International Journal of Multimedia Information Retrieval (2024) 13:30
https://doi.org/10.1007/s13735-024-00335-7

REGULAR PAPER

LSECA: local semantic enhancement and cross aggregation for video-text retrieval

Zhiwen Wang[1] · Donglin Zhang[1] · Zhikai Hu[2]

Received: 27 September 2023 / Revised: 27 April 2024 / Accepted: 28 May 2024 / Published online: 22 July 2024

Abstract

Recently video retrieval based on the pre-training models (e.g., CLIP) has achieved outstanding success. To further improve the search performance, most existing methods usually utilize the multi-grained contrastive fine tuning scheme. For example, frame features and word features are taken as fine-grained representations, aggregate features for frame features and [CLS] token for textual side are used as global representations. However, the above scheme still remains challenging. There are redundant and noise information in the raw output features of pre-training encoders, leading to suboptimal retrieval performance. Besides, a video usually correlates several text descriptions, while video embedding is fixed in previous works, which may also reduce the search performance. To conquer these problems, we propose a novel video-text retrieval model, named Local Semantic Enhancement and Cross Aggregation (LSECA). To be specific, we design a local semantic enhancement scheme, which utilizes global feature for video and keyword information for text to augment fine-grained semantic representations. Moreover, the cross aggregation module is proposed to enhance the interaction between video and text modalities. In this way, the local semantic enhancement scheme can increase the related representation of modalities and the developed cross aggregation module can make the representations of texts and videos more uniform. Extensive experiments on three popular text-video retrieval benchmark datasets demonstrate that our LSECA outperforms several state-of-the-art methods.

Keywords Video-text retrieval · Semantic enhancement · Cross aggregation · Multi-grained contrast

1 Introduction

With the rapid development of mobile device and Internet, short videos are becoming more and more important in modern life. Therefore, Text to Video Retrieval (TVR), a typical multi-modal task, has drown increasing attention [1–6]. The aim of this task is to rank videos (or texts) within the collection based on their relevance to a specific text or video, which enables users to efficiently and precisely retrieve their desired video content. In the past few decades, with the ongoing advancement of deep learning technology, remarkable progress has been made in the field of video retrieval [7–13]. However, due to the heterogeneity of text and video modalities, how to reduce the modality gap and improve performance is still an open problem.

To narrow this modality gap, some good methods have emerged, among which the method based on fine-tuning the pre-training models [14–16] has gained widespread attention. CLIP4Clip [17] is a notable method that leverages the robust semantic extraction capabilities of pre-training model CLIP [14] to align video with text in a shared feature space, enabling a direct comparison of video and text features. Compared to previous works [1, 3–8, 18–20], this method yields superior results. However, this method focuses on global information while ignoring fine-grained information. To mitigate the problem, some excellent works [21–25] are proposed, which employ frame [26] and word features as local information (as shown in Fig. 1a). For example, X-CLIP [22] excavates local and global representations and leverages cross-grained, coarse-grained and fine-grained contrastive learning scheme to further improve retrieval performance.

✉ Donglin Zhang
zhangdlin@jiangnan.edu.cn

Zhiwen Wang
6223112033@stu.jiangnan.edu.cn

Zhikai Hu
cszkhu@comp.hkbu.edu.hk

[1] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, Jiangsu, China

[2] Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR 999077, China

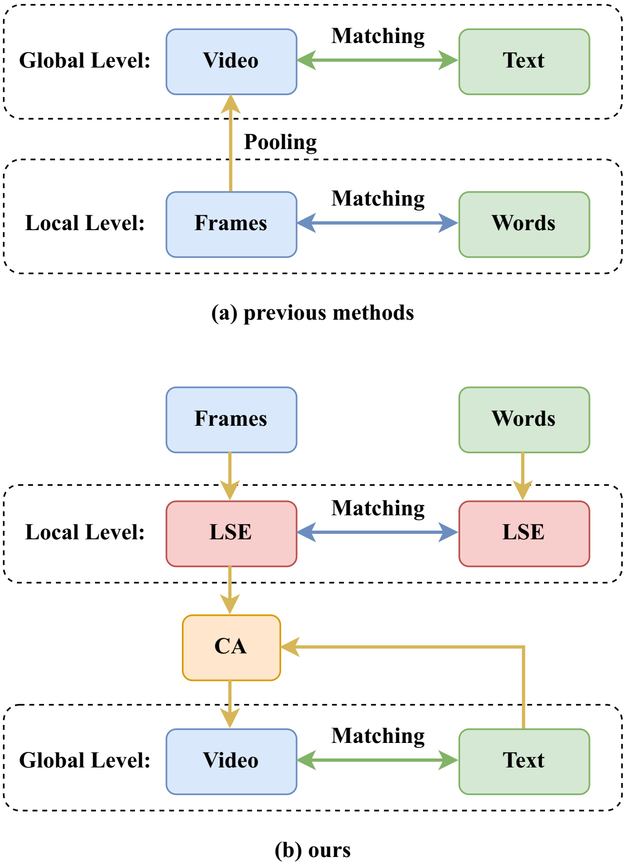

Fig. 1 Comparison of our proposed method with previous multi-grained contrastive methods. **a** The framework of previous methods (multi-grained retrieval), which uses the raw features of encoders for fine-grained matching. **b** Our LSECA framework semantically enhances the original fine-grained features via LSE (local semantic enhancement) modules, but also takes into account the correspondence between the two modalities to design the CA (cross aggregation) interaction module

However, the backbone model (e.g., CLIP [14]) is pre-trained with image-text pairs, the extracted raw frame and word features may not be suitable for video retrieval task, resulting in suboptimal search results.

Despite the above problems, the text-video retrieval task requires precise semantic alignment, but the multi-modal features output by the raw encoder have a lot of redundant and noisy information, which is a serious challenge for cross-modal matching. Specifically, with regard to video, certain frames in the sparsely sampled frames have high content overlap (i.e., redundant information), and a few frames have little semantic significance to the supplied text description (i.e., noisy information). For text modality, the given text contains prepositions (e.g., 'the', 'in', 'on', *etc*.) which may carry less semantic information than entities and verbs. Although these prepositions can help to understand the text, they have litte impact in the process of fine-grained matching and even lead to suboptimal performance. Besides, the video modality usually corresponds to multiple descriptions in the retrieval datasets [27–29], and videos often display more content than text. Therefore, how to conditionally filter frame features and enhance the interaction between the two modalities need to be addressed.

To solve the above problems, we develop a novel text to video retrieval model based on local semantic enhancement and cross aggregation, named LSECA. Previous approaches typically use more multimodal features [3] or more complex cross-granularity alignment [22] to achieve finer semantic matching. We use unimodal specific information (e.g., video global feature and keyword features) for local semantic enhancement. The proposed model can augment fine-grained semantic representation and facilitate interaction between video and text (as shown in Fig. 1b). Specifically, for video branch, we use pooled features as anchors to improve the semantics of fine-grained frame features. Besides, due to the uncertainty of frame content, we design an adapter-aware module to estimate the weight of each fine-grained representations. For text branch, we first extract the keywords of corresponding description with KeyBert [30] model and the fusion strategy between raw word and keyword representations is introduced to shift the focus more to semantic content. Local semantic enhancement is not enough and interaction between modalities is also required. Thus in the coarse-grained perspective, we further propose the cross aggregation module that fuses the frame-level features based on the text to enable interaction between two modalities. The above module can significantly improve the discimination of video and text representations. To demonstrate the effectiveness of the proposed LSECA, we conduct some extensive experiments on three mainstream text-video retrieval datasets, including MSRVTT [27], MSVD [28], and LSMDC [29]. The results illustrate that our proposed LSECA achieves significant improvment and outperforms several previous state-of-the-art methods.

