Neural Processing Letters (2024) 56:54
https://doi.org/10.1007/s11063-024-11468-5

MGSGA: Multi-grained and Semantic-Guided Alignment for Text-Video Retrieval

Xiaoyu Wu[1] · Jiayao Qian[1] · Lulu Yang[1]

Accepted: 27 November 2023

Abstract
In the text-video retrieval task, the objective is to calculate the similarity between a text and a video, and rank the relevant candidates higher. Most existing methods only consider the text-video semantic alignment in the global view. But using mean-pooling to obtain global semantics and simply aligning text and video in the global view may lead to semantic bias. In addition, some methods utilize offline object detectors or sentence parsers to obtain entity-level information in text and video and achieve local alignment. However, inaccurate detection introduces possible errors and such approaches prevent models from being trained end-to-end for retrieval. To overcome these limitations, we propose multi-grained and semantic-guided alignment for text-video retrieval in this paper, which can achieve fine-grained alignment based on video frames and text words, local alignment based on semantic centers, and global alignment. Specially, we explore summary semantics of text and video to guide the local alignment based on semantic centers for we believe that the importance of each semantic center is determined by summary semantics. We evaluate our approach on four benchmark datasets of MSRVTT, MSVD, ActivityNet Captions, and DiDeMo, achieving better performance than most existing methods.

Keywords Text-video retrieval · Contrastive learning · Semantic alignment · Cross-modal interaction · Common semantic space

Xiaoyu Wu, Jiayao Qian and Lulu Yang have contributed equally to this work.

✉ Xiaoyu Wu
wuxiaoyu@cuc.edu.cn

Jiayao Qian
qjy759@cuc.edu.cn

Lulu Yang
yangll05128@163.com

[1] State Key Laboratory of Media Convergence and Communication, Communication University of China, No.1 Dingfuzhuang East Street, Beijing 100024, Beijing, China

Published online: 17 February 2024

1 Introduction

With the rapid development of the information era, everyone can act as the publisher and disseminator of media content on the Internet, which makes the scale of information on the Internet grow explosively. Video and text are two essential forms of information with different modalities. Text-video retrieval is an important way for us to use the Internet. Users usually input keywords or descriptions to search for related videos, as shown in Fig. 1. For the text-video retrieval task, there is a natural semantic gap between the two modalities due to their heterogeneity. Although the human brain can process the two modalities with the help of the acquired cognitive and knowledge system to realize high-level semantic association and understanding, the machine still needs to rely on artificial intelligence to achieve this process. At present, the mainstream method is based on common semantic space, that is, the text embedding and video embedding are obtained by using encoders respectively, and then projected into a common semantic space, as shown in Fig. 2. For text-to-video retrieval, the similarities between a text query and candidate videos are calculated here. Then the candidate videos are ranked from high similarity to low. For video-to-text retrieval, the candidate texts are ranked according to the similarities with the query video from high to low.

Some earlier studies [1–6] consider the multi-stream information of video, regarding video as a synthesis of multi-modal, such as appearance, audio, and face. They take task-specific architecture (ResNet101 [7], Vggish [8], S3D [9],...) as offline encoders to get appearance embeddings, audio embeddings, motion embeddings, and so on. Then, the global video representation is obtained by early fusion at the embedding level. Text-video alignment is realized by calculating the similarity between global text and video representations. Another

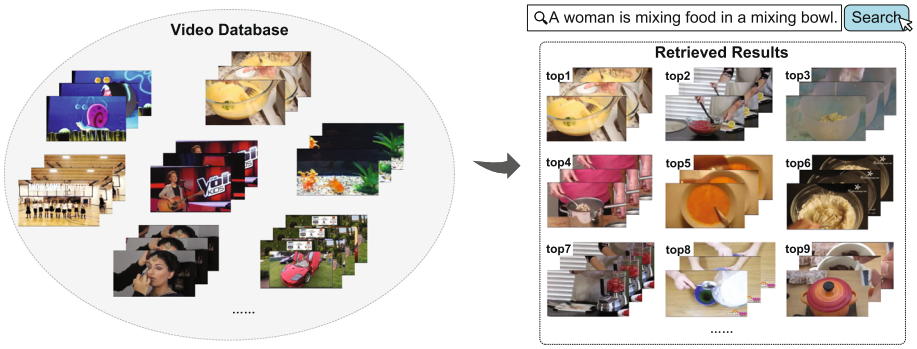

Fig. 1 Illustration of text-to-video retrieval. The user enters text, and the search engine returns several related videos from the video database

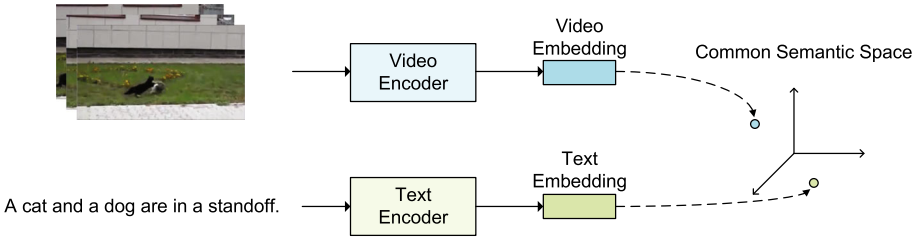

Fig. 2 Text-video retrieval method based on common semantic space. The embeddings of text and video are projected into the common semantic space to calculate the similarity

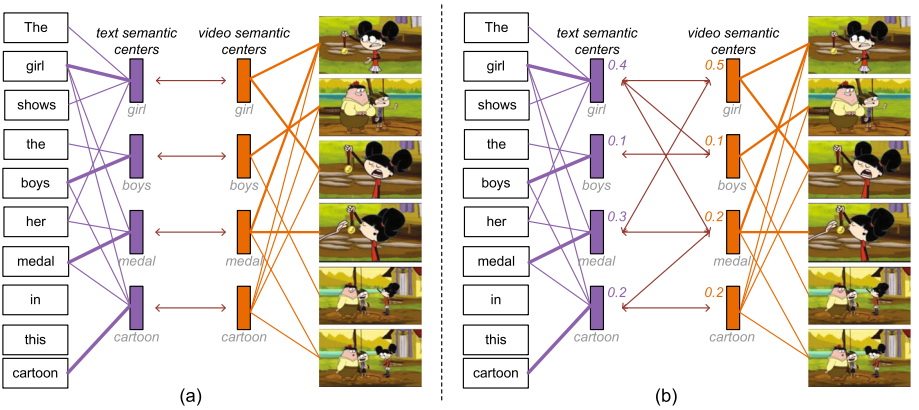

Fig. 3 Comparison between the local-level alignment of (**a**) T2VLAD [6] and (**b**) ours. **a** is the local alignment with one-to-one correspondence between semantic centers. And every semantic center is treated equally. In our view (**b**), there may be cross-modal semantic associations between centers of text and video. For example, the video semantic center "girl" should be semantically related to the text semantic center "medal". Because in the video frames, the girl always holds the medal. Besides, the semantic center "girl" should be the most important semantic center. Because "girl" is the subject of the sentence and the executor of the action. It is also the most frequent entity in all video frames. It should be given a higher importance score, as shown by the number next to the semantic center

way is to weight the similarities between text and various video modalities. The performance of these approaches is limited by the task-specific feature encoders and cannot be fine-tuned end-to-end to suit cross-modal retrieval task. Recently, with the success of Contrastive Language-Image Pretraining (CLIP) [10], some text-video retrieval methods [11–14] exploit CLIP as visual and text encoders, benefiting from large-scale pre-training with natural language supervision. Their common point is that the global video semantic is obtained by average-pooling the video frame representations, and aligns with the global text semantic. Actually, the text description could be relevant to a local video clip, which means that we should not simply assume that the average-pooled frame representations and the global text semantics are completely matched. We should construct more fine-grained text-video alignment.

