: Current research in cross-modal retrieval has primarily focused on aligning the global features of videos and sentences. However, video conveys a much more comprehensive range of information than text. Thus, text-video matching should focus on the similarities between frames containing critical information and text semantics. This paper proposes a cross-modal conditional feature aggregation model based on the attention mechanism. It includes two innovative modules: (1) A cross-modal attentional feature aggregation module, which uses the semantic text features as conditional projections to extract the most relevant features from the video frames. It aggregates these frame features to form global video features. (2) A global-local similarity calculation module calculates similarities at two granularities (video-sentence and frame-word features) to consider both the topic and detail features in the text-video matching process. Our experiments on the four widely used MSR-VTT, LSMDC, MSVD and DiDeMo datasets demonstrate the effectiveness of our model and its superiority over state-of-the-art methods. The results show that the cross-modal attention aggregation approach can effectively capture the primary semantic information of the video. At the same time, the global-local similarity calculation model can accurately match text and video based on topic and detail features.

Keywords: cross-modal retrieval; attention mechanism; video and text alignment; global-local similarity calculation

1. Introduction

The information obtained from various sources is commonly referred to as multi-modal information. At present, research in this field often begins with the aim of combining multi-modal information and maximizing the potential of different modal information in real life to create new cross-modal research areas [1–4]. The demand for content-based video retrieval has increased with the widespread use

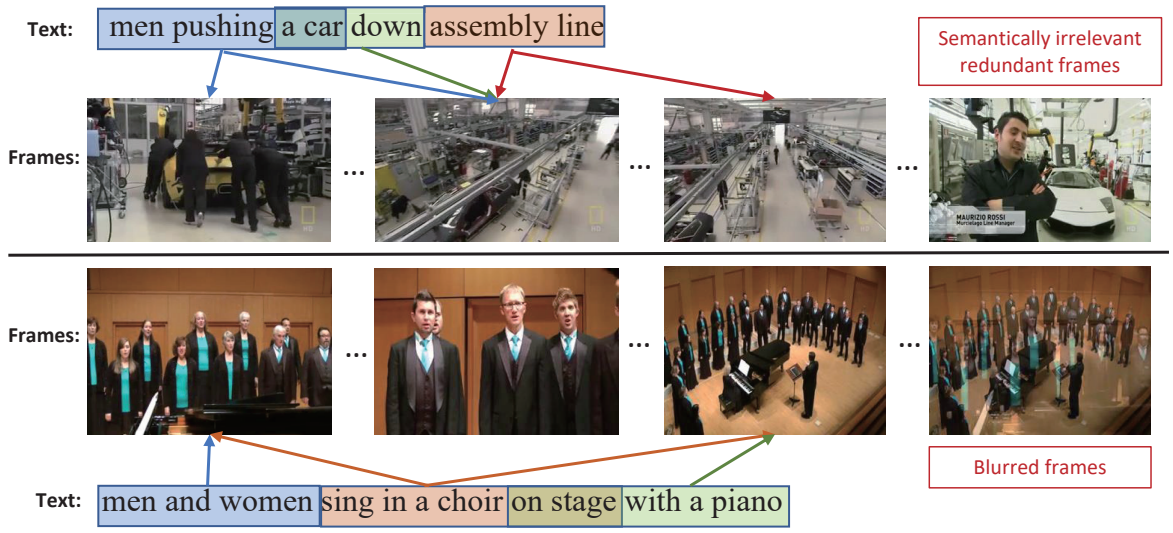

Figure 1. The difference between text and video in terms of information expression.

of video platforms such as TikTok and YouTube. Cross-modal retrieval [5–7] has gained significant attention from scholars in recent years.

However, text and video modes of conveying information differ significantly. The text conveys information through words or phrases, while video encompasses a broader range of information. The text only partially represents some of the content in a video [8]. When performing sentence and video matching, adopting a more precise approach may be necessary due to redundant frames in videos. For instance, Figure 1 shows two sample videos and their matching text from the MSR-VTT dataset [9]. Observations have shown that some video frames do not match the semantic meaning of the sentences and are deemed redundant components in the text-video matching procedure [10]. This highlights the presence of some bias in the results when matching videos with sentences.

To enhance the precision of text-video retrieval, a proven practical approach involves the elimination of redundant frames and a focused analysis of video subintervals that are semantically linked to the text, as demonstrated by Dong et al. [11]. Consequently, the matching model must possess the capability to extract critical textual content and correctly identify corresponding video segments.However, prevalent methods commonly resort to mean pooling or employ self-attention mechanisms [6, 12, 13] to derive global feature embeddings for a given sentence and video. These methods exhibit limitations regarding the definition and precise localization of keyframes. This limitation is significant, as it can negatively impact the performance of retrieval tasks, particularly when videos contain content that is not explicitly described in the accompanying text.

This paper presents a novel cross-modal conditional mechanism designed to address two critical issues: feature redundancy within videos and the imbalanced semantic alignment between the two modalities. To tackle these challenges, we propose two key components:

Firstly, we introduce a video feature aggregation module that leverages a text-conditioned attention mechanism. Here, the text features serve as a guiding condition, enhancing the emphasis on keyframes while diminishing the influence of redundant frames. The aggregation process combines numerical values by multiplying attention weights with frame features, resulting in refined video features.

Secondly, we introduce a global-local cross-modal conditional similarity calculation module. This

module considers video-sentence features as the global data and frame-word features as the local data. These features are input into the model for similarity calculation, a critical step for text-video matching. This approach enables us to effectively address the overarching topic and the finer details in the matching process.

2. Related work

2.1. Feature extraction

Text Feature Encoding Previous studies [14–17] examined the extraction of text features and have achieved exceptional outcomes. The Skip-Gram model [18] begins by considering the central word and predicts its surrounding words, producing text features. This approach not only cuts down the computational effort during the training phase but also elevates the quality of the word-to-vector representation. To enhance the matching between text and images, the Fisher Vector model [19] examines text representation and quantifies it using high-level statistical features for text feature extraction [20]. Furthermore, the GRU model [21] was introduced to address the issue of gradient disappearance in standard RNNs during text encoding. As a result, the GRU model has become a widely utilized text encoder. OpenAI has also made available a graphical pre-training model for CLIP based on contrast learning and has a transformer structure [22]. CLIP model has delivered similarly remarkable results in the coding of text.