The main contributions of this work are summarized as follows:

- We propose a novel framework LSECA for text-video retrieval, which not only enhances the fine-grained video and text representations but also fully considers the interaction between two modalities.
- For local semantic enhancement, we propose two effective strategies for video and text branches respectively. The cross aggregation module is introduced to achieve sufficient interaction between two modalities.
- Extensive experiments on three text-video retrieval datasets demonstrate the effectiveness of our method. Our LSECA achieves state-of-the-art performance on MSRVTT (47.1%), MAVD (46.9%), and LSMDC (23.4%).

2 Related work

2.1 Text to video retrieval

With the promotion and popularization of short video applications, accurate video similarity search is becoming increasingly important. Text-video retrieval task aims to find the most relevant video based on the given text information. However, unlike text-image cross-modal task, text-video retrieval task need to consider temporal information, making the retrieval task more diffcult. Some early approaches [3, 4, 6, 7, 11, 18, 31–34] extracted video as well as text features by using convolutional neural networks or experts. Despite these approches have demonstrated favorable outcomes, the performance of these methods is still limited due to end-to-end optimization issue. With the continuous development of the pre-training models (e.g., CLIP [14], ALIGN [15], CoCa [16], *etc*), the paradigm [17] of end-to-end video retrieval by fine-tuning models directly from raw video (or text) has gained a lot of attention. Numerous pretty works [17, 21, 22, 35–37] utilize the semantic extraction ability of CLIP learned from 400 M image-text pairs to adapt to video retrieval task. CLIP4Clip [17] leverages the knowledge obtaiend from the CLIP model and applies to the task of video retrieval. By employing the contrastive learning to compute the similarity scores, it achieves good performance and establishes a strong baseline for future research endeavors. Based on CLIP4Clip, CLIP2Video [36] proposes the temporal difference block and temporal alignment block to enhance the optimization of video and text representations. However, the above methods only use the global feature for contrastive learning, ignoring the semantic information and lack the interaction between two modality. Different from these approahces, we not only use fine-gained feature to improve the performance but also design the cross aggregation module to enhance the interaction between video and text.

2.2 Multi-grained representation learning

In recent years, there has been a proliferation of valuable studies [21, 22, 32, 33, 38–40] that employ multi-grained video and text representations to enhance retrieval performance. Concretely, for the text branch, the common approach [22] is to treat word embeddings as fine-grained features and [CLS] token as global feature. For the video branch, traditional methods [32, 33] utilize task-specific networks or experts to extract different types of features (*e.g.*, object, action, scene, audio, *etc*). However, these specific features extracted cannot be well adapted to retrieval task due to the end-to-end optimization issue. For example, T2VLAD [33] achieves better retrieval results by aligning local features though NetVlad and aligning global feature though aggregation. Recent works extract frame features as fine-grained representation of video by using image encoder due to the rapid development of pre-training image-language model. For instance, TS2-Net [21] proposes the token shift module to capture temporal movements and the token selection module to select tokens that contribute most to fine-grained semantic information. X-CLIP [22] presents the multi-grained contrastive learning to better utilize more semantic information for improving retrieval performance.

The above methods utilzie the output of the raw encoders for contrastive learning. However, the performance may be limited due to the heterogeneity between the video and image. In this paper, we utilize local semantic enhancement module to improve the video and text fine-grained representations and design the cross aggregation to enhance the interaction between two modalities, resulting in notable enhancement in retrieval performance.

3 Method

In this section, we detail the proposed LSECA, along with providing specific details regarding the text to video retrieval task. Concretely, in the retrieval task, given a set of descriptions and the same number of video clips, our goal is to obtain a semantic similarity matrix for the purpose of retrieving the videos. The architecture of LSECA is shown in Fig. 2. In Sect. 3.1, we firstly introduce the basic preliminary which consists mainly of the extraction of text and video features and the symbolic representation of each feature. We then elaborate the details of the proposed Local Semantic Enhancement module in Sect. 3.2, which consists of two parts, the video branch in Sect. 3.2.1 and the text branch in Sect. 3.2.2, respectively. Immediately following the Cross Aggregation module in Sect. 3.3. Finally, in Sect. 3.4, we describe the calculation of the multi-grained similarity and the objective function for optimization.

3.1 Preliminary

In general, given a set of video-text pairs $(\boldsymbol{V}, \boldsymbol{T})$ as the input data. For the video branch, we sample the video frames uniformly for a video, usually at a sampling rate of 1 frame per second. We use the image encoder of CLIP [14] which is a vision transformer architecture and initialized by the public checkpoints of ViT-B/32 to process frame image. Specifically, the frame image firstly is divided into multiple patches, add [CLS] token and position tokens which make encoder better extract semantic information from image. Finally, The [CLS] tokens from the last transformer layer are extracted as the frame-level features $\boldsymbol{f} = \left\{f_1, f_2, f_3, ..., f_{N_f}\right\}$, and N_f is the number of frames in the video. For a description $\boldsymbol{t} \in \boldsymbol{T}$, similar to the video side, the text encoder of CLIP is used to extract text features, the architecture of text encoder is also a

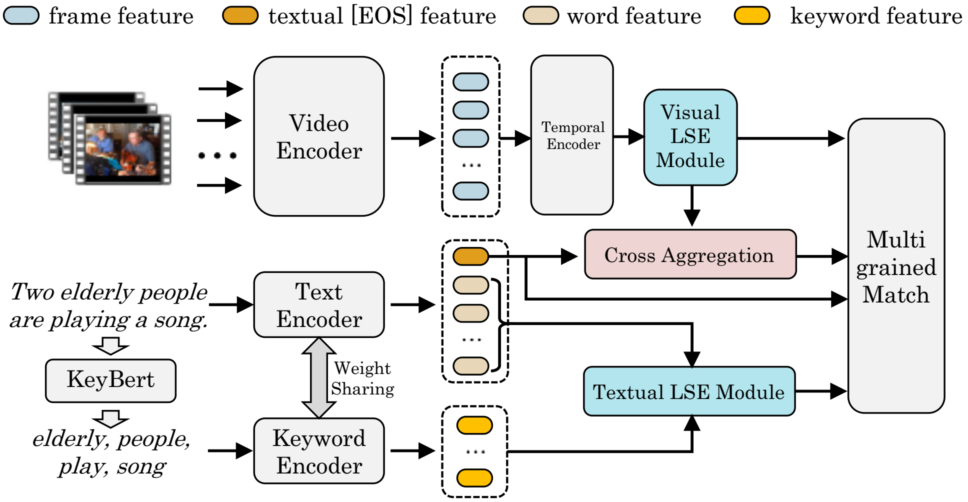

Fig. 2 An overview of our LSECA for text-video retrieval. In LSECA, We first extract the keywords of the given text description via the KeyBert [30] TransFormer. And we design two different local semantic enhancement(LSE) schemes for text and video, respectively. With the help of video representations and keywords, thus obtaining fine-grained representations that are richer and more compact in semantic information. In addition, to enhance the interaction between the two modalities, we propose the corss aggregation(CA) module

multi-layer transformer. We firstly split the given text description to word sequence by using the specific tokenizer [14]. Before being fed into the text encoder, the word sequence is padded with [BOS] and [EOS] tokens at the start and end of the sequence, respectively. Finally, the global textual feature t_{EOS} and word-level features $\boldsymbol{w}=\left\{w_1, w_2, w_3, ..., w_{N_w}\right\}$ are the output of the [EOS] token and corresponding word tokens from the final layer of the textual transformer, where N_w is the length of the description.

