Fine-grained Cross-modal Alignment Network for Text-Video Retrieval

Ning Han
Hunan University
ninghan@hnu.edu.cn

Jingjing Chen*
Fudan University
chenjingjing@fudan.edu.cn

Guangyi Xiao
Hunan University
guangyi.xiao@gmail.com

Hao Zhang
City University of Hong Kong
zhanghaoinf@gmail.com

Yawen Zeng
Hunan University
yawenzeng11@gmail.com

Hao Chen*
Hunan University
chenhao@hnu.edu.cn

ABSTRACT

Despite the recent progress of cross-modal text-to-video retrieval techniques, their performance is still unsatisfactory. Most existing works follow a trend of learning a joint embedding space to measure the distance between global-level or local-level textual and video representation. The fine-grained interactions between video segments and phrases are usually neglected in cross-modal learning, which results in suboptimal retrieval performances. To tackle the problem, we propose a novel Fine-grained Cross-modal Alignment Network (FCA-Net), which considers the interactions between visual semantic units (i.e., sub-actions/sub-events) in videos and phrases in sentences for cross-modal alignment. Specifically, the interactions between visual semantic units and phrases are formulated as a link prediction problem optimized by a graph auto-encoder to obtain the explicit relations between them and enhance the aligned feature representation for fine-grained cross-modal alignment. Experimental results on MSR-VTT, YouCook2, and VATEX datasets demonstrate the superiority of our model as compared to the state-of-the-art method.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**.

KEYWORDS

Fine-grained Cross-modal Alignment; Graph Auto-Encoder; Text-Video Retrieval

ACM Reference Format:
Ning Han, Jingjing Chen, Guangyi Xiao, Hao Zhang, Yawen Zeng, and Hao Chen. 2021. Fine-grained Cross-modal Alignment Network for Text-Video Retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475241

*Jingjing Chen and Hao Chen are the corresponding authors.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475241

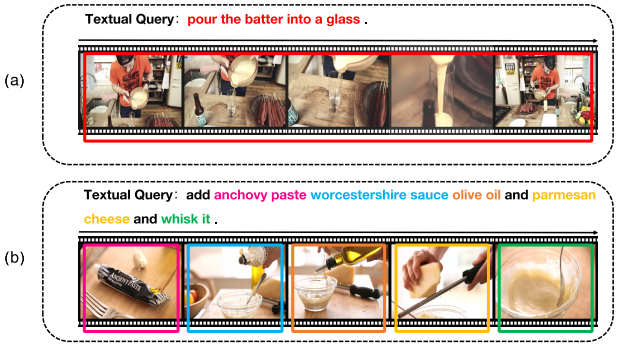

Figure 1: Example of text-video retrieval pairs. (a) shows a single action/event retrieval example. (b) presents a complex multiple sub-actions/sub-events retrieval example. However, existing approaches work well on a single action/event retrieval scenario but are not satisfactory for more realistic cases that involve coherent sub-action/sub-event segments of video and continuous semantic phrases of text. Therefore, we perform the visual-semantic-units-to-phrases interactions of video and text for text-video retrieval.

1 INTRODUCTION

Multimedia data (e.g., image, text, audio, and video) on the web have undergone exponential growth in recent years with the popularization of social media platforms (Facebook, Twitter) and video-sharing platforms (YouTube, TikTok). Therefore, users are overloaded by massive multimodal data [5, 22, 23, 30, 45], and this trend requires the exploration of advanced techniques for retrieving useful information across different modalities. As one of the most popular research topics in this domain, cross-modal retrieval between texts and videos has received growing interest from academics and the industry.

Cross-modal text-video retrieval is inherently a challenging problem. The major issue is the modality gap between texts and videos that hinders the alignment of related samples from different modalities. As a solution to this challenge, existing approaches [10, 12, 21, 28–31] mainly learn a common latent space to directly measure the distance between global- or local-level textual and video representations. However, these approaches coarsely capture the correspondence between modalities and are thus unable to capture the fine-grained interactions between video and text.

To better capture such fine-grained correspondences, recent studies have investigated cross-modal interaction methods [7, 40] that are based on different attention mechanisms to align the semantic spaces between videos and texts. Moreover, cross-modal interaction methods have been proven to be effective for image-text retrieval [19, 33], which can discover the fine-grained correspondence and thus achieves state-of-the-art performance. Yet, due to the large heterogeneity gap between videos and texts, existing attention-based models, e.g. [7], may not well capture coherent sub-action/sub-event segments of video. Meanwhile, existing works have largely neglected the fine-grained interactions between sub-action/sub-event segments of video and phrases in cross-modal learning. For realistic cases involving multiple semantically coherent sub-actions/sub-events, text-video retrieval results are not satisfactory.