Visual Feature Encoding Visual feature extraction is typically carried out using supervised or self-supervised research methods. Recently, there has been growing interest in using a transformer-based image encoder known as the ViT model [23]. While the application of transformers to feature extraction of video content [24, 25] is still in its early stages, it has shown potential for enhancing action classification in video text retrieval. Researchers have been exploring new and innovative approaches to enable models with better generalization capabilities [22, 49]. Text and video pairs obtained from the internet are collected and formed into large-scale datasets for training. One of the most successful methods is the CLIP model [22], which has achieved state-of-the-art performance in image feature extraction. The pre-trained CLIP model can learn more sophisticated visual concepts and use these features in retrieval tasks. To mitigate the impact of diverse topics in the dataset, a MIL-NCE model [26] based on the CLIP video encoder has been proposed and tested with positive results on the Howto100M [13] dataset. Furthermore, the ClipBERT [49] model, which is based on the MIL-NCE model, employs an end-to-end approach to streamline the pre-processing stage of video-text retrieval. This paper uses a pre-trained CLIP-based ViT model as our video encoder to extract visual features from the video frames. The effectiveness of the feature extraction has been verified through experimental evaluation.

2.2. Text-video retrieval

In cross-modal retrieval, text-video matching plays a key role in bridging vision and language. Text-video matching aims to learn the cross-modal similarity function [27] between text and video, so that related text-video pairs receive higher scores than unrelated ones. Establishing a semantic similarity model that effectively reduces the semantic gap between visual and textual information is crucial for the accuracy of this study [28]. Despite the complex matching patterns and vast semantic differences

between images and texts, this remains a challenging research topic. A common approach to overcome this challenge is mapping images and texts into a shared semantic space through a suitable embedding model, i.e., a joint latent space, and then computing cross-modal similarity in this shared space.

Text-video retrieval is typically achieved by integrating a pre-trained language model with a visual model to associate text features with visual features. When dealing with small datasets, incorporating a pre-trained model can improve performance. For instance, the Teachtext model [23] uses multiple text encoders to provide a complementary supervised signal for the retrieval model. MMT [29] and MDMMT [30] were early examples of using transformers for multi-modal video processing, integrating three modal features to accomplish the video retrieval task.

Additionally, some scholars have applied concepts from the data hashing field to tasks involving cross-modal data processing and information retrieval. The ROHLSE model [31] focuses on addressing label noise and exploiting semantic correlations in processing large-scale streaming data. This work presents an innovative approach for hashing streaming data. The DAZSH model [32] introduces a hashing method tailored to the zero-shot problem in cross-modal retrieval. Integrating data features with class attributes effectively captures relationships between known and unknown categories, facilitating the transfer of supervised knowledge. Moreover, a neural network-based approach [33] is designed to learn specific category centers and guide the hashing of multimedia data. Finally, the SKDCH model [34] proposes a semi-supervised knowledge distillation method for cross-modal hashing. It mitigates heterogeneity gaps and enhances discriminability by improving the triplet ranking loss. These studies collectively demonstrate the application of data hashing principles to tackle complex challenges in cross-modal data processing and information retrieval.

Recently, the CLIP model [22] utilized a rich text-image dataset to create a joint text-visual model, which the authors of the CLIP4CLIP model [6] leveraged through transfer learning to achieve state-of-the-art results in video retrieval tasks. In several studies based on the CLIP model [35], the model outperformed most other works [2, 12, 36], even in a zero-shot manner, showcasing its excellent generalization capabilities in text-video understanding.

Several video feature aggregation methods, including average pooling, self-attention, and multi-modal transformers [4, 6], are commonly used in CLIP-based studies and have been shown to match text and images effectively. However, there needs to be more research specifically focused on matching video sub-regions with words [49]. As noted in the previous section, many video frames are semantically irrelevant to the text in matching processes. Thus, using a cross-modal conditional attention mechanism to reduce the impact of redundant frames on retrieval results is the motivation behind this paper's research.

2.3. Cross-modal attention mechanism

In natural language processing, attention mechanisms are widely used to filter redundant information [37]. Similarly, attention mechanisms have been used to enhance the focus on visual and textual local features in cross-modal information-matching tasks. Some researchers have proposed a similarity attention filtration (SAF) module [38] based on attention mechanisms to match images with text. This module applies attention mechanisms to cross-modal feature alignment, aiming to eliminate the interference caused by redundant text-image pairs and enhance image retrieval accuracy. Owing to the remarkable performance of attention mechanisms in the cross-modal domain, certain researchers [39] have developed more intricate bidirectional focused attention networks, building upon this founda-

tion to enhance matching accuracy further. Concurrently, other scholars [40, 41] have introduced a recurrent attention mechanism to investigate the correspondence between fine-grained text regions and individual words.

The crucial aspect of implementing the attention model in text-video cross-modal inference lies in embedding the features of both text and video and subsequently identifying frames that align more effectively with text semantics, as demonstrated by Tan et al. [28]. We have incorporated a textual conditional attention module into our cross-modal matching model to achieve this. This module filters out extraneous semantic information within the frames by computing attention weights for each frame, using text semantics as a conditional projection.

3. Framework

Text-video retrieval can be defined as two tasks: one is retrieving semantically close text by the given video information as the input, named t2v. The other is retrieving semantically similar videos by the sentence given as the input, named v2t. Taking the t2v task as an example, a query text and a set of video sets to be queried are the input data. The model calculates the similarity score between the query text and each video in the video set and finds the video with the best semantic match to return. Similarly, v2t has a similar task. This paper mainly focuses on the t2v task as the leading study. We are dedicated to enhancing the accuracy of text-video retrieval tasks by implementing two pivotal strategies: filtering out irrelevant frames and aggregating key-frames to construct video features, followed by performing a global-local multi-modal feature matching approach.

Figure 2 illustrates the framework of our model for the text-video retrieval task. The text-video retrieval task is quantified into three main components: Data Embedding, Cross-modal Feature Extraction, and Similarity Calculation. In the Data Embedding phase, we feed the input data (including words and frames) into the text encoder ψ and the image encoder ϕ of the CLIP model, obtaining embedded data representations. The Cross-modal Feature Extraction section encompasses two critical steps. Firstly, we employ a self-attention mechanism to extract sentence features from the text. Secondly, we utilize a conditional attention mechanism to filter out redundant and aggregate frames semantically relevant to the text, thereby obtaining more precise video features. In the Similarity Calculation phase, we compute similarity at global and local granularities (i.e., video-sentence and frame-word features) to consider thematic and detail features during the text-video matching process. It is worth noting that the Cross-modal Feature Extraction and Similarity Calculation sections contain two innovative modules introduced in this paper, which are detailed as follows:

Cross-modal Conditional Attention Aggregation Module To process text input t, we pass it through a text encoder ψ to obtain its word embedding E_w. This embedding is then multiplied with the weight matrix query projection W_Q, to produce the text query vector Q_t. For video input v, it is passed through a video encoder to produce frame embedding E_f. This embedding is then multiplied with the key projection matrix W_K and the value projection matrix W_V, respectively, to obtain the key embedding of the frames K_v, and the value embedding of the frames V_v. Then we calculate the attention score of the video frames w_{att}, by taking the dot product of Q_t and K_v. The attention scores are used to weight the value vectors of the video frames V_v, to produce the self-attention frame feature embedding.