Some methods [15–17] adopt an object detection network on video frames to obtain entity embeddings and match the nouns extracted from the text. The inaccurate detection and word selection may cause possible errors in the retrieval task. Moreover, these approaches do not unify multi-grain semantic alignment, which leads to weak cross-modal associations. T2VLAD [6] assigns the video and text features to a set of shared centers, generating the center features for both video and text to calculate a local text-video similarity. It holds that the locally aligned semantic centers of video and text are in one-to-one correspondence, and all semantic centers are treated equally when calculating similarity, as shown in Fig. 3a. We believe that this alignment has semantic bias. As shown in Fig. 3b, on the one hand, there may be cross-correspondences and semantic associations between centers of text and video. On the other hand, different semantic centers make inconsistent contributions to global semantics, which can be quantified by the important scores generated by summary semantics.

In this paper, we present Multi-Grained and Semantic-Guided Alignment (MGSGA) for text-video retrieval. To get robust frame and word representations, MGSGA takes pretrained CLIP as visual and text encoders. In the multi-grained text-video alignment, firstly, we adopt a weighted token interaction module to align frame-word and calculate fine-grained similarity.

Secondly, we employ an aggregation network to get the embeddings of text semantic centers and video semantic centers with frame representations and word representations as input, respectively. We believe that summary semantics can determine the importance of each semantic center. To this end, we utilize the weights generated by video summary embedding and text summary embedding to guide the cross-modal alignment of local-level semantic centers. Finally, we use the aggregation network again to obtain the aggregated global text and video representation, achieving global-level text-video alignment. To sum up, MGSGA considers fine-grained, local, and global semantic alignment for text and video and applies summary semantics to guide the alignment of local semantic centers. Consequently, MGSGA achieves excellent performance on several standard benchmarks.

In this work, we make the following three contributions:

(i) We propose a multi-grained text-video alignment framework for text-video retrieval, which can achieve fine-grained alignment based on video frames and text words, local alignment based on semantic centers, and global alignment based on aggregation.
(ii) To make the local alignment better correlate with the summary semantic information, we present a semantic guidance mechanism. The summary semantic assigns weight to each semantic center. The semantic centers with higher weights are more dominant in the similarity calculation of the local alignment.
(iii) Experimental results show that our method outperforms most of the existing methods on the standard text-video retrieval benchmarks, including MSRVTT, MSVD, ActivityNet Captions and DiDeMo. We also conduct some ablation experiments to verify the effect of each component.

2 Related Works

2.1 CLIP-based Approaches for Text-Video Retrieval

A pioneering work in visual-language pretraining is Contrastive Language-Image Pretraining (CLIP) [10], which pretrained with 400 million text-image pairs to learn natural language supervision knowledge. CLIP has achieved remarkable performance in multiple downstream tasks. Based on CLIP, many researchers transfer the cross-modal information of text-image to the text-video retrieval task. Luo et al. [11] take the initiative to apply CLIP to text-video retrieval and explore three similarity calculators to realize time aggregation among video frames. CLIP2Video [12] adds a temporal difference block and a temporal alignment block on the basis of CLIP to model the temporal relationship and the alignment of text-video pairs, respectively. MKTVR [18] uses CLIP as feature extractors, utilizes machine translation models to construct multilingual text-video pairs, then transfers the knowledge from multilingual models to improve video retrieval performance. CLIP2TV [13] introduces momentum distillation and adopts multi-modal fusion to enhance cross-modal interaction. X-pool [19] uses cross-modal attention to make the text representations focus more on the frames that are more relevant. However, cross-modal attention brings additional computational costs in the test stage. Besides, some recent works [20–23] improve token encoding in Transformer [24] to reduce redundancy or model subtle motions. In summary, text-video retrieval methods based on CLIP constantly refresh the leaderboard of this task, which verify the capability of cross-modal knowledge learned from large-scale dataset composed of text-image pairs.

Our approach also benefits from the pretrained CLIP in feature extraction. However, our method proposes multi-grained text-video semantic alignment and innovatively designs a

Table 1 Classification of text-video retrieval methods based on multi-grained alignment.

Methods	Multi-experts	Global	Local		E2E
			PP	LE	
MEE [1]	√	√			
CE [2]	√	√			
MMT [3]	√	√			
MDMMT [4]	√	√			
MDMMT-2 [5]	√	√			
T2VLAD [6]	√	√		√	
Dual-encoding [30]		√		√	
HANet [15]		√	√		
HGR [17]		√	√		
ViSERN [31]		√	√		
BridgeFormer [32]		√	√		
CAMoE [14]		√	√	√	
HiSE [33]		√	√		
Ours		√		√	√

Multi-experts denotes using multiple architectures to extract video features. PP (pre-processing) denotes semantic parsing of text or extracting regions from videos. LE (learnable embeddings) refers to learning local information through training. E2E refers to using the raw text and video as model inputs, and the entire network can be fine-tuned end-to-end

semantic guidance mechanism to correct deviations in alignment. Experimental results show that our approach surpasses the performance of some CLIP-based methods.

2.2 Multi-grained Approaches for Text-Video Retrieval

In the text-image retrieval task, UNITER [25] and Oscar [26] extract region features of the image, and construct fine-grained alignment between word and region during the pre-training stage. TERAN [27] uses Faster-RCNN [28] to extract region features, and calculates the region-word fine-grained similarity matrix, obtaining the global-level similarity score through max-over-regions sum-over-words operation. HGAN [29] uses Resnet152 [7] and Faster-RCNN [28] to extract global and local features, respectively, and establish a feature graph to achieve global and local alignment. However, unlike images, videos represent dynamic events through continuous frames. Extracting region features frame by frame from videos is redundant and may not be effective. These methods cannot be extended to text-video retrieval task. Moreover, these methods either do not establish multi-grained alignment or do not establish semantic guidance between granularity.

For the text-video retrieval methods based on multi-grained alignment, we make classification among them based on whether to use multi-experts information of video, whether pre-processing is performed, and whether end-to-end fine-tuning is avaliable, as shown in Table 1.

Video is usually a synthesis of motion, appearance, audio, and other modalities. These modalities can be regarded as multi-grained information of video. Traditional text-video retrieval methods use mature architectures in action recognition, video classification, audio classification, and other tasks to extract video features, and define them as ‘experts’. They

put emphasis on how to integrate the information of these experts to get a strong global video representation, or calculate similarities with text representation separately and then aggregate in some way to achieve the retrieval model. MEE [1] uses the appearance, motion, face, and audio features extracted from video to calculate similarities with text features respectively, and weights the similarities to get a global similarity. CE [2] adds a collaborative gating mechanism based on MEE to enhance the interaction between various experts. MMT [3] applies Transformer to the video experts and uses multi-layer attention to strengthen temporal interaction within an expert and cross-modal interaction between experts, so as to obtain a powerful and compact video representation. Based on MMT, MDMMT [4] combines several video description datasets in an attempt to obtain a text-video retrieval model with multi-domain generalization ability. MDMMT-2 [5] adds a double positional encoding on the foundation of MDMMT to better integrate the various video experts. Besides, T2VLAD [6] applies two aggregation methods at different granularity for experts to align text-video pairs globally and locally. However, the performance of such methods is limited by the feature extractors, which are trained for specific downstream tasks. The feature extraction part cannot be finetuned in the retrieval task to learn the cross-modal information between text and video.

Some works focus on exploring the fine-grained semantics of both video and text to achieve retrieval models that consider coarse and fine granularity. Dong et al. [30] uses CNN and biGRU as encoder to obtain global encoding, temporal-aware encoding, and local-enhanced encoding, and projects the multi-level encoding of text and video into a common space for similarity calculation. Texts often contain elements such as nouns and verbs, which are supposed to be semantically matched with entities and actions in video. Based on this view, some works [15–17, 33] parse text into events, actions, and entities, construct semantic role graph and conduct attention-based graph reasoning, so as to perform hierarchical matching with video. ViSERN [31] adopts offline ResNet101 [7] to extract regional features from video frames, constructs semantic relationships, and uses GCN for graph reasoning. BridgeFormer [32] introduces Multiple Choice Questions (MCQ) task in the training stage, which enhances the video representation and enables it to answer the noun and verb questions in MCQ task. CAMoE [14] splits nouns and verbs from texts first. Fusion expert, entity expert, and action expert are extracted from video through attention networks, which are matched with the global text embedding, noun embedding, and verb embedding, respectively. The loss functions of the three matches are averaged to realize the retrieval model.