3.2 Local semantic enhancement

3.2.1 Visual fine-grained representation

Firstly, we need to obtain the feature of the entire video based on the obtained frame features in Sect. 3.1. However, frame-level features are extracted from separate frames without considering the temporal interaction among frames, which only include spatial features of each single frame. And it is essential to be able to understand the video content correctly. Therefore, we follow the [17] and utilize the temporal transformer to model the temporal relationships between frames. Specifically, we add a position token for each frame feature before fed into model and the outputs of the temporal transformer are average pooled to obtain final video-level features, which can be formulate as:

$$f_i^{'} = TransEnc(f_i + p_i), \tag{1}$$

$$\boldsymbol{v} = \frac{1}{N_f}\sum_{i}^{N_f} f_i^{'}, \tag{2}$$

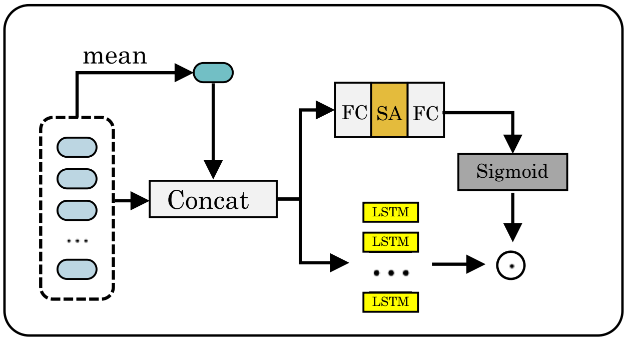

Fig. 3 The details of visual LSE module. We enhance the frame features by utiling the pooled feature and design adapter-aware module to adjust the enhanced features

where $\boldsymbol{p}$ is the added position tokens for frame features $\boldsymbol{f}$, N_f is the number of sampled frames, and $\boldsymbol{v}$ is the final video-level feature.

Earlier TVR works mainly focus on fine-grained and coarse-grained contrastive learning, which compute similarity using the raw output of CLIP encoder. However, the output of raw encoder may not be well suited for video retrieval task due to the heterogeneity between the video and image. To this end, we develop the visual semantic enhancement module in the proposed LSECA, which differs from prior approaches.

The video frames obtained by uniform sampling contain a lot of redundant information, which is detrimental to cross-modal matching. Therefore, we adjust the local frame features from a global perspective thereby achieving semantic enhancement. As shown in Fig. 3, given the video-level feature $\boldsymbol{v}$ and frame-level features $\boldsymbol{f}^{'}=\left\{f_1^{'}, f_2^{'}, f_3^{'}, ..., f_{N_f}^{'}\right\}$, we first concatenate global video feature $\boldsymbol{v}$ with each frame-level feature f_i, generating the input of the visual semantic enhancement module $\hat{f}_i = [\boldsymbol{v}, f_i^{'}]$. Moreover, we utilize the LSTM as main part of the visual local semantic enhancement module to generatea sequence of global-guide frame embeddings $\boldsymbol{f}^g = \left\{f_1^g, f_2^g, f_3^g, ..., f_{N_f}^g\right\}$. In addition, The information across frame and video is partially matched, and it is not appropriate to treat them equally. Thus, we propose the adapter-aware module to adjust the enhanced features and reduce the impact on the final similarity calculation. The whole process can be formulated as:

$$f_i^g = LSTM(\hat{f}_i) \cdot W_a, \tag{3}$$

where f_i^g is the fine-grained feature after visual semantic enhancement, W_a is the weights estimated by the adapter-aware module, which add soft labels to video fine-grained features to filter out unnecessary frames by comparing each frame with its video context. To be specific, as shown in Fig. 3, the adapter-aware module consists two linear FC layers, a self-attention layer, and a sigmoid activation function layer to calculate the corresponding weights. In this case, the self-attention layer is able to provide a global view of the fine-grained features, and the sigmoid layer can generate smooth adaptive weights for these features in the end.

3.2.2 Textual fine-grained representation

Due to the difference between two modalities, it is not advisable to adopt the same enhancement strategy as the video branch. In real video retrieval scenarios, we usually focus more on words with more semantic information, such as entities, actions, scenes, and so on. In light of this, we propose the local semantic enhancement strategy that relies on the utilization of keyword-guide. Specifically, we utilize the KeyBert [30] transformer to extract keywords of corresponding textual description, which is an effective and easy-to-use keyword extraction technique that leverages BERT embeddings to create keywords and keyphrases. The formulation can be represented as follows:

$$\boldsymbol{w}^{'} = KeyEnc(KeyBert(\boldsymbol{t})), \tag{4}$$

where $\boldsymbol{w}^{'}$ is the keyword features and $KeyEnc(\cdot)$ is the keyword encoder that is a standard transformer encoder with 12 layers and 8 attention heads the same structure as text encoder of CLIP [14]. With the exception of the final linear projection layer, the weight parameters are shared between keyword and text encoders.

For the textual local semantic enhancement module, we design a cross-attention strategy for raw word features $\boldsymbol{w} = \left\{w_1, w_2, w_3, ..., w_{N_w}\right\}$ and extracted features features $\boldsymbol{w}^{'} = \left\{w_1^{'}, w_2^{'}, w_3^{'}, ..., w_{N_k}^{'}\right\}$, where N_k is the number of keywords in the description. As shown in Fig. 4a, the word features are the keys and values and the keyword features are the queries, and it can be formulated as follows:

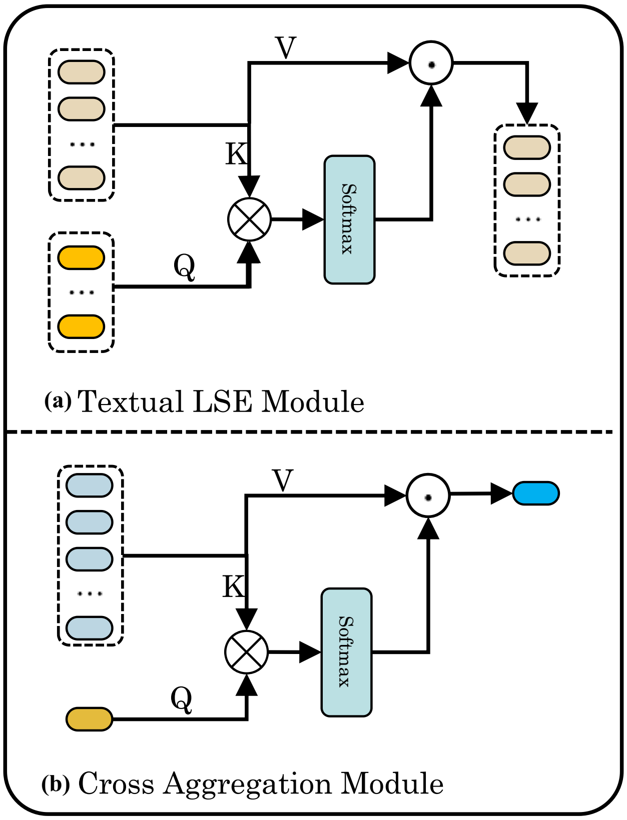

Fig. 4 The details of textual LSE module and cross aggregation module. The core algorithm of both modules is the cross-attention mechanism. Enhanced text features and text-guided video aggregation features are obtained through the guidance of keywords as well as text features, respectively

$$\boldsymbol{w}^k = CrossAtten(\boldsymbol{w} \cdot W_K, \boldsymbol{w} \cdot W_V, \boldsymbol{w}^{'} \cdot W_Q), \tag{5}$$

where $CrossAtten(\cdot)$ is the cross-attention mechanism which dynamically assigns the importance of different elements of the input features based on the relationship between two modalities, thus better capturing their interdependencies. And W_Q, W_K and W_V are trainable projection matrices and $\boldsymbol{w}^k$ is the enhanced text fine-grained representations. Finally, semantic content can be enhanced in term of word features through keyword guidance.