Global-Local Similarity Matching Module proposes a cross-similarity calculation module to perform the text-video matching task. The module integrates cosine similarity and conditional probability

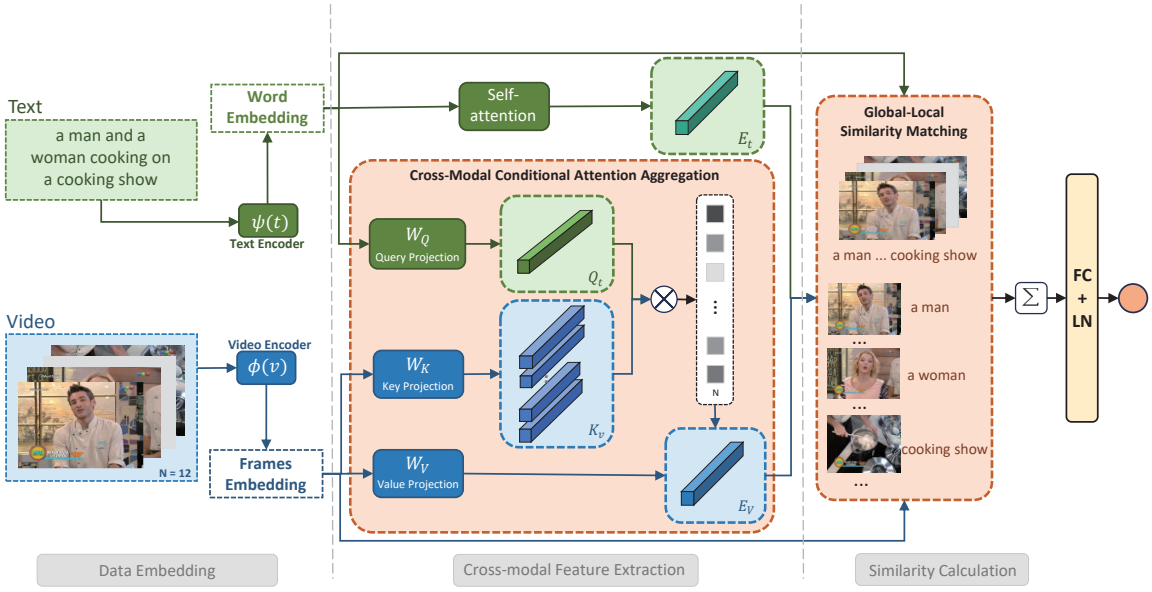

Figure 2. A Brief illustration of our proposed approach.

models to compute the similarity scores between the different modal data feature embeddings, considering their mutual dependence. The global feature data (text-video) and local feature data (word-frame) are fed into the model separately, producing their similarity scores. The model then aggregates the global and local similarity scores through self-attention to obtain the final matching scores.

4. Methodology

In this section, we concentrate on the methods for implementing the model presented in the paper. To facilitate a comprehensive understanding of our model, we commence by elucidating the procedure for utilizing the pre-trained CLIP model to encode text and video in Section 4.1. Subsequently, the following two sections introduce pivotal functional components of our model: the Cross-modal Conditional Attention Aggregation Module (Section 4.2) and the Global-Local Similarity Matching Module (Section 4.3). Section 4.2 describes the method for incorporating attention mechanisms into cross-modal feature aggregation to enhance the relevance of video features to text semantics. In Section 4.3, we highlight the limitations of traditional similarity computation method for cross-modal feature matching and propose a novel method for computing the global-local similarity of correlated cross-modal features. Finally, we present the implementations of training objectives with both the two modules in Section 4.4.

4.1. Data embedding

The video can be considered a sequence of images, with each video frame being an individual image. In this study, many pre-trained models have been found to extract features from text and images effectively, enabling cross-modal semantic understanding [6, 22]. These models have been pre-trained on large and diverse datasets, allowing us to leverage their excellent performance in feature extraction to simplify the training process of our work.

CLIP models trained on large, richly typed datasets have demonstrated exceptional feature extraction abilities and robust performance in downstream tasks. Numerous studies have shown that CLIP performs well in extracting the rich semantic features of input information [22]. In the task of video feature extraction, individual video frames are embedded in CLIP's joint latent space as images. The video features are obtained by aggregating the embedded features of the individual frames. In this paper, we learn a new joint latent space based on the CLIP model to serve as an encoder for our standard video-text feature extraction.

Given text t and video v as inputs, we first preprocess the video into quantifiable frames v^{f_n} and input these frames into the CLIP model as images. CLIP then outputs a text embedding E_t and a frame embedding $E_v^{f_n}$ as encodings. By aggregating the sequence of frame embeddings S_F, we can obtain the video embedding E_V:

$$E_t = \psi(t) \in \mathbb{R}^d \tag{4.1}$$

$$E_v^{f_n} = \phi(v^{f_n}) \in \mathbb{R}^d \tag{4.2}$$

$$Set_F = [E_v^{f_1}, E_v^{f_2}, \cdots, E_v^{f_n}] \in R^{n \times d} \tag{4.3}$$

where ψ is CLIP's text encoder, and ϕ is CLIP's image encoder. Set_F is the set of frames feature embedding.

Then we can obtain the video's feature embedding by a temporal aggregation function ρ:

$$E_V = \rho(Set_F) \tag{4.4}$$

Obviously, E_t and $E_v^{f_n}$ are the two outputs of CLIP.

4.2. Cross-modal conditional attention aggregation

Previous research has typically used average pooling or self-attention mechanism when calculating the video embedding by aggregating the frame embeddings [12, 29]. However, this approach results in a video embedding that contains many redundant visual features that need to be more relevant to the semantic features of the text. This is because the text has much less semantic information than the video. As a result, these aggregate methods can negatively impact the accuracy of the final similarity computation results.

The aggregation of frame features to obtain the video embedding for use in the similarity calculation model often results in the inclusion of redundant visual features that need to be more relevant to the semantic features of the text. This can negatively impact the accuracy of the final similarity computation results.