However, these multi-grained text-video retrieval methods base on semantic relationships need to adopt sentence parser technology to extract semantic roles. The possible errors of the parser may lead to inaccurate semantic relationships and then affect the semantic matching. The methods of using region features as local information require pre-extraction. Both of them hinder the ability of the model to be fine-tuned end-to-end. As shown in Table 1, our approach avoids introducing additional steps, but instead learns more accurate semantic relationships through fine-tuning the entire network end-to-end and constructs multi-grained and semantic-guided alignment through hierarchical design and interaction.

3 Methods

3.1 Overview

As shown in Fig. 4, we propose multi-grained and semantic-guided alignment for text-video retrieval, which aligns text and video features in fine-grained, local and global granularity. In

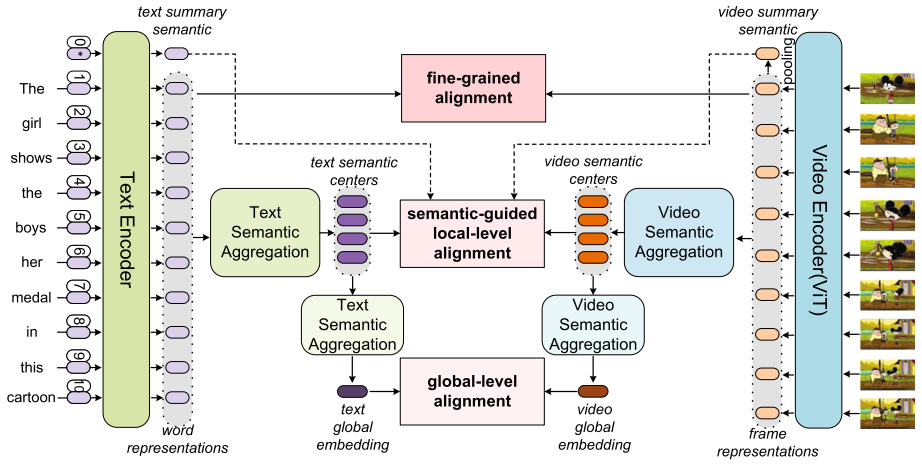

Fig. 4 Overview of MGSGA. The proposed method for cross-modal retrieval can align text and video embeddings at fine granularity, local level, and global level. Specially, in the semantic-guided local-level alignment, we exploit summary semantics to assign important scores to each semantic center. For clarity, we show the case when the number of semantic centers for each modality $K = 4$

the text-video retrieval, given a video set $\mathcal{V}$ and a text set $\mathcal{T}$, the goal of semantic alignment is to measure the similarity $s(v_i, t_j)$ between video $v_i \in \mathcal{V}$ and text $t_j \in \mathcal{T}$. To begin with, the CLIP model pretrained with 400 million text-image pairs is applied to extract text and video features (Sect. 3.2). Consequently, multi-grained and semantic-guided text-video alignment is adopted to align the two modalities from different granularities(Sect. 3.3). We consider that summary semantics can assist local-level alignment. So we obtain summary semantic embeddings from existing representations without extra computation to guide the alignment of local semantics. Finally, we use contrastive learning to construct three loss functions corresponding to the three stages of the multi-grained text-video alignment to form the total loss function for our method (Sect. 3.4).

3.2 Feature Representation

We sample the video v_i uniformly to get N frames, denoted as $v_i = \{v_i^1, v_i^2, ..., v_i^n, ..., v_i^N\}$, where $n \in [1, N]$. The visual encoder F of CLIP (ViT-B/16) takes frames as input. The output from the last layer at [CLS] token is regarded as frame representations. The frame representations of the video v_i is denoted as $z_i = F(v_i) = \{z_i^1, z_i^2, ..., z_i^n, ..., z_i^N\}$, with $z_i \in \mathbb{R}^{N \times d}$.

For the text t_j, we denote it as $t_j = \{t_j^1, t_j^2, ..., t_j^l, ..., t_j^L\}$, where $j \in [1, L]$ and L is the number of words. The text encoder G of CLIP (ViT-B/16) is applied to get the word representations, denoted as $w_j = G(t_j) = \{w_j^1, w_j^2, ..., w_j^l, ..., w_j^L\}$, with $w_j \in \mathbb{R}^{L \times d}$. In addition, the output w_j^0 from G at [EOS] token is treated as the summary embedding of text t_j.

3.3 Multi-grained and Semantic-Guided Text-Video Alignment

Firstly, z_i and w_j are frame representations and word representations respectively, which can be regarded as fine-grained semantics. Text-video alignment at fine granularity makes full use of the frame and word information. Secondly, we obtain multiple local semantic centers through an aggregation network to realize a higher level of alignment. In the local-level text-video alignment, we use the weights generated from mean-pooling of frame representations to guide the similarity calculation of video-to-text. We use the weights generated from [EOS] token of text to guide the similarity calculation of text-to-video. Finally, the semantic aggregation network is applied again to obtain the global representations of text and video to achieve global-level alignment.

3.3.1 Fine-Grained Text-Video Alignment

Given the sequences of frame representations $z_i = \{z_i^1, z_i^2, ..., z_i^n, ..., z_i^N\}$ and word representations $w_j = \{w_j^1, w_j^2, ..., w_j^l, ..., w_j^L\}$, we adopt a weighted token interaction module for fine-grained alignment.

Specifically, the frame-word similarity matrix $S_{f-w} \in \mathbb{R}^{N \times L}$ is calculated via inner product. The nth row in the matrix represents the similarities between frame representation z_i^n and all word representations $\{w_j^l\}_{l=1}^L$. We take the maximum value as the similarity between the current frame v_i^n and the text t_j as a single frame-text similarity. A weight estimation module composed of MLP and Softmax is applied to get the importance scores of frames, which takes the frame representations z_i as input and outputs N scores. The importance scores are utilized to weight the frame-text similarities to get the fine-grained video-to-text similarity score:

$$s_{fgrained}^{v2t}(v_i, t_j) = \sum_{n=1}^{N} f_{vw,\theta}^{n}(z_i) max_{l=1}^{L} \left\{ \left(\frac{z_i^n}{\|z_i^n\|_2}\right)^T \left(\frac{w_j^l}{\|w_j^l\|_2}\right) \right\}, \tag{1}$$

where $f_{vw,\theta}(\cdot)$ is the weight estimation module with frame representations z_i as input and T denotes transposition.

Similarly, fine-grained text-to-video similarity score can be calculated:

$$s_{fgrained}^{t2v}(v_i, t_j) = \sum_{l=1}^{L} f_{tw,\theta}^{l}(w_j) max_{n=1}^{N} \left\{ \left(\frac{z_i^n}{\|z_i^n\|_2}\right)^T \left(\frac{w_j^l}{\|w_j^l\|_2}\right) \right\}. \tag{2}$$

3.3.2 Semantic-Guided Local-level Text-Video Alignment

In fine-grained text-video alignment, we establish the semantic matching relationship between frames and words. However, sentences usually contain some modal words, prepositions, etc., which serve no practical purpose in the semantic matching of events. Although some works [14, 17, 32] employ a sentence parser to label the parts of speech, it increases the computational cost from additional steps and introduces possible errors. In addition, although the video is uniformly sampled with a certain sampling interval, it may still contain some of the same frames due to the temporal redundancy of video. Therefore, in the local-level text-video alignment, we introduce the idea of clustering and use a semantic aggregation network to adaptively generate local representations of text and video.

Specifically, assuming that there are K semantic centers in a video and a text, denoted as $\{c_p^{v_i}\}_{p=1}^K$ and $\{c_q^{t_j}\}_{q=1}^K$, respectively. Given the frame representation z_i^n, its contribution to the pth video semantic center $c_p^{v_i}$ is represented as $a_{n,p}^i$:

$$a_{n,p}^i = \frac{exp(z_i^n (c_p^{v_i})^T + b_p^{v_i})}{\sum_{k=1}^K exp(z_i^n (c_k^{v_i})^T + b_k^{v_i})}, \tag{3}$$

where $b_p^{v_i}$ is a learnable bias term. After that, we use the following formula to calculate the embeddings of video semantic centers:

$$c_p^{v_i} = normalize\left(\sum_{n=1}^N a_{n,p}^i (z_i^n - c'^{v_i}_p)\right), \tag{4}$$

where $c'^{v_i}_p$ is learnable and has the same size as $c_p^{v_i}$. They belong to the same semantic center, which was proposed in [34] to enhance the adaptive ability of aggregation.