3.3 Cross aggregation

In addition to local semantic enhancement, we also consider the interaction between two modalities. Upon examination of the retrieval dataset, we find that a video typically corresponds to multiple descriptions. Besides, the video embedding is fixed in previous approaches, whereas the semantic focus of the different related texts is different. Therefore, merely relying on global video features is insufficient for effectively cross-modal retireval.

Inspired by [37], we use the similar cross aggregation module for enhanced frame features $\boldsymbol{f}^g$ to interact with specific text feature $\boldsymbol{t}_{EOS}$. Specifically, as shown in Fig. 4b, the enhanced frame features are projected as key and value by two different linear layer and the query are the result of the projection of text feature $\boldsymbol{t}_{EOS}$. Finally, the output of cross aggregation module is the text-guided video aggregated feature v_{ca}, which can be formulated as follows:

$$\hat{\boldsymbol{v}} = CrossAtten(\boldsymbol{f}^g \cdot W_K^{'}, \boldsymbol{f}^g \cdot W_V^{'}, \boldsymbol{t}_{EOS} \cdot W_Q^{'}), \quad (6)$$

$$\boldsymbol{v}_{ca} = LN_1(LN_2(\hat{\boldsymbol{v}}) + Dropout(\hat{\boldsymbol{v}})), \quad (7)$$

where W_Q, W_K and W_V are trainable projection matrices. Similar to the Sect. 3.2.2, $LN_1(\cdot)$ and $LN_2(\cdot)$ are the LayerNorm layers. Besides, Dropout is a dropout layer, which not only makes train more stable but also prevents overfitting risk.

3.4 Multi-grained similarity and objective function

After the above features processing, we obtain enhanced frame features $\{f_i^g\}_{i=1}^{N_f}$, enhanced words features $\{w_i^k\}_{i=1}^{N_w}$, aggregated video feature $\boldsymbol{v}_{ca}$, and raw text feature $\boldsymbol{t}_{EOS}$. For coarse-grained similarity calculation, we directly use matrix multiplication between the video feature $\boldsymbol{v}_{ca}$ and text feature t_{EOS}, which can be represented as follows:

$$S_{coarse} = (\boldsymbol{v}_{ca})^{\top} \cdot \boldsymbol{t}_{EOS}. \quad (8)$$

For fine-grained similarity calculation, the fine-grained embeddings of video is the enhanced frame features $\boldsymbol{f}^g = \{f_i^g\}_{i=1}^{N_f}$, where the N_f is the number of sampled frames. The fine-grained embeddings of text is the enhanced text features $\boldsymbol{w}^k = \{w_i^k\}_{i=1}^{N_k}$. Following the [41], we calculate a similarity matrix which is defined as $A = [a_{i,j}]^{N_f \times N_k}$, where $a_{i,j}$ is computed from the cosine similarity of f_i^g and w_j^k and represents the fine-grained similarity score between the f_i^g and the w_j^k. Beisides, we choose the maximum value $\max\limits_j a_{ij}$ and $\max\limits_i a_{ij}$ in each row and column as the score that each fine-grained feature contributes to the final similarity calculation. At the same time we use the computed adaptive weights to pool the corresponding scores over all frames and words. Finally, it can be formulated as:

$$S_{fine} = \frac{1}{2}\left(\sum_{i=1}^{N_f} \omega_f^i \max_j a_{i,j} + \sum_{j=1}^{N_k} \omega_t^j \max_i a_{i,j}\right), \quad (9)$$

where $[\omega_f^0, \omega_f^1, ..., \omega_f^{N_f}] = \Phi(f^g)$ and $[\omega_t^0, \omega_t^1, ..., \omega_t^{N_k}] = \Psi(w^k)$ are the corresponding weights of the video frames and text words and they facilitate fine-grained cross-modal alignment. Specifically, $\Phi(\cdot)$ and $\Psi(\cdot)$ have the same structure, both consist of an FC layer and a Softmax layer. The first term of the whole equation is to represent the video to text retrieval similarity and the second term is opposite. Therefore, the final similarity score $\boldsymbol{S}$ of LSECA contains multi-grained contrastive similarity scores, which can be represented as follows:

$$\boldsymbol{S} = \alpha S_{coarse} + (1 - \alpha) S_{fine}, \quad (10)$$

where α is the trade-off hyper-parameter of total similarity, S_{coarse} and S_{fine} are the global and local similarity, respectively.

Based on the above schemes, given a batch of B text-video pairs, our LSECA can calculate a $B \times B$ similarity matrix during the training process. The InfoNCE loss based the similarity matrix is used to optimize the whole LSECA, which can be formulated as:

$$\mathcal{L}_{v2t} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(S_{v_i,t_i}/\tau)}{\sum_{j=1}^{B} \exp(S_{v_i,t_j}/\tau)}, \quad (11)$$

$$\mathcal{L}_{t2v} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(S_{v_i,t_i}/\tau)}{\sum_{j=1}^{B} \exp(S_{v_j,t_i}/\tau)}, \quad (12)$$

$$\mathcal{L} = (\mathcal{L}_{v2t} + \mathcal{L}_{t2v})/2, \quad (13)$$

where B is the pre-set batch size, τ is a temperature hyper-parameter which makes the training process converge more rapidly. The loss function $\mathcal{L}$ is utilized to increase the similarity of the positive pairs and decrease the similarity of the negative pairs, thereby shortening the distance between relevant video-text representations and separating the irrelevant text-video representations during the training process.

4 Experiments

4.1 Experimental settings

We conduct experiments on three mainstream text to video retrieval datasets to demonstrate the effectiveness of our LSECA.

MSRVTT [27] contains about 10K YouTube video clips, each with 20 caption descriptions. The duration of each video clip in this collection varies between 10 and 40 s. Following the dataset splits from [3], we train models with associated captions on the Training-9K set and report results on the test 1K-A set.

MSVD [28] contains 1,970 videos with 80K captions, with about 40 captions on average per video. Videos tend to be 40 s or less in length.There are 1,200, 100, and 670 videos in the train, validation, and test sets, respectively. The training as well as the inference on this dataset is in multi-sentence mode, which is slightly different from the other two datasets and can be found in the source code.

LSMDC [29] contains 118,081 videos and captions, which are extracted from 202 movies. The length of each video ranges from 2 to 30 s. We follow the split of [3] and there are 109,673, 7,408, and 1,000 videos in the train, validation, and test sets, respectively.