This module uses the attention mechanism to extract the video features. We combine the semantic text features to compute the attention weights for the keyframes. This enhances the crucial information in the frames and filters out redundant information, resulting in video features. Firstly, we project the text embedding E_t as a query vector $Q_t \in \mathbb{R}^{1 \times d_a}$. The video embedding obtained from Section 4.1 is then projected as a key vector $K_F \in \mathbb{R}^{1 \times d_a}$ and a value vector $V_F \in \mathbb{R}^{1 \times d_a}$ through dot product operations with matrices $W_K \in \mathbb{R}^{d \times d_a}$ and $W_V \in \mathbb{R}^{d \times d_a}$, respectively. The calculations are defined as follows:

$$Q_t = W_Q \cdot E_t \tag{4.5}$$

$$K_F = W_K \cdot Set_F \tag{4.6}$$

$$V_F = W_V \cdot Set_F \tag{4.7}$$

where W_Q, W_K and W_V are the parameter matrices obtained from the neural network training.

Finally, by utilizing the cross-modal attention feature aggregation module, we obtain the joint text-video semantic attention scores for each frame, represented as S_{f_n} .

$$S_V = [S_{f_1}, S_{f_2}, \cdots, S_{f_n}] = \text{softmax}\left(\frac{Q_t K_v^T}{\sqrt{d_a}}\right) V_v \tag{4.8}$$

The above equation is the main idea of the aggregation function ρ, and the input video features embedding E_V can finally be calculated as follows:

$$E_V = S_{f_1} E_{f_1} + S_{f_2} E_{f_2} + \cdots + S_{f_n} E_{f_n} \tag{4.9}$$

4.3. Global-local similarity matching

In Section 4.1, the CLIP encoder obtains the text feature embedding E_t and the set of frame feature embeddings Set_F. Section 4.2 then leverages the attention mechanism to aggregate the frame embeddings and get the text-conditional video embedding E_v. Although this approach incorporates semantic text features into the video feature embedding, conventional similarity computation models, such as cosine similarity, can only improve the matching accuracy to a certain extent. It may still need to look at the local semantics expressed in specific keyframes. This section considers the consistency of structure and text word features in semantic expression to address this issue. It combines the similarity computation of both video and sentences to perform text-video matching.

Vector Similarity Function The previous methods of calculating the similarity between features of two different modal data often relied on cosine or Euclidean distance [40]. While these methods can capture relevance to a degree, they cannot detect finer local correspondences between the vectors. Our proposed similarity representation function aims to address this issue by leveraging the local features of the vectors and using cosine similarity calculation as the core component. This enables a more in-depth analysis of the correlation information between the feature representations from different modalities. The similarity function is formulated as follows:

$$f(\alpha_1, \alpha_2; W_{\text{sim}}) = \frac{W_{\text{sim}}|\alpha_1 - \alpha_2|^2}{||W_{\text{sim}}|\alpha_1 - \alpha_2|^2||_2} \tag{4.10}$$

where $||\alpha_1 - \alpha_2||^2$ is the square operation of each element in the result $\alpha_1 - \alpha_2$, and $||W_{sim}|\alpha_1 - \alpha_2|^2||_2$ is the l_2- operation of $W_{sim}|\alpha_1 - \alpha_2|^2$. The W_{sim} in the equation is a learnable parameter matrix to obtain the similarity vector.

Text-Video Global Similarity Calculation According to the similarity Eq (4.10), we replace α_1 and α_2 with the text feature embedding E_t and the video feature embedding E_V, respectively.

$$Sim^g = f(E_V, E_t; W_g) = \frac{W_g|E_V - E_t|^2}{||W_g|E_V - E_t|^2||_2} \tag{4.11}$$

where W_g is the parameter matrix that aims to learn the global similarity through training.

Frame-Text Local Similarity Calculation To exploit the local semantic information in frames, we propose a similarity calculation regarding the similarity between the video's local frames and words.

First, we obtain the cosine similarity C_{ij} of the frame feature vector v_i and the word vector t_j:

$$C_{ij} = \frac{v_i^T \cdot t_j}{\|v_i \cdot t_j\|} \tag{4.12}$$

Then, softmax is used to normalize the cosine similarity to obtain the local feature weights β_{ij}.

$$\beta_{ij} = \frac{\max(0, C_{i,j})}{\sqrt{\sum_{i=1}^{n}(\max(0, C_{i,j})^2}} \tag{4.13}$$

After obtaining the attention weights, we calculate the frames feature representation containing the words' semantic information:

$$V_i^f = \sum_{i=1}^{n} \beta_{ij} v_i \tag{4.14}$$

Finally, we compute the frame-text local similarity representation between V_i^f and t_j using Eq (4.10):

$$sim_j^l = f(V_i^f, t_j; W_l) \tag{4.15}$$

where W_l is also the parameter matrix like W_g.

Local similarity represents the association between capturing a specific word and the frames that make up the video, using finer-grained visual semantic alignment to improve similarity prediction.

4.4. Training objective

We take the widely used ranking loss function [42] as the training objective in our cross-modal retrieval task. Its goal is to evaluate the relative distance between input samples and optimize model training by incorporating the similarity calculation results into the ranking loss. The similarity computation model is defined as $sim()$, with positive samples (V, T) being the matched video-text pairs and the negative samples being mismatched pairs:

$$V^{'} = argmax_{r \neq v}(r, T) \tag{4.16}$$

$$T^{'} = argmax_{w \neq T}(V, w) \tag{4.17}$$

The loss is obtained referring to the ranking loss function:

$$Loss = \omega_1 Loss_{loc} + \omega_2 Loss_{glo} \tag{4.18}$$

where:

$$Loss_l(v_a, v_p, v_n) = \sum \max(0, s(v_a, v_p) - s(v_a, v_n) + \alpha) \tag{4.19}$$

$$Loss_g(v_a, v_p, v_n) = \max(0, s(v_a, v_p) - s(v_a, v_n) + \alpha) \tag{4.20}$$

where v_a is the anchor sample, representing the reference vector. v_p is the sample I or T that matches the reference sample. v_n is the sample $I^{'}$ or $T^{'}$ that does not match the reference sample. Vector parameters in the $Loss_l$ function refer to the frame or text local feature vectors. Vector parameters within the $Loss_g$ function refer to the video and text local feature vectors.

5. Experiments

To validate the effectiveness of our model, in this section, we demonstrate experiments on four widely used text-video retrieval datasets: MSR-VTT [9], LSMDC [44], MSVD [43] and DiDeMo [12]. The model's performance is evaluated by testing its performance in terms of different recall rates, ranking results, and comparing the results with experimental results from existing studies.