Correspondingly, the embeddings of text semantic centers are obtained by the following formula:

$$c_q^{t_j} = normalize(\sum_{l=1}^L \frac{exp(w_j^l (c_q^{t_j})^T + b_q^{t_j})}{\sum_{k=1}^K exp(w_j^l (c_k^{t_j})^T + b_k^{t_j})} (w_j^l - c'^{t_j}_q)), \tag{5}$$

where $b_q^{t_j}$ is a learnable bias term and $c'^{t_j}_q$ is learnable and has the same size as $c_q^{t_j}$.

Given the text and video semantic center representations $\{c_p^{v_i}\}_{p=1}^K$ and $\{c_q^{t_j}\}_{q=1}^K$, respectively, different from the one-to-one correspondence between video semantic centers and text semantic centers in [6], we believe that there may be cross-correspondences between semantic centers of two modalities. Besides, there should be summary semantics that direct the retrieval model to focus more on the more important semantic centers. Motivated by these, we propose a semantic-guided text-video alignment module at the local level, as shown in Fig. 5.

For frame representations $\{z_i^n\}_{n=1}^N$, mean-pooling is applied to get the video summary embedding z_i^0. Then we employ MLP and Softmax to z_i^0 to calculate the weight of each video semantic center, which is used to guide the calculation of video-to-text similarity. Specifically, we obtain a cross-modal center-level similarity matrix $S_c \in \mathbb{R}^{K \times K}$ via inner product firstly. The pth row in the matrix denotes the similarities between the video semantic center $c_p^{v_i}$ and all text semantic centers $\{c_q^{t_j}\}_{q=1}^K$. Then the maximum value is selected since it measures the similarity between the current video semantic center $c_p^{v_i}$ and the most relevant text semantic center to $c_p^{v_i}$. Lastly, the weights generated from video summary embedding are used to weight the similarities we selected in each row to get the local-level video-to-text similarity score as follows:

$$z_i^0 = mean - pooling(z_i^1, z_i^2, ..., z_i^N), \tag{6}$$

$$s_{local}^{v2t}(v_i, t_j) = \sum_{p=1}^K f_{vw',\theta'}^p(z_i^0) max_{q=1}^K \{(\frac{c_p^{v_i}}{\|c_p^{v_i}\|_2})^T (\frac{c_q^{t_j}}{\|c_q^{t_j}\|_2})\}, \tag{7}$$

where $f_{vw',\theta'}(\cdot)$ is the weight estimation module with video summary embedding z_i^0 as input.

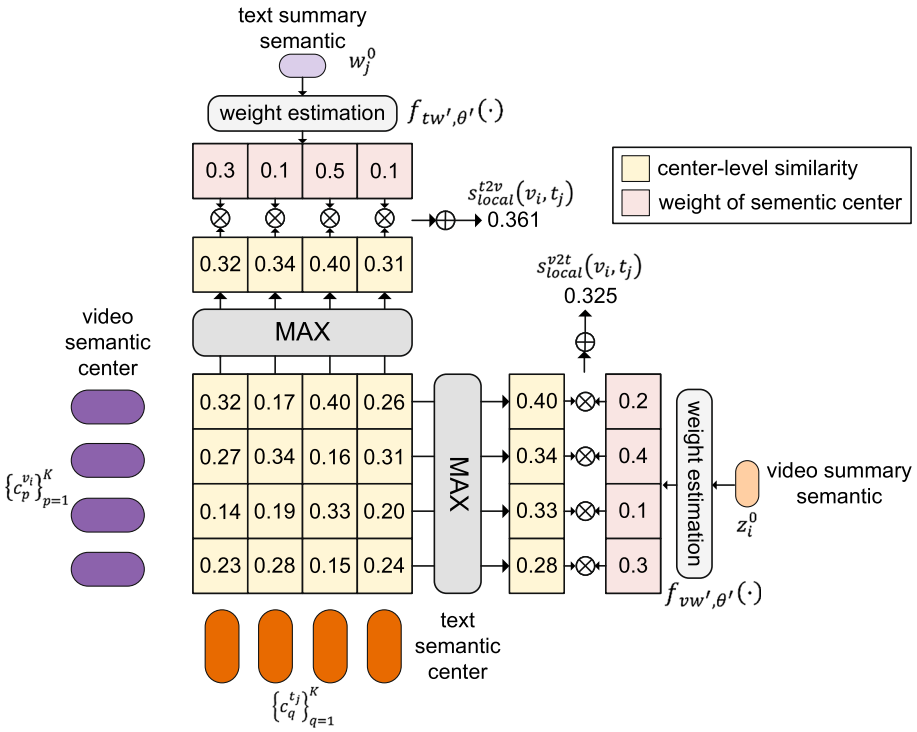

Fig. 5 Illustration of semantic-guided local-level text-video alignment based on semantic centers

The representation w_j^0 of text [EOS] token is used to generate the weights of text semantic centers. Similarly, we calculate the local-level text-to-video similarity score:

$$s_{local}^{t2v}(v_i,t_j)=\sum_{q=1}^{K} f_{tw',\theta'}^{q}(w_j^0)\, max_{p=1}^{K}\{(\frac{c_p^{v_i}}{\|c_p^{v_i}\|_2})^T(\frac{c_q^{t_j}}{\|c_q^{t_j}\|_2})\}, \tag{8}$$

where $f_{tw',\theta'}(\cdot)$ is the weight estimation module with text summary embedding w_j^0 as input.

3.3.3 Global-Level Text-Video Alignment

Given the sequence of video semantic center representations $\{c_p^{v_i}\}_{p=1}^K$, we employ the semantic aggregation network used in Sect. 3.3.2 to get the aggregated video embedding, just modifying the K to 1 in the formula 5. The obtained aggregated video embedding $g^{v_i} \in \mathbb{R}^{1\times d}$ comprehensively considers multiple semantic centers.

Similarly, a text semantic aggregation network is utilized to get the aggregated text embedding $g^{t_j} \in \mathbb{R}^{1\times d}$. Thus, the global-level text-video alignment similarity score is calculated by inner product:

$$s_{global}(v_i,t_j)=\frac{(g^{v_i})^T\cdot g^{t_j}}{\|g^{v_i}\|\|g^{t_j}\|}. \tag{9}$$

3.4 Text-Video Contrastive Learning

During training, for a given batch consisting of B text-video pairs, we adopt the InfoNCE loss [35] to optimize the retrieval model. InfoNCE loss maximizes the similarity between the matched text-video pairs and minimizes the similarity between other pairs. For text-video fine-grained alignment based on frames and words, the loss function is calculated as follows:

$$\mathcal{L}_{fgrained} = \mathcal{L}^{v2t}_{fgrained} + \mathcal{L}^{t2v}_{fgrained}, \tag{10}$$

$$\mathcal{L}^{v2t}_{fgrained} = -\frac{1}{B}\sum_{i}^{B} log\frac{exp(s^{v2t}_{fgrained}(v_i, t_i))}{\sum_{j=1}^{B} exp(s^{v2t}_{fgrained}(v_i, t_j))}, \tag{11}$$

$$\mathcal{L}^{t2v}_{fgrained} = -\frac{1}{B}\sum_{i}^{B} log\frac{exp(s^{t2v}_{fgrained}(v_i, t_i))}{\sum_{j=1}^{B} exp(s^{t2v}_{fgrained}(v_j, t_i))}. \tag{12}$$

For text-video local alignment based on semantic centers, the loss function is as follows:

$$\mathcal{L}_{local} = \mathcal{L}^{v2t}_{local} + \mathcal{L}^{t2v}_{local}, \tag{13}$$

$$\mathcal{L}^{v2t}_{local} = -\frac{1}{B}\sum_{i}^{B} log\frac{exp(s^{v2t}_{local}(v_i, t_i))}{\sum_{j=1}^{B} exp(s^{v2t}_{local}(v_i, t_j))}, \tag{14}$$

$$\mathcal{L}^{t2v}_{local} = -\frac{1}{B}\sum_{i}^{B} log\frac{exp(s^{t2v}_{local}(v_i, t_i))}{\sum_{j=1}^{B} exp(s^{t2v}_{local}(v_j, t_i))}. \tag{15}$$