Evaluation Metrics To evaluate the performance of our proposed LSECA, we choose recall at Rank K (R@K, higher is better), median rank (MdR, lower is better), and mean rank (MnR, lower is better) as retrieval performance metrics. To be specific, R@K refers to the percentage of the first *K* retrieved videos that correspond to the text description among all the videos to be retrieved, i.e., the ability of the model to find the target video during retrieval. Referring to previous work [17], we use R@1, R@5 and R@10 as specific recall metrics As a result, the higher R@K indicates better performance. Median Rank (MdR) is the median retrieved rank of the ground truth. Similarly, Mean Rank (MnR) is the mean retrieved rank of the ground truth. Thus, the lower MdR and MnR indicate better performance. In addition, we added SumR (R@1+R@5+R@10) as a composite metric.

Implementation Details We conduct extensive experiments on 2 NIVIDIA GeForce RTX 3090 24GB GPUs using PyTorch library. Following the previous work [17], we initialize the text encoder and video encoder by using the public CLIP checkpoint (ViT-B/32). The frame sampling rate of videos is 1 FPS. The text description length is set to 32, the video length is set to 12 for all datasets and the number of keywords is 5. The initial learning rate for text encoder and video encoder of CLIP is 1e-7 and the initial learning rate for other modules is 1e-4. Then we decay the learning rate using the cosine schedule strategy and use the Adam optimizer to optimize the whole model. We train the model for 5 epochs with above settings and set the temperature τ is 0.01. We conduct ablation, comparison and qualitative experiments on the MSR-VTT dataset, which is more popular and competitive compared with other datasets.

4.2 Comparison with state-of-the-art

In this subsection, we compare the proposed LSECA with the previous state-of-the-art (SOTA) works on the three datasets, namely MSRVTT, MSVD and LSMDC. Results of the experiments on these datasets are presented in Tables 1, 2, 3. We can see that the LSECA obtains significant improvement on all three datasets. Furthermore, Table 1 shows the retrieval results of our method and comparisons with other SOTA model on MSRVTT 1K. To be specific, compared to the baseline CLIP4Clip-seqTransf [17], LSECA obtains 47.1 R@1 (5.9% improvement) and gets higher performance in all other metrics(e.g., 74.9 R@5, 5.0% improvement) in text to video retrieval with ViT-B/32 checkpoint. Comparing with other SOTA methods we also get the highest SumR. Therefore, our LSECA has significantly improved compared to the baseline [17], and also obtains competitive performance compared to other SOTA models. Tables 2 and 3 show results for the MSVD dataset and LSMDC dataset, respectively. Our LSECA also achieves good performance improvement compared to CLIP4Clip-seqTransf [17], which demonstrates the effectiveness and generalization ability of our proposed LSECA. The proposed LSECA can achieve good performance may be attributed to the following reasons:

- We optimize the fine-grained features compared to some previous works [21, 22, 32, 35–37]. For the video side, we utilize the video representations to semantically enhance the frame features so as to filter out irrelevant information and make the corresponding local information more prominent. For the text side, we process word features with KeyBert [30] extracted keywords as anchors to reduce the impact of irrelevant words on retrieval performance.
- We consider the uncertain matching problem between text and video, that is, video usually corresponds to multiple text descriptions, and a single text can only correspond to a portion of the elements in the video. Thus we design the cross aggregation module to alleviate this problem well, so as to obtain good performance.

4.3 Ablation study

In this section, we provide detailed ablation studies to further clarify the effects of each part of our design. The MSRVTT dataset is selected as the testbed, the results and analyses are as follows.

Table 1 Retrieval performance comparison on the MSR-VTT 1K validation set

Model	Text-to-video retrieval					Video-to-text retrieval					SumR
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	
CE [4]	20.9	48.8	62.4	6.0	28.2	20.6	50.3	64.0	5.3	25.1	267.0
ClipBERT [5]	22.0	46.8	59.9	6.0	–	–	–	–	–	–	–
MMT [3]	26.6	57.1	69.6	4.0	24.0	27.0	57.5	69.7	3.7	21.3	307.5
Frozen [6]	31.0	59.5	70.5	3.0	–	–	–	–	–	–	–
HiT [7]	30.7	60.9	73.2	2.6	–	32.1	62.7	74.1	3.0	–	333.7
BridgeFormer [8]	37.6	64.8	75.1	–	–	–	–	–	–	–	–
TMVM [18]	36.2	64.2	75.7	3.0	–	34.8	63.8	73.7	3.0	–	348.4
CLIP4Clip-MeanP [17]	43.1	70.4	80.8	2.0	16.2	43.1	70.5	81.2	2.0	12.4	389.1
CLIP4Clip-seqLSTM [17]	42.5	70.8	80.7	2.0	16.7	42.8	71.0	80.4	2.0	12.3	388.2
CLIP4Clip-seqTransf [17]	44.5	71.4	81.6	2.0	15.3	42.7	70.9	80.6	2.0	11.6	391.7
CenterCLIP [35]	44.2	71.6	82.1	2.0	15.1	42.8	71.7	82.2	2.0	10.9	394.6
CAMoE [32]	44.6	72.6	81.8	2.0	13.3	45.1	72.4	83.1	2.0	10.0	399.6
CLIP2Video [36]	45.6	72.6	81.7	2.0	14.6	43.5	72.3	82.1	2.0	10.2	397.8
X-Pool [37]	46.9	72.8	82.2	2.0	14.3	–	–	–	–	–	–
X-CLIP [22]	46.1	73.0	83.1	2.0	13.2	46.8	73.3	84.0	2.0	9.1	406.3
TS2-Net [21]	47.0	74.5	83.8	2.0	13.0	45.3	74.1	83.7	2.0	9.2	408.4
UCOFIA [42]	47.1	74.3	–	–	–	–	–	–	–	–	–
TeachCLIP [43]	46.8	**74.9**	82.9	2.0	–	–	–	–	–	–	–
MSIA [44]	47.2	73.8	**84.1**	2.0	–	–	–	–	–	–	–
PromptSwitch [45]	**47.8**	73.9	82.2	2.0	14.1	46.0	74.3	**84.8**	2.0	**8.5**	409.0
UATVR [46]	47.5	73.9	83.5	2.0	**12.3**	46.9	73.8	83.8	2.0	8.6	409.4
LSECA (Ours)	47.1	**74.9**	82.8	**2.0**	14.9	**47.5**	**75.4**	83.4	**2.0**	12.3	**411.1**

The best results for each evaluation metrics are in bold

"↑" denotes that higher is better. "↓" denotes that lower is better. And the CLIP4Clip-seqTransf is the baseline

Table 2 Results of text-to-video retrieval on the MSVD

Model	Text-to-video retrieval				
	R@1	R@5	R@10	MdR	MnR
NoiseEst [47]	13.1	35.7	47.7	12.0	–
CE [4]	19.8	49.0	63.8	6.0	–
Multi Cues [48]	20.3	47.8	–	–	–
Frozen [6]	33.7	64.7	76.3	3.0	–
SUPPORT [49]	28.4	60.0	72.9	4.0	–
TT-CE+ [50] 21.6	48.6	62.9	6.0	–	
CLIP [14]	37.0	64.1	73.8	3.0	–
TMVM [18]	36.7	67.4	81.3	2.5	–
BridgeFormer [8]	43.6	73.9	84.9	–	–
CLIP4Clip [17]	45.2	75.5	84.3	2.0	10.3
CAMoE [32]	46.9	76.1	85.5	2.0	9.9
TS2-Net [21]	44.6	75.8	–	2.0	–
PromptSwitch [45]	46.3	75.8	–	2.0	–
LSECA	46.9	76.8	85.7	2.0	9.9