5.1. Datasets

MSR-VTT dataset was created by collecting 257 popular video queries from a commercial search engine, with each query including 118 videos. The current version of MSR-VTT offers 10,000 web video clips, totaling 41.2 hours and 200,000 clip-sentence pairs, and each video is annotated with approximately 20 captions. To compare with previous work, 7000 videos were selected for training [13], and 1000 videos were selected for testing [43], following the commonly used segmentation method in current studies. Since no validation set was provided, 1000 videos were randomly selected from MSR-VTT to form the validation set.

LSMDC dataset comprises 118,081 video clips extracted from 202 movies, ranging from two to 30 seconds. The validation set includes 7,408 clips, and the evaluation is performed on a separate test set consisting of 1000 videos from movies that are distinct from those in the training and validation sets.

MSVD dataset comprises 1970 videos ranging from 10 to 25 seconds, and each video is annotated with 40 captions. The videos feature various subjects, including people, animals, actions, and scenes. Each video was annotated by multiple annotators, with approximately 41 annotated sentences per clip and a total of 80,839 sentences. The standard splitting [6] was used, with 1,200 videos for training, 100 videos for validation, and 670 videos for testing.

DiDeMo dataset comprises 10,000 flickr videos, each annotated with 40,000 sentences. In the test set, there are 1000 videos. As per the approach in references, we assess paragraph-to-video retrieval, wherein all sentence descriptions for a video are concatenated to form a single query. Notably, this dataset includes localization annotations (ground truth proposals), and our reported results incorporate these ground truth proposals.

5.2. Implementation details

Data Pre-processing. Different datasets have varying video durations and frame sizes, making standardizing the model input format challenging. This study extracts 12 frames from each video according to a specified time window to resolve this issue. It uses them as representatives of the video content, ensuring a uniform input shape for the model. Additionally, to ensure consistency with previous work [2, 6, 12] and facilitate testing, the pixel size of each video frame was adjusted to 224×224.

Model Settings. The study employs the CLIP model as its backbone and initializes all encoder parameters based on the pre-trained weights of the CLIP model, as described in [22]. For each video, the ViT-B/32 image encoder of the CLIP model is used to obtain the frame embeddings, while the transformer text encoder of the CLIP model is used to obtain the text embeddings. The CLIP encoder has an output size of 512, which also determines the attention size of the three projection dimensions, which is set to 512. The weight matrices W_q, W_k, and W_v are randomly initialized, and the bias values

Table 1. Results of comparative experiments on text-to-video retrieval (R@1/5/10) on four widely used public datasets.

Method	MSR-VTT			LSMDC			MSVD			DiDeMo		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CE [22]	20.9	48.8	62.4	11.2	26.9	34.8	19.8	49	63.8	16.1	41.1	-
MMT [29]	26.6	57.1	69.6	12.9	29.9	40.1	-	-	-	-	-	-
Frozen [12]	31	59.5	70.5	15.0	30.8	39.8	33.7	64.7	76.3	34.6	65.0	74.7
HIT-pretrained [47]	30.7	60.9	73.2	14.0	31.2	41.6	-	-	-	-	-	-
MDMMT [30]	38.9	69.0	79.7	18.8	39.5	47.9	-	-	-	-	-	-
All-in-one [48]	37.9	68.1	77.1	-	-	-	-	-	-	32.7	61.4	73.5
ClipBERT [49]	22.0	46.8	59.9	-	-	-	-	-	-	20.4	48.0	60.8
CLIP-straight [35]	31.2	53.7	64.2	11.2	22.7	29.2	37	64.1	73.8	-	-	-
CLIP2Video [50]	30.9	55.4	66.8	-	-	-	47.0	76.8	85.9	-	-	-
Singularity [51]	42.7	69.5	78.1	-	-	-	-	-	-	53.1	79.9	88.1
LAVENDER [52]	40.7	66.9	77.6	26.1	46.4	57.3	46.3	76.9	86.0	53.4	78.6	85.3
CLIP4Clip-meanP [53]	43.1	70.4	80.8	20.7	38.9	47.2	46.2	76.1	84.6	43.4	70.2	80.6
CLIP4Clip-seqTransf [53]	44.5	71.4	**81.6**	22.6	41.0	49.1	45.2	75.5	84.3	42.8	68.5	79.2
VINDLU [54]	**46.5**	71.5	80.4	-	-	-	-	-	-	**61.2**	85.8	91.0
ours	45.3	**72.5**	81.3	**26.5**	**47.1**	**57.4**	**47.6**	**77.2**	**86.0**	60.7	**86.1**	**92.2**

Table 2. The impact of feature aggregation module (Module 1) configurations. Mean, Self-att, and Cross-Modal respectively denote the employment of mean-based feature aggregation, self-attention-based feature aggregation, and cross-modal feature aggregation conditioned on text semantics in the Feature Aggregation Module.

Test Model	Aggregation Method			Result				
	Mean	Self-Att	Cross-modal	R@1	R@5	R@10	MdR	MnR
1	✓			41.8	70.9	83.5	3.0	13.7
2		✓		45.3	74.5	84.7	2.0	12.3
3			✓	47.6	77.2	86.0	2.0	10.0

are set to 0. The output units of the fully connected layer are also set to 512, and a dropout of 0.3 is applied, as described in [45]. The study employs the Adam optimizer [46] for training, with an initial learning rate of 0.00002, and the learning rate is decayed using a cosine schedule, as described in [22].

The recall [12, 12, 29] represents the ratio of the valuable fraction in the detection results to that in the dataset. Recall at K was used to measure the model's performance, and recall at 1 ($R@1$), recall at 5 ($R@5$), and recall at 10 ($R@10$) were used as evaluation metrics during testing.

5.3. Results and analysis

In this section, we present the results of the retrieval performance of our model on the MSR-VTT, LSMDC, MSVD and DiDeMo datasets. The aim is to showcase the superiority of our model in comparison to other existing models.

5.3.1. Comparisons on four datasets

Table 1 presents the results of comparative experiments in text-to-video retrieval (R@1/5/10) across four widely utilized public datasets.

Table 3. The impact of similarity calculation module (Module 2) configurations. Local and Global respectively signify the utilization of local similarity calculation and global similarity calculation or Global-Local similarity calculation.

Text Model	Similarity Computation Method		Result				
	Local	Global	R@1	R@5	R@10	MdR	MnR
1	✓		45.2	75.2	85.6	3.0	11.5
2		✓	44.8	73.7	84.5	3.0	12.2
3	✓	✓	47.6	77.2	86.0	2.0	10.0

Figure 3. Visualization results for two examples of different aggregation strategies on MSVTT. The bars show the attention weight values for each frame. cross-Modal Attention Aggregation is marked in blue. The orange and gray markers are Mean Aggregation and Self-attention Aggregation without textual semantic involvement, respectively.