For text-video global alignment, the loss function is as follows:

$$\mathcal{L}_{global} = \mathcal{L}^{v2t}_{global} + \mathcal{L}^{t2v}_{global}, \tag{16}$$

$$\mathcal{L}^{v2t}_{global} = -\frac{1}{B}\sum_{i}^{B} log\frac{exp(s_{global}(v_i, t_i))}{\sum_{j=1}^{B} exp(s_{global}(v_i, t_j))}, \tag{17}$$

$$\mathcal{L}^{t2v}_{global} = -\frac{1}{B}\sum_{i}^{B} log\frac{exp(s_{global}(v_i, t_i))}{\sum_{j=1}^{B} exp(s_{global}(v_j, t_i))}. \tag{18}$$

Finally, the loss function for the proposed multi-grained and semantic-guided text-video alignment is as follows:

$$\mathcal{L} = \mathcal{L}_{fgrained} + \alpha\mathcal{L}_{local} + \beta\mathcal{L}_{global}, \tag{19}$$

Where α and β are the weighting parameters to balance losses at different granularity. It can be set manually. As shown in Sect. 4.4, We tested multiple groups of values in the experiment, and determined the values shown in Sect. 4.2 according to the experimental effect.

4 Experiments

4.1 Datasets and Evaluation Metrics

4.1.1 Datasets

MSRVTT [36] is a text-video cross-modal dataset established by Microsoft. It consists of 10,000 videos and 20 sentences for each video. In the officially provided split (MSRVTT full), the dataset is divided into three parts: training, validation, and testing. We use MSRVTT-9k and MSRVTT-7k splits, which are more commonly used in text-video retrieval. The training set contains 9000 and 7000 videos respectively. They have the same test set, which contains 1000 videos.

MSVD [37] dataset contains 1970 videos, ranging in length from 1 to 62 s. We use the split that was officially provided. The number of videos in training, validation, and test set is 1200, 100, and 670, respectively.

ActivityNet Caption [38] contains 20,000 videos with a total duration of 849 h. The average length of each text is 13.48 words. The ActivityNet Captions dataset focuses more on the actions in the video. The complexity and variability of video actions pose new challenges to the temporal correlation and action semantic understanding ability of text-video retrieval models.

DiDeMo [39] dataset contains 10,000 videos, each containing 40 sentences. Following the practice of [40, 41], the sentences for a video are concatenated to form a query.

The above four datasets represent medium, small, and larger datasets in the text-video cross-modal retrieval task, which can effectively test the performance of the model under different data scales.

4.1.2 Evaluation Metric

We use the standard metrics. R@K, recall rate in rank K, is the proportion of matched samples in the top-K retrieved results. We report R@1, R@5, and R@10 following [6, 11, 19]. We also report the median rank (MdR) and mean rank (MnR) of the matched samples. Higher R@K and lower MdR or MnR indicate better performance.

4.2 Implementation Details

We use CLIP (ViT-B/16) to initialize the parameters of text encoder and video encoder. We uniformly sample 12 frames for each video. The max length of text is 32. A two-layer MLP is adopted to get the weights of frames and words in the fine-grained text-video alignment. The number of semantic centers in the local-level text-video alignment K is set to be 3. In the total loss function, α and β are set to 0.2 and 0.1, respectively. The optimizer is Adam [42]. The initial learning rate is 1e−7 for CLIP and 1e−4 for other modules. The batch size is 16 and we run 5 epochs as the same as CLIP4Clip [11].

4.3 Comparisons with Other Methods

In the text-video retrieval paradigm, the similarity scores of video-to-text can be obtained by transposing the text-to-video similarity scores matrix, thus completing the video-to-text retrieval task. In the following experimental results, we have listed the experimental results

of text-to-video retrieval. Please refer to Appendix A for the experimental results of the video-to-text retrieval task.

Table 2 shows the results of MSRVTT-9k. The performance of the multi-grained and semantic-guided alignment retrieval model proposed in this paper is superior to the approaches based on multi-experts, such as CE, MMT, and T2VLAD, indicating the effect of pretrained CLIP in transferring the cross-modal semantic knowledge of text-image to text-video retrieval. CLIP4Clip-seqTranf adopts a temporal transformer to model temporal relationship. CLIP2Video utilizes a temporal alignment block to realize sequential alignment. Our approach captures the multi-grained semantic relationship, which performs 3.2% and 2.1% higher on R@1 than CLIP4Clip-seqTranf and CLIP2Video, respectively. Compared with X-Pool which utilizes cross-modal attention for feature interaction, our approach exceeds it by 0.8% and 1.7% on R@1 and R@5 respectively, which improves the performance while avoiding the extra computational cost brought by the single-stream model in the test stage. Compared to the multi-grained based methods CAMoE and HiSE, our approach achieves significant advantages. As shown in Table 3, on MSRVTT-7k, our method has an increase of 3.3% over CLIP4Clip-meanP on R@1, reducing MnR to 12.4, achieving better performance than most existing methods.

The comparison with existing methods on MSVD is shown in Table 4. MSVD is a small dataset containing only 1970 videos. Compared with CAMoE, which uses three attention networks, our approach significantly outperforms on R@5 and R@10. Our method achieves 0.4%, 1.7%, 2.2% higher than CLIP4Clip-meanP on R@1, R@5, and R@10, respectively. Our method also reduces the MdR metric of MSVD to 9.1, which is smaller than the previous models.

In Table 5, we show the results of our method on the ActivityNet Captions dataset. Compared with TS2-Net and HBI, our method achieved better R@1, achieving 42.5.

As shown in Table 6, on DiDeMo dataset, our method achieves more than 5.3% and 2.6% improvement over other methods on R@5 and R@10 of text-to-video retrieval.

From the performance comparison on the above four datasets, our approach significantly outperforms the other methods, which demonstrates the importance of multi-grained text-video semantic alignment. Multi-grained semantics is a natural attribute of text and video. It is how the human brain processes information from the two modalities. Therefore, it's essential to establish multi-grained text-video alignment in cross-modal retrieval.

4.4 Ablations and Analysis

To further verify the effect of each module of the multi-grained and semantic-guided text-video alignment, we conduct ablation experiments on MSRVTT-9k.

4.4.1 Effect of Each Granularity

To fully examine the impact of different alignment modules, we conduct an ablation study as shown in Table 7. We start with the fine-grained text-video alignment based on frames and words and add other modules successively. 'local' refers to the pure local-level alignment based on semantic centers, without the guidance mechanism of summary semantics. We can see that adding the local-level alignment directly to the fine-grained alignment does not improve retrieval performance significantly. After applying the guidance mechanism of summary semantics to adjust the local-level alignment based on semantic centers, the R@1 of text-to-video retrieval is increased to 47.3%. After adding the global-level text-

Table 2 Text-to-video retrieval performance comparisons to SOTAs on the MSRVTT-9k dataset

*Method	text-to-video				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Non-CLIP for feature representations					
CE [2]	20.9	48.8	62.4	6.0	28.2
MMT [3]	26.6	57.1	69.6	4.0	24.0
T2VLAD [6]	29.5	59.0	70.1	4.0	–
LaT [44]	35.3	61.3	72.9	3.0	–
BridgeFormer [32]	37.6	64.8	75.1	3.0	–
MILES [45]	37.7	63.6	73.8	3.0	–
MDMMT [4]	38.9	69.0	79.7	2.0	16.5
RaP [46]	40.9	67.2	76.9	2.0	–
CLIP for feature representations					
CLIP [10]	31.2	53.7	64.2	4.0	–
COTS [47]	36.8	63.8	73.2	2.0	–
CLIP4Clip-tightTransf [11]	40.2	71.5	80.5	2.0	13.4
CLIP4Clip-seqLSTM [11]	42.5	70.8	80.7	2.0	16.7
CLIP4Clip-meanP [11]	43.1	70.4	80.8	2.0	16.2
CLIP4Clip-seqTransf [11]	44.5	71.4	81.6	2.0	15.3
CAMoE [14]	44.6	72.6	81.8	2.0	13.3
HiSE [33]	45.0	72.7	81.3	2.0	–
CLIP2Video [12]	45.6	72.6	81.7	2.0	14.6
LAFF [48]	45.8	71.5	82.0	–	–
TokenFlow [49]	46.1	72.7	82.0	2.0	13.6
EMCL-Net [50]	46.8	73.1	83.1	2.0	–
X-Pool [19]	46.9	72.8	82.2	2.0	14.3
TABLE [51]	47.1	74.3	82.9	2.0	13.4
Ours	**47.7**	**74.5**	**83.7**	**2.0**	**11.9**

Here we report results without any post-processing operations (e.g., [14] or [43]) during inference (as in subsequent tables)

video alignment as our entire model, the R@1 of text-to-video retrieval increase to 47.7%. It demonstrates the effect of each granularity of our proposed multi-grained alignment for text-video retrieval. Also, the semantic guidance mechanism plays an important role in the local-level alignment.