Table 3 Results of text-to-video retrieval on the LSMDC

Model	Text-to-video retrieval				
	R@1	R@5	R@10	MdR	MnR
CT-SAN [51]	5.1	16.3	25.2	46.0	–
CE [4]	11.2	26.9	34.8	25.3	96.8
Frozen [6]	15.0	30.8	39.8	20.0	–
STG [52]	10.3	23.1	33.9	28.0	65.9
JS-Fusion [53]	9.1	21.2	34.1	36.0	–
MMT [3]	12.9	29.9	40.1	19.3	75.0
HiT [7]	14.0	31.2	41.6	18.5	–
BridgeFormer [8]	17.9	35.4	44.5	15.0	–
CLIP4Clip [17]	22.6	41.0	49.1	11.0	61.0
CAMoE [32]	22.5	42.6	50.1	–	56.9
TS2-Net [21]	23.4	42.3	50.1	10.0	56.9
QB-Norm [54]	22.4	40.1	45.9	11.0	–
LSECA	23.4	43.1	50.4	10.0	56.0

4.3.1 Ablation about components

To validate the effectiveness of each component, we conduct the ablation experiments with the 1k-A test split on the MSR-VTT 1K validation set. The results are shown in Table 4, and we obtain some important observations: We first investigate the impact of Visual Local Semantic Enhancement (VLSE) module. The global video embedding is utilized

Table 4 Component-wise evaluation of our framework on the MSR-VTT 1K validation set

Method			Text-to-video retrieval					Video to text retrieval				
			R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Baseline			44.5	71.4	81.6	2.0	15.3	42.7	70.9	80.6	2.0	11.6
+ fine-grained match			45.8	72.6	83.0	2.0	13.6	45.3	73.0	82.2	2.0	11.5
VLSE	TLSE	CA										
✓			46.7	74.3	82.5	2.0	12.8	45.6	73.6	82.6	2.0	11.3
	✓		46.5	72.9	82.8	2.0	14.0	45.8	74.0	82.8	2.0	11.1
✓	✓		47.0	73.6	83.6	2.0	13.6	45.4	72.8	83.1	2.0	11.8
✓	✓	✓	47.1	74.9	82.8	2.0	14.9	47.5	75.4	83.4	2.0	12.3

The baseline method is the CLIP4Clip-seqTransf [17], which only use the global features to calculate similarity between video and text. The "+ fine-grained match" is to retrieve videos based on the baseline using the original fine-grained features following the same fine-grained similarity computation method to validate the effectiveness of our LSECA components

to assist frame-level features for obtaining more semantic information. Similar to obtaining a synopsis, we use it to adaptively enhance the semantic information of the sampled frame features, which can be associated with entities, actions, backgrounds, and other information in the synopsis. From the experimental results in Table 4, it can be seen that our proposed enhancement module significantly improves the retrieval performance. Furthermore, we conduct experiment to testify the impact of Textual Local Semantic Enhancement (TLSE) module. In a real-world scenario, for videos, we tend to summarize their content, but for text, we are more inclined to extract its key points due to the heterogeneity between the two modalities. Therefore, we extract keywords to guide fine-grained features of the text towards the semantic center. Not disappointing our expectations, the experimental results in Table 4 also fully support our design. In addition, we simultaneously enhance both video and text features to achieve better retrieval performance. Finally, due to the fact that videos represent more content than text, to retrieve more accurately videos, we also consider the interaction between two modalities and propose the Cross Aggregation (CA) module based on the corresponding text. The results show that our model achieves the better performance. It demonstrates that the three parts are beneficial for semantic enhancement and cross-modality interaction.

4.3.2 Effect of the number of keywords

The **k** controls the size of keywords features $\{w'_1, w'_2, w'_3, ..., w'_{N_k}\}$. We start with a small size and increase it to large ones. In Table 5, overall performance improves and then decreases. By analysed, on the one hand, we find that fewer keywords limit the ability to enhance fine-grained features. On the other hand, the guidance ability of keywords decreases as the size increases. From Table 5, We set the keywords size k = 5 to achieve the better performance in practice.

Table 5 Ablation studies for the number of Keywords k on the MSR-VTT 1K validation set

The number	Text-to-video retrieval				
	R@1	R@5	R@10	MdR	MnR
k = 1	45.7	73.1	83.1	2.0	14.3
k = 3	46.3	73.8	82.4	2.0	15.1
k = 5	47.1	74.9	82.8	2.0	14.9
k = 7	46.9	74.2	82.6	2.0	14.6

4.3.3 Effect of different visual local semantic enhancement strategy

As shown in Fig. 5, we design four fusion schemes for frame-level semantic enhancement. To investigate the effect of the four fusion structures, i.e. "Sum + Trans", "Concat + Trans", "Sum + LSTM" and "Concat + LSTM" on retrieval performance, we perform some ablation experiments to compare them with each other in Table 6. "Sum" means that each fine-grained frame feature is added to the global video feature to obtain the combined features. "Concat" denotes to cascade the frame features with the video feature to obtain a longer feature which the dimension is 1024. "Trans" and "LSTM" means the fusion network structures which can fuse processed features thus achieving local semantic enhancement. From Table 6, we summarize the following observations: 1) When we use simple approach, Sum, compared to the Concat Obtains poor retrieval results. This may be because our goal is to use global features as anchors to guide semantic enhancement, yet the Sum operation causes two features of the same dimension to be confused together, damaging the original semantic information, resulting in poor retrieval performance compared to Concat. 2) From Table 6, we can also find that using LSTM for semantic enhancement can achieve better retrieval results compared to Transformer. By analyzing LSTM, Transformer, and input features, concating the

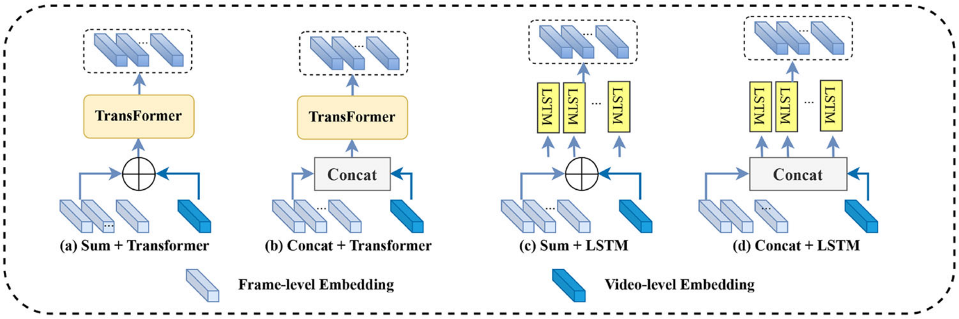

Fig. 5 Illustration of four fusion strategies. "Sum" and "Concat" represent the combination solutions between fine-grained frame features and global video feature. "LSTM" and "TransFormer" the two feature fusion architectures we apply. The effects of different combination solutions and fusion architectures on local semantic enhancement are analyzed experimentally

Table 6 Ablation studies for the different visual local semantic enhancement strategies on the MSR-VTT 1K validation set

The number	Text-to-video retrieval				
	R@1	R@5	R@10	MdR	MnR
Baseline	44.5	71.4	81.6	2.0	15.3
Sum + Trans	44.7	71.1	81.2	2.0	15.4
Concat + Trans	46.3	72.8	82.6	2.0	15.1
Sum + LSTM	45.1	72.2	82.1	2.0	14.2
Concat + LSTM	47.1	74.9	82.8	2.0	14.9

Table 7 Ablation studies for the adapter-aware module and the adaptive weights on the MSR-VTT 1K validation set

Method	Text-to-video retrieval		
	R@1	R@5	R@10
w/o adapter-aware module	46.9	72.4	82.8
w/o $\Phi(\cdot)$ and $\Psi(\cdot)$	46.5	73.1	82.6
LSECA	47.1	74.9	82.8

features can enhance the increase in frame level feature similarity, while Transformer, based on key, Query, and Value for temporal interaction. As a result, the results of the semantic enhancement effect does not align with initial expectations, and it is not as effective as LSTM. In summary, the experimental results show that proper fusion strategy between video and frame features can obtain better fine-grained representations.