Comparing our method's results with existing approaches, we observe that on the MSR-VTT, LSMDC, MSVD, and DiDeMo datasets, our average accuracy rates are 66.4% (+ 0.3%), 43.7% (+ 0.4%), 70.3% (+ 0.2%) and 79.7% (+ 0.4%), respectively. These scores surpass the performance of the models listed in the table across all four datasets, thus validating the effectiveness of the approach presented in this paper.

More accurately, on the LSMDC and DiDeMo datasets, we observed that our model's R@1 results were lower than those of the VINDLU model. Upon analysis, it was discovered that the VINDLU model focuses on effective video-and-language pretraining, utilizing the jointly trained CC3M + WebVid2M dataset containing content domains that are more aligned with MSR-VTT, such as sports, news, and human actions. Consequently, the VINDLU model outperforms our model on the R@1 metric. However, due to our model's enhancements in capturing video themes and details, our overall performance excels over VINDLU on the R@5 and R@10 metrics.

Additionally, it is worth noting that only on the MSR-VTT dataset, the R@10 results of the CLIP4Clip-seqTransf model are slightly higher than our model's results. On all other datasets and

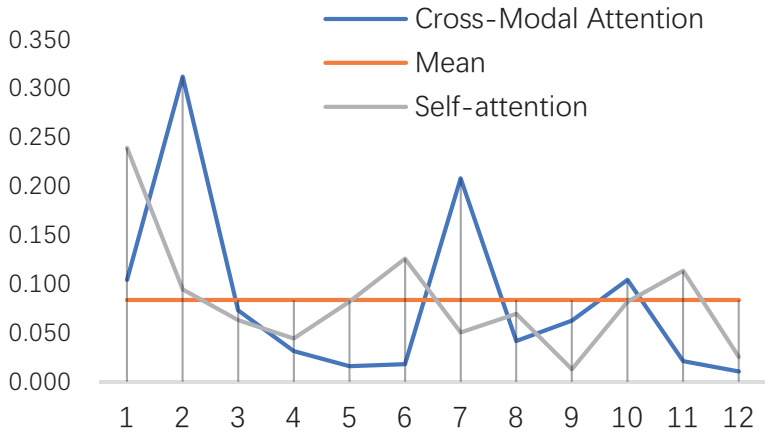

Figure 4. The trend of the weights corresponding to the key frames of the first example in Figure 3.

metrics, our model outperforms CLIP4Clip-seqTransf. Therefore, it can be considered that our model exhibits better stability in terms of performance compared to CLIP4Clip-seqTransf. Since both CLIP4Clip-seqTransf and our method use CLIP as the backbone, we can attribute the improvement in model performance to the fact that CLIP4Clip-seqTransfer employs a text-agnostic visual feature extraction approach, whereas our model utilizes a frame feature aggregation approach conditioned on text semantics.

Furthermore, on the LSMDC dataset, the retrieval task is more challenging due to the inherently vague textual descriptions of movie scenes. This conclusion can be drawn from the generally lower retrieval scores achieved by previous methods on this dataset. However, our approach outperforms the models listed in the table across all metrics. This demonstrates the significance of our model's ability to aggregate video features conditioned on text semantics. It learns the features of frames most relevant to the text semantics and suppresses the interference of redundant frames in feature aggregation.

5.3.2. Ablation studies

In this section, a series of ablation experiments are conducted to explore the two modules' effects to understand the model's advantages.

Module 1. The embedding module for video feature acquisition, which utilizes a cross-modal attention mechanism to aggregate frame features.

Module 2. The global-local similarity-based computation module.

The comparison experiments were performed on the MSR-VTT dataset.

Cross-modal Feature Aggregation

Table 2 presents the results of the ablation study on the cross-modal feature aggregation module for video feature extraction. The different configurations for the ablation experiments are shown in the table.

In this set of experiments, we compare the performance of our cross-modal aggregation method with that of Mean Aggregation and Self-attention Aggregation. The Mean Aggregation method calculates an unweighted average of the frame feature embeddings, while the Self-attention Aggregation method computes aggregation weights without utilizing textual semantic information and aggregates the frame

Figure 5. Visualization of text-to-video search results on MSR-VTT. The first row is the query text, the second row is the corresponding Ground Truth. the third, fourth and fifth rows are the retrieval results for Top1–3.

features using a focused mechanism.

The results of these experiments, as shown in Table 3, reveal an improvement in $R@1$ values ranging from 1% to 6%. This indicates that our cross-modal attention-based approach to acquiring video features leads to a more accurate capture of the relationships between video frames and text semantics.

Global-Local Similarity Calculation

In the ablation experiments of the similarity calculation module, Table 3 demonstrates the impact of various strategies on similarity analysis and score prediction. The results indicate that using video features obtained from the cross-modal attention feature aggregation method (as outlined in Section 4.3) as input data for the similarity calculation module slightly decreases performance compared to using frame-word local features. This suggests two things: (1) the aggregation process may result in a loss of detailed features, and (2) the slight performance decrease also implies that the aggregated video features can effectively capture the features present in the frames. The global-local similarity

calculation approach leads to an improvement of 1–3% in $R@1$ compared to using either method individually.

Figure 3 displays the attention weights of selected video frames generated by the cross-modal feature aggregation model. As can be observed from the examples, the model's attention mechanism can distinguish the relative importance of each frame's content, assigning lower weights to frames with limited correlation to textual information. In comparison, the self-attention aggregation method can recognize frames with crucial information but fails to differentiate between frames with subtle differences. On the other hand, the mean weighting aggregation method doesn't differentiate between frame.

The line graph in Figure 4 showcases the trend of the weight assigned to the key frames of the first example shown in Figure 3. The results demonstrate that the cross-modal Attention mechanism effectively identifies the frames relevant to the critical information in the video as it assigns higher weights to these frames. On the other hand, the mean aggregation method presents a flat trend, with no significant fluctuations in the weight assignments. In comparison, the self-attention method appears less responsive to the changes in the frame content, leading to a more moderate trend in the graph.

5.3.3. Qualitative results

The results in Figure 5 show the effectiveness of the text-to-video model developed in this study. The first row displays the input query text, while the second shows the ground truth. The remaining rows (3–5) present each query's top 1–3 ranked results. The retrieved video frames are visually similar to the ground truth and semantically align with the given text query, demonstrating the ability of the model to match textual and visual information.