4.4.2 Effect of the Number of Centers

For the local alignment based on semantic centers, we also designed a group of experiments by setting different semantic center numbers ($K = 2, 3, 4, 5, 6$), as shown in Fig. 6a. Comprehensively considering the R@1 of text-to-video and video-to-text retrieval, we believe that the effect of $K = 3$ and 4 is similar, which is better than the effect of $K = 5$ or 6. Considering the calculation efficiency, we choose to set the value of K to 3. In fact, the results obtained from the experiment are consistent with the basic composition of the event and human perception. Events are reflected in text as subjects, predicates, and their relation-

Table 3 Text-to-video retrieval performance comparisons to SOTAs on the MSRVTT-7k dataset

Method	text-to-video				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Non-CLIP for feature representations					
HGR [17]	9.2	26.2	36.5	24.0	164.0
ViSERN [31]	18.1	48.4	61.3	6.0	28.6
COTS [47]	32.1	60.8	70.2	3.0	–
RaP [46]	38.5	64.0	74.4	3.0	–
CLIP for feature representations					
CLIP4Clip-tightTransf [11]	37.8	68.4	78.4	2.0	17.2
CLIP4Clip-seqLSTM [11]	41.7	68.8	78.7	2.0	16.6
CLIP4Clip-seqTransf [11]	42.0	68.6	78.7	2.0	16.2
CLIP4Clip-meanP [11]	42.1	71.9	81.4	2.0	15.7
X-Pool [19]	43.9	72.5	**82.3**	2.0	14.6
Ours	**45.4**	**73.9**	81.7	**2.0**	**12.4**

Table 4 Text-to-video retrieval performance comparisons to SOTAs on the MSVD dataset

Method	text-to-video				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Non-CLIP for feature representations					
ViSERN [31]	18.1	48.4	61.3	6.0	28.6
CE [2]	19.8	49.0	63.8	6.0	23.1
LaT [44]	40.0	74.6	84.2	2.0	–
RaP [46]	45.4	74.8	83.6	2.0	–
CLIP for feature representations					
CLIP4Clip-tightTransf [11]	40.0	71.5	82.1	2.0	13.3
LAFF [48]	45.4	76.0	84.6	–	–
CLIP4Clip-seqLSTM [11]	46.2	75.3	84.5	2.0	10.2
CLIP4Clip-meanP [11]	46.2	76.1	84.6	2.0	10.0
DiffusionRet [52]	46.6	75.9	84.1	2.0	15.7
CAMoE [14]	46.9	76.1	85.5	–	9.8
CLIP2Video [12]	47.0	76.8	85.9	2.0	9.6
X-Pool [19]	47.2	77.4	86.0	2.0	9.3
DiCoSA [53]	**47.4**	76.8	86.0	2.0	9.1
Ours	46.6	**77.8**	**86.8**	**2.0**	**9.1**

ships, and in videos as objects and their relationships. Thus, it can be clustered into three semantic centers. Moreover, the semantic guidance mechanism adjusts the weight of these parts throughout the entire event, so the model can focus more on the more important parts.

Table 5 Text-to-video retrieval performance comparisons to SOTAs on the **ActivityNet Captions** dataset

Method	text-to-video				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Non-CLIP for feature representations					
CE [2]	18.2	47.7	–	6.0	23.1
ClipBERT [41]	21.3	49.0	63.5	6.0	–
T2VLAD [6]	23.7	55.5	–	4.0	–
MMT [3]	28.7	61.4	–	3.3	16.0
CLIP for feature representations					
CLIP4Clip-seqLSTM [11]	40.1	72.2	–	2.0	7.3
CLIP4Clip-seqTransf [11]	40.5	72.4	–	2.0	7.5
CLIP4Clip-meanP [11]	40.5	72.4	–	2.0	7.4
TS2-Net [21]	41.0	**73.6**	84.5	2.0	8.4
HBI [54]	42.2	73.0	84.6	2.0	6.6
Ours	**42.5**	73.1	**85.0**	**2.0**	**6.5**

Table 6 Text-to-video retrieval performance comparisons to SOTAs on the DiDeMo dataset

Method	text-to-video				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Non-CLIP for feature representations					
CE [2]	16.1	41.1	–	8.3	43.7
LaT [44]	32.6	61.3	71.6	3.0	–
MILES [45]	36.6	63.9	74.0	3.0	–
BridgeFormer [32]	37.0	62.2	73.9	3.0	–
RaP [46]	42.9	71.2	80.2	2.0	–
CLIP for feature representations					
CLIP4Clip-tightTransf [11]	25.8	52.8	66.3	5.0	27.3
CLIP4Clip-seqTransf [11]	42.8	68.5	79.2	2.0	18.9
CLIP4Clip-seqLSTM [11]	43.4	69.9	80.2	2.0	17.5
CLIP4Clip-meanP [11]	43.4	70.2	80.6	2.0	17.5
TS2-Net [21]	41.8	71.6	82.0	2.0	14.8
CAMoE [14]	43.8	71.4	79.9	2.0	16.3
HiSE [33]	44.1	69.9	80.3	2.0	–
Ours	**44.4**	**75.2**	**82.9**	**2.0**	**12.4**

4.4.3 Parameter Sensitivity of Weighting Parameters

The parameter α indicates the importance of L_{local}. We evaluate the scale range setting $\alpha \in [0.1, 0.5]$ as shown in Fig. 6b. When $\alpha = 0.2$, R@1 in both directions achieves the best performance. So we set $\alpha = 0.2$ as the default in practice. We present the influence of hyper-parameter β in Fig. 6c. We find that the model performs best when β is set to 0.1. So we adopt $\beta = 0.1$ in practice.

Table 7 Ablation study of each granularity on the MSRVTT-9k (text-to-video)

Method				text-to-video				
Multi-grained			Semantic- guided	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Fgrained	Local	Global						
√				47.0	**75.8**	82.8	2.0	12.8
√	√			46.9	75.4	82.9	2.0	12.3
√	√	√		47.3	75.1	83.2	2.0	12.0
√	√		√	47.3	75.2	**83.7**	2.0	13.2
√	√	√	√	**47.7**	74.5	**83.7**	**2.0**	**11.9**

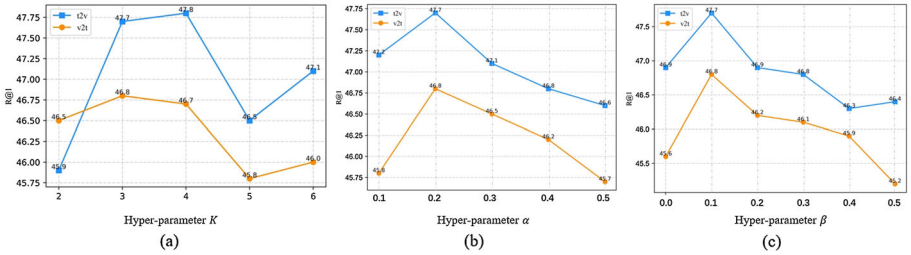

Fig. 6 Parameter sensitivity study on MSRVTT-9k. Effect of (**a**) number of centers K; (**b**) the weighting parameter α; (**c**) the weighting parameter β

Table 8 Effect of MLP layers in global-guided mechanism on MSRVTT-9k

MLP layers	text-to-video					video-to-text				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
1FC	46.3	74.6	**84.0**	2.0	12.0	46.3	75.5	**84.5**	2.0	8.4
2FC	**47.7**	74.5	83.7	2.0	**11.9**	**46.8**	**75.6**	84.3	2.0	**8.1**
3FC	46.9	**74.7**	**84.0**	2.0	12.3	45.8	74.7	**84.5**	2.0	8.5

4.4.4 Effect of MLP Layers in the Semantic Guidance Mechanism

In the semantic-guided local-level text-video alignment, we adopt the simple and efficient MLP to generate weights for semantic centers, which is composed of linear layers and ReLU. We test MLP consisting of 1, 2, and 3 linear layers, respectively. As shown in Table 8, we found that each of them achieves good performance, while the MLP with two linear layers performs better, scoring 47.7% and 46.8% on R@1 of text-to-video and video-to-text respectively.