4.3.4 Effect of the adapter-aware module and the adaptive weights

In Table 7, we testify the impact of the adapter-aware module in visual LSE module and adaptive weights in the Eq. 9. There are decreases in overall performance after removal. Specifically, after removing the adapter aware module, R@5 decrease from 74.9% to 72.4%. And without adaptive weights, R@1 also decreased by 0.6%. Therefore, these two parts are helpful for representation learning as well as cross-modal alignment.

Table 8 Ablation studies for the cross-attention module in textual LSE module on the MSR-VTT 1K validation set

Method	Text-to-video retrieval		
	R@1	R@5	R@10
Replaced with Transformer	45.8	72.0	81.9
LSECA	47.1	74.9	82.8

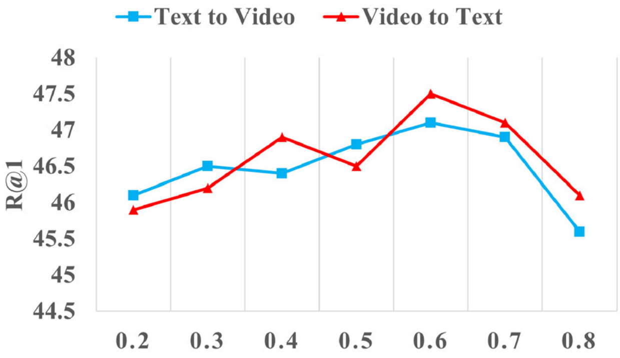

Fig. 6 Effect of the trade-off hyper-parameters α on MSRVTT 1K validation set

4.3.5 Effect of the cross-attention module in textual LSE module

As shown in Table 8, we further compare our method with other interaction methods. For the transformer, we cascade the word and keyword features, input them into the transformer, and output them as enhanced local text features. From the experimental results, we can see that the values of R@1, R@5 and R@10 degrade to some extent. Our approach improves the representation of textual local features and

Query: *a* *girl* *and a* *man* *are* *talking* *to each other.*		Similarity		
Videos	Keywords	Ours	UATVR	UCOFIA
	girl	31.65	29.29	30.20
	man	29.11	27.54	29.85
	talking	30.89	29.65	30.02

Fig. 7 Our top-3 text-to-video retrieval visualization results on MSR-VTT. And we also visualize the other state-of-the-art methods(UATVR [46] and UCOFIA [42])

obtains good retrieval performance by using keyword features as queries and reassigning semantics in word features through a cross-attention mechanism.

4.4 Parameter sensitivity analysis

The hyper-parameter α is used to trade off S_{coarse} and S_{fine} in Eq. 10. Intuitively the matching scores of different granularity features may contribute differently to the final retrieval. So we conduct experiments with the value range setting $\alpha \in [0.2, 0.8]$ as shown in Fig. 6. And we can observe that our proposed LSECA achieves the best retrieval performance when $\alpha = 0.6$ is adopted.

4.5 Qualitative analysis

To visually validate the effectiveness of our proposed LSECA, we show a typical text-to-video retrieval example in Fig. 7 and make the comparsion with the UATVR [46] and UCOFIA [42]. Our model can find the correct video based on keyword guidance from similar videos. The similarity between the third video and query calculated by UATVR [46] is highest, leading to incorrect retrieval results. Although UCOFIA [42] retrieves the correct video, it did not distinguish well between hard negative pairs. Local semantic enhancement makes it possible to find key information about videos and text, and cross aggregation aids in the process of information filtering. Thus, LSECA performs well in visual and textual content understanding, achieving good retrieval results.

5 Conclusion

In this paper, we have proposed a new framework LSECA which not only considers the interaction between two modalities but also enhances the fine-grained video and text representations. For the heterogeneity between video and text, we have proposed different local semantic enhancement schemes, which utlizies global embedding of the video and keywords of the text as anchors to guide fine-grained features to highlight semantic information. Moreover, we have also designed the cross interaction module for frame and text features, which can achieve sufficient interaction between two modalities. Experiments have shown that LSECA achieves significant improvements on three standard text-video retrieval datasets, verifying the effectiveness and generalization of our proposed method. In this paper, the design of the semantic enhancement module for text embedding is slightly simplistic, and the keywords can bring much more than that to the retrieval task, we are working towards this direction in our future work.

Acknowledgements This work was supported by NSFC (62202204) and the Fundamental Research Funds for the Central Universities (JUSRP123032).

Author Contributions ZW, DZ and ZH conceived and designed the analysis of the study. ZW performed the experiment. DZ, ZW and ZH wrote the main manuscript text. DZ and ZH validate the study. All authors reviewed the manuscript.

Data Availability Statements The MSRVTT dataset is given in [27]. The MSVD is given in [28]. The LSMDC is given in [29]. The source code supporting the results of this study are available upon request from the authors.

Declarations

Conflict of interest The authors declare that there is no Conflict of interest regarding the content of the study.

References

1. Wang J, Hua Y, Yang Y, Kou H (2023) Spsd: similarity-preserving self-distillation for video-text retrieval. Int J Multimed Inf Retr 12(2):32
2. Mithun NC, Li J, Metze F, Chowdhury AKR (2019) Joint embeddings with multimodal cues for video-text retrieval. Int J Multimed Inf Retr 8:3–18
3. Gabeur V, Sun C, Alahari K, Schmid C (2020) Multi-modal transformer for video retrieval. In Computer vision–ECCV 2020: 16th

european conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, pp. 214–229. Springer
4. Liu Y, Albanie S, Nagrani A, Zisserman A (2019) Use what you have: video retrieval using representations from collaborative experts. arXiv preprint arXiv:1907.13487
5. Lei J, Li L, Zhou L, Gan Z, Berg TL, Bansal M, Liu J (2021) 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, pp. 7331–7341
6. Bain M, Nagrani A, Varol G, Zisserman A (2021) 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, pp. 1728–1738
7. Liu S, Fan H, Qian S, Chen Y, Ding W, Wang Z (2021) Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 11915–11925
8. Ge Y, Ge Y, Liu X, Li D, Shan Y, Qie X, Luo P (2022) Bridging video-text retrieval with multiple choice questions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16167–16176
9. Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C (2021) Vivit: a video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 6836–6846
10. Bertasius G, Wang H, Torresani L (2021) Is space-time attention all you need for video understanding? In: Proceedings of the 38th international conference on machine learning, vol 139. PMLR, pp 813–824
11. Zhang D, Wu X-J, Yu J (2021) Discrete bidirectional matrix factorization hashing for zero-shot cross-media retrieval. In: Chinese conference on pattern recognition and computer vision (PRCV), pp. 524–536. Springer
12. Dong J, Li X, Xu C, Ji S, He Y, Yang G, Wang X (2019) Dual encoding for zero-example video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9346–9355
13. Dzabraev M, Kalashnikov M, Komkov S, Petiushko A (2021) Mdmmt: multidomain multimodal transformer for video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3354–3363
14. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J et al. (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp 8748–8763. PMLR
15. Jia C, Yang Y, Xia Y, Chen Y-T, Parekh Z, Pham H, Le Q, Sung Y-H, Li Z, Duerig T (2021) Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning, pp 4904–4916. PMLR
16. Yu J, Wang Z, Vasudevan V, Yeung L, Seyedhosseini M, Wu Y (2022) Coca: contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917
17. Luo H, Ji L, Zhong M, Chen Y, Lei W, Duan N, Li T (2022) Clip4clip: an empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing 508:293–304
18. Lin C, Ancong W, Liang J, Zhang J, Ge W, Zheng W-S, Shen C (2022) Text-adaptive multiple visual prototype matching for video-text retrieval. Adv Neural Inf Process Syst 35:38655–38666
19. He F, Wang Q, Feng Z, Jiang W, Lü Y, Zhu Y, Tan X (2021) Improving video retrieval by adaptive margin. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, pp 1359–1368
20. Sun C, Myers A, Vondrick C, Murphy K, Schmid C (2019) Videobert: a joint model for video and language representation learning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7464–7473
21. Liu Y, Xiong P, Xu L, Cao S, Jin Q (2022) Ts2-net: Token shift and selection transformer for text-video retrieval. In: European conference on computer vision, pp 319–335. Springer
22. Ma Y, Xu G, Sun X, Yan M, Zhang J, Ji R (2022) X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In: Proceedings of the 30th ACM international conference on multimedia, pp 638–647
23. Yang J, Bisk Y, Gao J (2021) Taco: token-aware cascade contrastive learning for video-text alignment. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 11562–11572
24. Yao L, Huang R, Hou L, Lu G, Niu M, Xu H, Liang X, Li Z, Jiang X, Xu C (2021) Filip: fine-grained interactive language-image pretraining. arXiv preprint arXiv:2111.07783
25. Li L, Chen Y-C, Cheng Y, Gan Z, Yu L, Liu J (2020) Hero: hierarchical encoder for video+ language omni-representation pre-training. arXiv preprint arXiv:2005.00200
26. Tong X-Y, Xia G-S, Fan H, Zhong Y, Datcu M, Zhang L (2019) Exploiting deep features for remote sensing image retrieval: a systematic investigation. IEEE Transactions on Big Data 6(3):507–521
27. Xu J, Mei T, Yao T, Rui Y (2016) Msr-vtt: a large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5288–5296
28. Wu Z, Yao T, Fu Y, Jiang Y-G (2017) Deep learning for video classification and captioning. In: Frontiers of multimedia research. Association for Computing Machinery and Morgan & Claypool, pp 3–29. https://doi.org/10.1145/3122865.3122867
29. Rohrbach A, Rohrbach M, Schiele B (2015) The long-short story of movie description. In: Pattern recognition: 37th german conference, GCPR 2015, Aachen, Germany, October 7-10, 2015, Proceedings 37, pp 209–221. Springer
30. Sharma P, Li Y (2019) Self-supervised contextual keyword and keyphrase retrieval with self-labelling. Preprints. https://doi.org/10.20944/preprints201908.0073.v1
31. Zhang D, Xiao-Jun W (2020) Scalable discrete matrix factorization and semantic autoencoder for cross-media retrieval. IEEE Transactions on Cybernetics 52(7):5947–5960
32. Cheng X, Lin H, Wu X, Yang F, Shen D (2021) Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. arXiv preprint arXiv:2109.04290
33. Wang X, Zhu L, Yang Y (2021) T2vlad: global-local sequence alignment for text-video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5079–5088
34. Zhang D, Wu X-J, Xu T, Yin H-F (2021) DAH: discrete asymmetric hashing for efficient cross-media retrieval. IEEE Trans Knowl Data Eng 35(2):1365–1378
35. Zhao S, Zhu L, Wang X, Yang Y (2022) Centerclip: token clustering for efficient text-video retrieval. In: Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, pp 970–981
36. Fang H, Xiong P, Xu L, Chen Y (2021) Clip2video: mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097
37. Gorti SK, Vouitsis N, Ma J, Golestan K, Volkovs M, Garg A, Yu G (2022) X-pool: Cross-modal language-video attention for text-video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5006–5015
38. Zhang D, Wu X-J, Xu T, Kittler J (2023) Watch: two-stage discrete cross-media hashing. IEEE Trans Knowl Data Eng 35(6):6461–6474

39. Zhang D, Xiao-Jun W, Yin H-F, Kittler J (2021) Moon: multi-hash codes joint learning for cross-media retrieval. Pattern Recogn Lett 151:19–25
40. Zhang D, Wu X-J, Liu Z, Yu J, Kitter J (2021) Fast discrete cross-modal hashing based on label relaxation and matrix factorization. In: 2020 25th International conference on pattern recognition (ICPR), pp 4845–4850. IEEE
41. Wang Q, Zhang Y, Zheng Y, Pan P, Hua X-S (2022) Disentangled representation learning for text-video retrieval. arXiv preprint arXiv:2203.07111
42. Wang Z, Sung Y-L, Cheng F, Bertasius G, Bansal M(2023) Unified coarse-to-fine alignment for video-text retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 2816–2827, October
43. Tian K, Zhao R, Xin Z, Lan B, Li X (2024) Holistic features are almost sufficient for text-to-video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
44. Chen L, Deng Z, Liu L, Yin S (2024) Multilevel semantic interaction alignment for video–text cross-modal retrieval. IEEE Trans Circuits Syst Video Technol. https://doi.org/10.1109/TCSVT.2024.3360530
45. Deng C, Chen Q, Qin P, Chen D, Wu Q (2023) Prompt switch: efficient clip adaptation for text-video retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 15648–15658, October
46. Fang B, Wu W, Liu C, Zhou Y, Song Y, Wang W, Shu X, Ji X, Wang J(2023) Uatvr: uncertainty-adaptive text-video retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 13723–13733, October
47. Amrani E, Ben-Ari R, Rotman D, Bronstein A (2021) Noise estimation using density estimation for self-supervised multimodal learning. In: Proceedings of the AAAI conference on artificial intelligence 35:6644–6652
48. Mithun NC, Li J, Metze F, Roy-Chowdhury AK (2018) Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In: Proceedings of the 2018 ACM on international conference on multimedia retrieval, pp 19–27
49. Patrick M, Huang P-Y, Asano Y, Metze F, Hauptmann A, Henriques J, Vedaldi A (2020) Support-set bottlenecks for video-text representation learning. arXiv preprint arXiv:2010.02824
50. Croitoru I, Bogolin S-V, Leordeanu M, Jin H, Zisserman A, Albanie S, Liu (2021) Teachtext: Crossmodal generalized distillation for text-video retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 11583–11593
51. Yu Y, Ko H, Choi J, Kim G (2017) End-to-end concept word detection for video captioning, retrieval, and question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3165–3173
52. Song X, Chen J, Zuxuan W, Jiang Y-G (2021) Spatial-temporal graphs for cross-modal text2video retrieval. IEEE Trans Multimedia 24:2914–2923
53. Yu Y, Kim J, Kim G (2018) A joint sequence fusion model for video question answering and retrieval. In: Proceedings of the European conference on computer vision (ECCV), pp 471–487
54. Bogolin S-V, Croitoru I, Jin H, Liu Y, Albanie S (2022) Cross modal retrieval with querybank normalisation. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pp 5194–5205

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.