The first column in Figure 5 demonstrates the model's aptitude in retrieving videos accurately related to the query text. The query "doing craft", is reflected in the captions of the retrieved videos, all of which pertain to "craft" and feature a "woman". This indicates that the model can efficiently match text and video topics during retrieval. The second column showcases the model's focus on the critical elements shared between the text and video modalities, as the top-ranked retrieval result, despite not being the ground truth, contains the crucial information from the query, namely a "woman" and a "laptop". Similarly, both the top 2 and top 3 ranked videos in the last column depict a "student" and a "teacher" in a "classroom".

The utilization of cross-modal feature aggregation and global-local similarity calculation in the model elevates the accuracy and sophistication of text-to-video retrieval results. This allows the model to concentrate on the topics and visual aspects of the videos, resulting in a more precise and refined retrieval outcome.

6. Conclusions

This paper improves the performance of text-video matching by implementing two modules: the cross-modal attention feature aggregation module and the global-local similarity calculation module. The cross-modal attention feature aggregation module leverages the pre-trained CLIP model's multimodal feature extraction capabilities to extract highly relevant video features, focusing on the frames most pertinent to the text. Meanwhile, the global-local similarity calculation module calculates similarities based on the video-sentence and frame-word granularities, allowing for a more nuanced con-

sideration of both the topic and detail features in the matching process. The experimental results, conducted on the benchmark dataset, clearly demonstrate the efficacy of our proposed modules in capturing both topic and detail features, leading to improvement in text-video matching accuracy. This work contributes to multi-modal representation learning, highlighting the potential of advanced feature aggregation and similarity calculation techniques in enhancing text-video matching. Further research may be necessary to realize our methods in real-world applications fully.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors would like to acknowledge the support provided by Aerospace HongKa Intelligent Technology (Beijing) CO., LTD.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16x16 words: Transformers for image recognition at scale, preprint, arXiv: 2010.11929.
2. Y. Liu, S. Albanie, A. Nagrani, A. Zisserman, Use what you have: Video retrieval using representations from collaborative experts, preprint, arXiv: 1907.13487.
3. X. Wang, J. Wu, J. Chen, L. Li, Y. F. Wang, W. Y. Wang, Vatex: A large-scale, high-quality multilingual dataset for video-and-language research, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2019), 4581–4591. https://doi.org/10.1109/ICCV.2019.00468
4. L. Zhu, Y. Yang, Actbert: Learning global-local video-text representations, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 8746–8755. https://doi.org/10.1109/CVPR42600.2020.00877
5. F. C. Heilbron, V. Escorcia, B. Ghanem, J. C. Niebles, Activitynet: A large-scale video benchmark for human activity understanding, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 961–970. https://doi.org/10.1109/CVPR.2015.7298698
6. H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, et al., Univl: A unified video and language pre-training model for multimodal understanding and generation, preprint, arXiv: 2002.06353.
7. N. Shvetsova, B. Chen, A. Rouditchenko, S. Thomas, B. Kingsbury, R. S. Feris, et al., Everything at once-multi-modal fusion transformer for video retrieval, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 20020–20029. https://doi.org/10.1109/CVPR52688.2022.01939

8. N. C. Mithun, J. Li, F. Metze, A. K. Roy-Chowdhury, Learning joint embedding with multimodal cues for cross-modal video-text retrieval, in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, (2018), 19–27. https://doi.org/10.1145/3206025.3206064
9. J. Xu, T. Mei, T. Yao, Y. Rui, Msr-vtt: A large video description dataset for bridging video and language, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 5288–5296. https://doi.org/10.1109/CVPR.2016.571
10. Y. Yuan, T. Mei, W. Zhu, To find where you talk: Temporal sentence localization in video with attention based location regression, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **33** (2019), 9159–9166. https://doi.org/10.1609/aaai.v33i01.33019159
11. J. Dong, X. Li, C. Xu, S. Ji, Y. He, G. Yang, et al., Dual encoding for zero-example video retrieval, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 9346–9355. https://doi.org/10.1109/CVPR.2019.00957
12. M. Bain, A. Nagrani, G. Varol, A. Zisserman, Frozen in time: A joint video and image encoder for end-to-end retrieval, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 1728–1738. https://doi.org/10.1109/ICCV48922.2021.00175
13. A. Miech, D. Zhukov, J. B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic, Howto100m: Learning a text-video embedding by watching hundred million narrated video clips, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2019), 2630–2640. https://doi.org/10.1109/ICCV.2019.00272
14. S. Wang, R. Wang, Z. Yao, S. Shan, X. Chen, Cross-modal scene graph matching for relationship-aware image-text retrieval, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, (2020), 1508–1517. https://doi.org/10.1109/WACV45572.2020.9093614
15. F. Shang, C. Ran, An entity recognition model based on deep learning fusion of text feature, *Inf. Process. Manage.*, **59** (2022), 102841. https://doi.org/10.1016/j.ipm.2021.102841
16. Y. Zhou, R. Zhang, C. Chen, C. Li, C. Tensmeyer, T. Yu, et al., Towards language-free training for text-to-image generation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 17907–17917. https://doi.org/10.1109/CVPR52688.2022.01738
17. W. Li, S. Wen, K. Shi, Y. Yang, T. Huang, Neural architecture search with a lightweight transformer for text-to-image synthesis, *IEEE Trans. Network Sci. Eng.*, **9** (2022), 1567–1576. https://doi.org/10.1109/TNSE.2022.3147787
18. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, preprint, arXiv: 1301.3781.
19. F. Perronnin, C. Dance, Fisher kernels on visual vocabularies for image categorization, in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, (2007), 1–8. https://doi.org/10.1109/CVPR.2007.383266
20. B. Klein, G. Lev, G. Sadeh, L. Wolf, Associating neural word embeddings with deep image representations using Fisher vectors, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 4437–4446. https://doi.org/10.1109/CVPR.2015.7299073
21. R. Kiros, R. Salakhutdinov, R. S. Zemel, Unifying visual-semantic embeddings with multimodal neural language models, preprint, arXiv: 1411.2539.

22. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., Learning transferable visual models from natural language supervision, preprint, arXiv: 2103.00020.