4.4.5 Visualization Results

Figure 7 shows the quantitative results of our retrieval method. It can be seen from the retrieved videos that our model can capture subtle information and gain more accurate retrieval results than the model that only focuses on global semantics. Specifically, for #$query1$, compared to the model without semantic guidance mechanism, MGSGA focuses more on more discriminative semantics such as "classroom" and "children", rather than just on "man" and "speak", thus excluding the top4 video in "w/o semantic guided alignment". For #$query2$,

Fig. 7 Quantitative results of text-video retrieval. The blue number is the similarity score. Each video is represented by 3 randomly sampled frames. The ground-truth video is marked with a red box and a red star. The left column is the retrieval results of the multi-grained and semantic-guided text-video alignment proposed in this paper. The middle column is the retrieval results of the model without semantic guidance mechanism. The right column is the retrieval results only using global semantics (CLIP4Clip). (Color figure online)

the five retrieved videos of our model all embody "people" and "beach". Due to the presence of "beach" semantics, top 5 video in "MGSGA" is ranked as top 4 in "w/o semantic-guided alignment", but the information that "singing" is more important in the overall semantics of this text is ignored. The last three of the first five videos retrieved by "w/o multi-grained alignment" are all related to text but not strongly related. Due to the multi-grained and semantic-guided text-video alignment, MGSGA performs well in cross-modal association, thus obtaining better retrieval results.

5 Conclusion

In this paper, we explore the multi-grained semantic correspondence between text and video, A multi-grained and semantic-guided text-video alignment framework is proposed for cross-modal retrieval, which aligns the embeddings of text and video in the common semantic space from fine-grained based on frames and words, local-level based on semantic centers, and the global-level. Specially, we introduce the semantic guidance mechanism provided by the summary semantics in the local-level alignment module. It makes the model focus on the more important semantic center from the perspective of summary semantics. Performance on MSRVTT, MSVD, ActivityNet Captions, and DiDeMo show that our method exceeds most existing methods. We also conduct ablation experiments and qualitative analysis to validate our method.

The current research on text-video retrieval only exploits the annotation inside the dataset. In the future, we will consider introducing external knowledge into the text-video cross-modal retrieval model, and integrating external knowledge and the information inside the dataset in the architecture to improve the retrieval performance.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection and experiments were performed by Jiayao Qian. Analysis and conclusion were performed by Xiaoyu Wu, Jiayao Qian and Lulu Yang. The first draft of the manuscript was written by Jiayao Qian and all authors commented and modified on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work is supported by National Key R&D Program of China (No. 2021YFF0900701), National Natural Science Foundation of China (No.61801441), and in part by discipline construction project of "Beijing top notch" discipline (Internet information of Communication University of China).

Availability of Data and Materials The datasets we used are open and can be downloaded through official ways.

Code availability The code required to reproduce these findings cannot be shared at this time as the code also forms part of an ongoing study.

Declarations

Conflict of interest The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Appendix A Video-to-Text Retrieval Results

Table 9, 10, 11, and 12 present the video-to-text retrieval results of MGSGA on MSR-VTT-9k, MSVD, and DiDeMo. Table 13 shows the ablation study of each granularity on the MSRVTT-9k(video-to-text).

Table 9 Video-to-text retrieval performance comparisons to SOTAs on the MSRVTT-9k dataset

Method	video-to-text				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Non-CLIP for feature representations					
CE [2]	20.6	50.3	64.0	5.3	25.1
MMT [3]	27.0	57.5	69.7	3.7	21.3
T2VLAD [6]	31.8	60.0	71.1	3.0	–
LaT [44]	35.4	61.3	72.4	3.0	–
CLIP for feature representations					
CLIP [10]	27.2	51.7	62.6	5.0	–
CLIP4Clip-tightTransf [11]	40.6	69.5	79.5	2.0	13.6
CLIP4Clip-seqTransf [11]	42.7	70.9	80.6	2.0	11.6
CLIP4Clip-seqLSTM [11]	42.8	71.0	80.4	2.0	12.3
CLIP4Clip-meanP [11]	43.1	70.5	81.2	2.0	12.4
CLIP2Video [12]	43.5	72.3	82.1	2.0	10.2
X-Pool [19]	44.4	73.3	84.0	2.0	9.0
CAMoE [14]	45.1	72.4	83.1	2.0	10.0
TokenFlow [49]	45.4	73.5	83.7	2.0	10.0
EMCL-Net [50]	46.5	73.5	83.5	2.0	–
HiSE [33]	46.6	73.3	82.3	2.0	–
TABLE [51]	**47.2**	74.2	84.2	2.0	11.0
Ours	46.8	**75.6**	**84.3**	**2.0**	**8.1**

Table 10 Video-to-text retrieval performance comparisons to SOTAs on the MSVD dataset

Method	video-to-text				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Non-CLIP for feature representations					
ViSERN [31]	24.3	46.2	59.5	7.0	34.6
LaT [44]	39.7	75.6	85.4	2.0	–
CLIP for feature representations					
CLIP4Clip-seqLSTM [11]	52.5	74.0	78.1	1.0	14.7
CLIP4Clip-tightTransf [11]	54.3	85.3	91.0	1.0	6.0
CLIP4Clip-meanP [11]	56.6	79.7	84.3	1.0	7.6
CLIP2Video [12]	58.7	85.6	91.6	1.0	**4.3**
DiffusionRet [52]	**60.3**	86.4	92.0	1.0	4.5
Ours	**60.2**	**87.2**	**92.7**	**1.0**	4.8

Table 11 Video-to-text retrieval performance comparisons to SOTAs on the ActivityNet Captions dataset

Method	video-to-text				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Non-CLIP for feature representations					
CE [2]	17.7	46.6	–	6.0	24.4
T2VLAD [6]	24.1	56.6	–	4.0	–
MMT [3]	28.9	61.1	–	4.0	17.1
CLIP for feature representations					
CLIP4Clip-meanP [11]	42.5	74.1	85.8	2.0	6.6
CLIP4Clip-seqLSTM [11]	42.6	73.4	85.6	2.0	6.7
CLIP4Clip-seqTransf [11]	41.4	73.7	85.3	2.0	6.5
HBI [54]	42.4	73.0	**86.0**	2.0	6.5
Ours	**42.4**	**73.2**	85.8	**2.0**	**6.4**

Table 12 Video-to-text retrieval performance comparisons to SOTAs on the DiDeMo dataset

Method	video-to-text				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Non-CLIP for feature representations					
CE [2]	15.6	40.9	–	8.2	42.4
LaT [44]	32.7	61.1	72.7	3.0	–
CLIP for feature representations					
CLIP4Clip-tightTransf [11]	21.5	51.1	64.8	5.0	22.4
CLIP4Clip-seqTransf [11]	41.4	68.2	79.1	2.0	12.4
CLIP4Clip-seqLSTM [11]	42.4	69.2	79.2	2.0	11.8
CLIP4Clip-meanP [11]	42.5	70.6	80.2	2.0	11.6
HiSE [33]	**43.8**	70.4	79.2	2.0	–
Ours	43.2	**73.2**	**82.7**	**2.0**	**8.7**

Table 13 Ablation study of each granularity on the MSRVTT-9k **(video-to-text)**

Method				**video-to-text**				
Multi-grained			Semantic- guided	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
fgrained	local	global						
√				45,6	75.3	84.4	2.0	8.9
√	√			45.6	75.4	84.5	2.0	8.9
√	√	√		45.9	**75.7**	**84.9**	2.0	8.7
√	√		√	46.0	75.5	84.4	2.0	8.8
√	√	√	√	**46.8**	75.6	84.3	**2.0**	**8.1**

References

1. Miech A, Laptev I, Sivic J (2018) Learning a text-video embedding from incomplete and heterogeneous data. arXiv preprint arXiv:1804.02516
2. Liu Y, Albanie S, Nagrani A, Zisserman A (2019) Use what you have: video retrieval using representations from collaborative experts. In: 30th British machine vision conference, p 279
3. Gabeur V, Sun C, Alahari K, Schmid C (2020) Multi-modal transformer for video retrieval. In: European conference on computer vision, pp 214–229
4. 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
5. Kunitsyn A, Kalashnikov M, Dzabraev M, Ivaniuta A (2022) Mdmmt-2: multidomain multimodal transformer for video retrieval, one more step towards generalization. arXiv preprint arXiv:2203.07086
6. 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
7. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
8. Gemmeke JF, Ellis DP, Freedman D, Jansen A, Lawrence W, Moore RC, Plakal M, Ritter M (2017) Audio set: an ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing, pp 776–780
9. Xie S, Sun C, Huang J, Tu Z, Murphy K (2018) Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification. In: Proceedings of the European conference on computer vision, pp 305–321
10. 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
11. 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
12. Fang H, Xiong P, Xu L, Chen Y (2021) Clip2video: mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097
13. Gao Z, Liu J, Chen S, Chang D, Zhang H, Yuan J (2021) Clip2tv: an empirical study on transformer-based methods for video-text retrieval. arXiv preprint arXiv:2111.05610
14. 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
15. Wu P, He X, Tang M, Lv Y, Liu J (2021) Hanet: hierarchical alignment networks for video-text retrieval. In: Proceedings of the 29th ACM international conference on multimedia, pp 3518–3527
16. Satar B, Hongyuan Z, Bresson X, Lim JH (2021) Semantic role aware correlation transformer for text to video retrieval. In: 2021 IEEE international conference on image processing, pp 1334–1338
17. Chen S, Zhao Y, Jin Q, Wu Q (2020) Fine-grained video-text retrieval with hierarchical graph reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10638–10647
18. Madasu A, Aflalo E, Stan GBM, Tseng S, Bertasius G, Lal V (2023) Improving video retrieval using multilingual knowledge transfer. In: Advances in information retrieval-45th European conference on information retrieval, pp 669–684
19. 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
20. Zhao S, Zhu L, Wang X, Yang Y (2022) Centerclip: token clustering for efficient text-video retrieval. In: The 45th international ACM SIGIR conference on research and development in information retrieval, pp 970–981
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
22. Zhang B, Jin X, Gong W, Xu K, Zhang Z, Wang P, Shen X, Feng J (2023) Multimodal video adapter for parameter efficient video text retrieval. arXiv preprint arXiv:2301.07868
23. Jiang H, Zhang J, Huang R, Ge C, Ni Z, Lu J, Zhou J, Song S, Huang G (2022) Cross-modal adapter for text-video retrieval. arXiv preprint arXiv:2211.09623
24. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems 30, Annual conference on neural information processing systems, pp 5998–6008
25. Chen Y, Li L, Yu L, Kholy AE, Ahmed F, Gan Z, Cheng Y, Liu J (2020) UNITER: universal image-text representation learning. In: European conference on computer vision, pp 104–120

26. Li X, Yin X, Li C, Zhang P, Hu X, Zhang L, Wang L, Hu H, Dong L, Wei F, Choi Y, Gao J (2020) Oscar: object-semantics aligned pre-training for vision-language tasks. In: European conference on computer vision, pp 121–137
27. Messina N, Amato G, Esuli A, Falchi F, Gennaro C, Marchand-Maillet S (2021) Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. ACM Trans Multimed Comput Commun Appl 17(4):128–112823
28. Ren S, He K, Girshick RB, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems 28: annual conference on neural information processing systems 2015, pp 91–99
29. Guo J, Wang M, Zhou Y, Song B, Chi Y, Fan W, Chang J (2023) HGAN: hierarchical graph alignment network for image-text retrieval. IEEE Trans Multimed 25:9189–9202
30. 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
31. Feng Z, Zeng Z, Guo C, Li Z (2020) Exploiting visual semantic reasoning for video-text retrieval. In: Proceedings of the twenty-ninth international joint conference on artificial intelligence, pp 1005–1011
32. 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
33. Wang H, Xu D, He D, Li F, Ji Z, Han J, Ding E (2022) Boosting video-text retrieval with explicit high-level semantics. In: Proceedings of the 30th ACM international conference on multimedia, pp 4887–4898
34. Arandjelovic R, Gronat P, Torii A, Pajdla T, Sivic J (2016) NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5297–5307
35. Oord Avd, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748
36. 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
37. Chen D, Dolan WB (2011) Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, pp 190–200
38. Krishna R, Hata K, Ren F, Fei-Fei L, Carlos Niebles J (2017) Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision, pp 706–715
39. Anne Hendricks L, Wang O, Shechtman E, Sivic J, Darrell T, Russell B (2017) Localizing moments in video with natural language. In: Proceedings of the IEEE international conference on computer vision, pp 5803–5812
40. 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
41. 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
42. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: 3rd international conference on learning representations, ICLR 2015
43. Bogolin S, Croitoru I, Jin H, Liu Y, Albanie S (2022) Cross modal retrieval with querybank normalisation. In: IEEE/CVF conference on computer vision and pattern recognition, pp 5184–5195
44. Bai J, Liu C, Ni F, Wang H, Hu M, Guo X, Cheng L (2022) Lat: latent translation with cycle-consistency for video-text retrieval. arXiv preprint arXiv:2207.04858
45. Ge Y, Ge Y, Liu X, Wang J, Wu J, Shan Y, Qie X, Luo P (2022) Miles: visual bert pre-training with injected language semantics for video-text retrieval. In: European conference on computer vision, pp 691–708
46. Wu X, Gao C, Lin Z, Wang Z, Han J, Hu S (2022) Rap: redundancy-aware video-language pre-training for text-video retrieval. In: Findings of the Association for computational linguistics, pp 3036–3047
47. Lu H, Fei N, Huo Y, Gao Y, Lu Z, Wen J-R (2022) Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 15692–15701
48. Hu F, Chen A, Wang Z, Zhou F, Dong J, Li X (2022) Lightweight attentional feature fusion: a new baseline for text-to-video retrieval. In: European conference on computer vision, pp 444–461
49. Zou X, Wu C, Cheng L, Wang Z (2022) Tokenflow: rethinking fine-grained cross-modal alignment in vision-language retrieval. arXiv preprint arXiv:2209.13822

50. Jin P, Huang J, Liu F, Wu X, Ge S, Song G, Clifton DA, Chen J (2022) Expectation-maximization contrastive learning for compact video-and-language representations. In: Advances in neural information processing systems, vol 35, pp 30291–30306
51. Chen Y, Wang J, Lin L, Qi Z, Ma J, Shan Y (2023) Tagging before alignment: integrating multi-modal tags for video-text retrieval. In: Thirty-Fifth conference on innovative applications of artificial intelligence, pp 396–404
52. Jin P, Li H, Cheng Z, Li K, Ji X, Liu C, Yuan L, Chen J (2023) Diffusionret: generative text-video retrieval with diffusion model. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2470–2481. https://doi.org/10.48550/arXiv.2303.09867
53. Jin P, Li H, Cheng Z, Huang J, Wang Z, Yuan L, Liu C, Chen J (2023) Text-video retrieval with disentangled conceptualization and set-to-set alignment. In: Proceedings of the thirty-second international joint conference on artificial intelligence, pp 938–946
54. Jin P, Huang J, Xiong P, Tian S, Liu C, Ji X, Yuan L, Chen J (2023) Video-text as game players: hierarchical banzhaf interaction for cross-modal representation learning. In: IEEE/CVF conference on computer vision and pattern recognition, pp 2472–2482