23. I. Croitoru, S. V. Bogolin, M. Leordeanu, H. Jin, A. Zisserman, S. Albanie, et al., Teachtext: Crossmodal generalized distillation for text-video retrieval, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 11583–11593. https://doi.org/10.1109/ICCV48922.2021.01138

24. A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, Vivit: A video vision transformer, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 6836–6846. https://doi.org/10.1109/ICCV48922.2021.00676

25. G. Bertasius, H. Wang, L. Torresani, Is space-time attention all you need for video understanding?, *ICML*, **2** (2021), 4. https://doi.org/10.48550/arXiv.2102.05095

26. A. Miech, J. B. Alayrac, L. Smaira, I. Laptev, J. Sivic, A. Zisserman, End-to-end learning of visual representations from uncurated instructional videos, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 9879–9889. https://doi.org/10.1109/CVPR42600.2020.00990 .

27. X. Duan, W. Huang, C. Gan, J. Wang, W. Zhu, J. Huang, Weakly supervised dense event captioning in videos, *Adv. Neural Inf. Process. Syst.*, **31** (2018). https://doi.org/10.48550/arXiv.1812.03849

28. R. Tan, H. Xu, K. Saenko, B. A. Plummer, Logan: Latent graph co-attention network for weakly-supervised video moment retrieval, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, (2021), 2083–2092. https://doi.org/10.1109/WACV48630.2021.00213

29. V. Gabeur, C. Sun, K. Alahari, C. Schmid, Multi-modal transformer for video retrieval, in *Computer Vision–ECCV 2020: 16th European Conference*, (2020), 214–229. https://doi.org/10.48550/arXiv.2007.10639

30. M. Dzabraev, M. Kalashnikov, S. Komkov, A. Petiushko, Mdmmt: Multidomain multimodal transformer for video retrieval, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 3354–3363. https://doi.org/10.1109/CVPRW53098.2021.00374

31. L. Li, Z. Shu, Z. Yu, X. J. Wu, Robust online hashing with label semantic enhancement for cross-modal retrieval, *Pattern Recognit.*, **145** (2023), 109972. https://doi.org/10.1109/ICME55011.2023.00173

32. Z. Shu, K. Yong, J. Yu, S. Gao, C. Mao, Z. Yu, Discrete asymmetric zero-shot hashing with application to cross-modal retrieval, *Neurocomputing*, **511** (2022), 366–379. https://doi.org/10.1016/j.neucom.2022.09.037

33. Z. Shu, Y. Bai, D. Zhang, J. Yu, Z. Yu, X. J. Wu, Specific class center guided deep hashing for cross-modal retrieval, *Inf. Sci.*, **609** (2022), 304–318. https://doi.org/10.1016/j.ins.2022.07.095

34. M. Su, G. Gu, X. Ren, H. Fu, Y. Zhao, Semi-supervised knowledge distillation for cross-modal hashing, *IEEE Trans. Multimedia*, **25** (2021), 662–675. https://doi.org/10.1109/TMM.2021.3129623

35. J. A. Portillo-Quintero, J. C. Ortiz-Bayliss, H. Terashima-Marín, A straightforward framework for video retrieval using clip, in *Mexican Conference on Pattern Recognition*, (2021), 3–12. https://doi.org/10.1007/978-3-030-77004-4_1

36. M. Patrick, P. Y. Huang, Y. Asano, F. Metze, A. Hauptmann, J. Henriques, et al., Support-set bottlenecks for video-text representation learning, preprint, arXiv: 2010.02824.

37. Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, et al., Camp: Cross-modal adaptive message passing for text-image retrieval, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2019), 5764–5773. https://doi.org/10.1109/ICCV.2019.00586

38. H. Diao, Y. Zhang, L. Ma, H. Lu, Similarity reasoning and filtration for image-text matching, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **35** (2021), 1218–1226. https://doi.org/10.1609/aaai.v35i2.16209

39. Y. Liu, H. Liu, H. Wang, F. Meng, M. Liu, BCAN: Bidirectional correct attention network for cross-modal retrieval, *IEEE Trans. Neural Netw. Learn. Syst.*, (2023). https://doi.org/10.1109/TNNLS.2023.3276796

40. S. K. Gorti, N. Vouitsis, J. Ma, K. Golestan, M. Volkovs, A. Garg, et al., X-pool: Cross-modal language-video attention for text-video retrieval, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 5006–5015. https://doi.org/10.1109/CVPR52688.2022.00495

41. H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, J. Han, Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 12655–12663. https://doi.org/10.1109/CVPR42600.2020.01267

42. F. Faghri, D. J. Fleet, J. R. Kiros, S. Fidler, Vse++: Improving visual-semantic embeddings with hard negatives, preprint, arXiv: 1707.05612.

43. Y. Yu, J. Kim, G. Kim, A joint sequence fusion model for video question answering and retrieval, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 471–487. https://doi.org/10.1007/978-3-030-01234-2_29

44. A. Rohrbach, M. Rohrbach, N. Tandon, B. Schiele, A dataset for movie description, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 3202–3212. https://doi.org/10.1109/CVPR.2015.7298940

45. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.*, **15** (2014), 1929–1958.

46. Z. Xie, I. Sato, M. Sugiyama, Stable weight decay regularization, preprint, arXiv: 2011.11152.

47. S. Liu, H. Fan, S. Qian, Y. Chen, W. Ding, Z. Wang, Hit: Hierarchical transformer with momentum contrast for video-text retrieval, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 11915–11925. https://doi.org/10.1109/ICCV48922.2021.01170

48. J. Wang, Y. Ge, R. Yan, Y. Ge, K. Q. Lin, S. Tsutsui, et al., All in one: Exploring unified video-language pre-training, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 6598–6608. https://doi.org/10.1109/CVPR52729.2023.00638

49. J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, et al., Less is more: Clipbert for video-and-language learning via sparse sampling, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (2021), 7331–7341. https://doi.org/10.1109/CVPR46437.2021.00725

50. H. Fang, P. Xiong, L. Xu, Y. Chen, Clip2video: Mastering video-text retrieval via image clip, preprint, arXiv: 2106.11097.

51. J. Lei, T. L. Berg, M. Bansal, Revealing single frame bias for video-and-language learning, preprint, arXiv: 2011.11152.

52. L. Li, Z. Gan, K. Lin, C. C. Lin, Z. Liu, C. Liu, et al., Lavender: Unifying video-language understanding as masked language modeling, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 23119–23129. https://doi.org/10.1109/CVPR52729.2023.02214

53. H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, et al., Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning, *Neurocomputing*, **508** (2022), 293–304. https://doi.org/10.1016/j.neucom.2022.07.028

54. F. Cheng, X. Wang, J. Lei, D. Crandall, M. Bansal, G. Bertasius, Vindlu: A recipe for effective video-and-language pretraining, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 10739–10750. https://doi.org/10.1109/CVPR52729.2023.01034

AIMS Press