In this work, we address the problem of cross-modal retrieval by performing fine-grained cross-modal alignment between videos and texts. Specifically, we propose to model the interactions between video segments and text phrases for fine-grained alignment. We present an example to explain the necessity of doing so. For example, for text-video retrieval based on a simple query (Figure 1(a)), e.g., "pour the batter into a glass," a common retrieval system considering motion and semantical compositions returns relevant videos containing the motion "pour" and objects "batter" and "glass". However, a complex query (Figure 1(b)) may consist of multiple semantic phrases that correspond to multiple coherent actions (i.e., "add anchovy paste," "add worcestershire sauce," "add olive oil," "add parmesan cheese," "whisk it"). During the matching procedure, they rely more on identifying the most discriminative global feature of the text and video than on establishing fine-grained features at the segment and phrase levels. In this case, comprehensive context and correspondences are likely to be poorly understood. Although fine-grained modeling for video and text is necessary, it progresses slowly because of two major stumbling blocks. First, videos in real life contain a diverse range of compositional objects with complex mutual interactions, and each object/action bears a different level of importance when grounded with textual entities. Therefore, modeling fine-grained visual contents with regard to grounded textual semantics is a non-trivial task. Second, videos usually contain richer semantical content than a textual query (i.e., "A video is worth a thousand words"). A typical approach to discovering information embedded in video frames is to extract frame-level features and aggregate the frame-level features into a video-level feature. However, this process leads to serious information redundancy because of some noisy or meaningless semantics with respect to the textual query. Therefore, discovering and enhancing valuable semantic information while restraining useless semantic information between video and text is a complicated task.

To this end, we propose a Fine-grained Cross-modal Alignment Network (FCA-Net) for cross-modal text-video retrieval. The proposed network considers the interactions between visual semantic units in videos and phrases in sentences for cross-modal alignment. Figure 2 presents an overview of the proposed FCA-Net. Specifically, we first decode the videos into coherent visual semantic units on the basis of an ordered clustering of embedded framewise video features. Meanwhile, we employ the StanfordCoreNLP toolkit [26] to obtain a set of phrases from the sentence and adopt Bidirectional Encoder Representations from Transformers (BERT) [8] to extract phrase-level features. Then, with the new representation, the interactions between visual semantic units and phrases are formulated as a link prediction problem optimized by a Graph Auto-Encoder (GAE) [2, 15]. In this way, the explicit relations between visual semantic units and phrases are obtained and utilized to enhance the embedding features for cross-modal alignment. The whole framework is trained in an end-to-end manner with triplet rank loss as the loss function.

Our contributions are summarized as below:

- We highlight the importance of exploring the correspondences of videos and texts at multiple granularities and propose a novel FCA-Net that considers the fine-grained cross-modal interactions between visual semantic units and phrases for cross-modal text-video retrieval.
- We formulate the fine-grained interactions between different modalities as a link prediction problem and introduce GAE to model such interactions.
- We conduct extensive experiments on three standard benchmarks and verify the effectiveness of our proposed method.

2 RELATED WORK

2.1 Text-Video Retrieval

We briefly review the representative methods of cross-modal text-video retrieval, which follows a trend of learning a joint embedding space to measure the distance between textual and video representation. These methods roughly fall into two categories: 1) cross-modal interaction-free methods [9, 10, 12, 21, 28, 30–32, 39] and 2) cross-modal interaction methods [7, 13, 24, 34, 40, 41, 44].

Cross-modal interaction-free methods typically encode video and textual queries and accordingly map them into a common latent space where the video-text similarity can be measured directly with ranking loss variants. For instance, Miech et al. [28, 30] propose a large HowTo100M [30] dataset to improve video-text representations by leveraging large-scale pretraining. They also find that leveraging the contrastive loss can address visually misaligned narrations from uncurated instructional videos and improve video-text representations. Yang et al. [39] present a base tree-augmented cross-modal encoding model, which designs a tree-augmented query encoder to derive structure-aware query representation and a temporal attentive video encoder to model the temporal characteristics of videos. Dong et al. [9, 10] adopt three branches, i.e., mean pooling, Bi-GRU and CNN, to encode sequential videos and texts and learn a hybrid common space for video-text similarity prediction. Patrick et al. [32] present a generative objective to improve the instance discrimination limitations of contrastive learning and thereby enhance its performance in text-video retrieval. Other works [12, 21, 31] achieve performance improvements by fusing other modalities, such as motion and audio features, into video embedding.

Cross-modal interaction methods, which utilize deep networks to learn complex nonlinear transformation and perform interactions between videos and textual sentences, adopt traditional ranking loss to maximize their correlation over the network outputs. Yu et al. [41] adopt a concept word detector based on a semantic

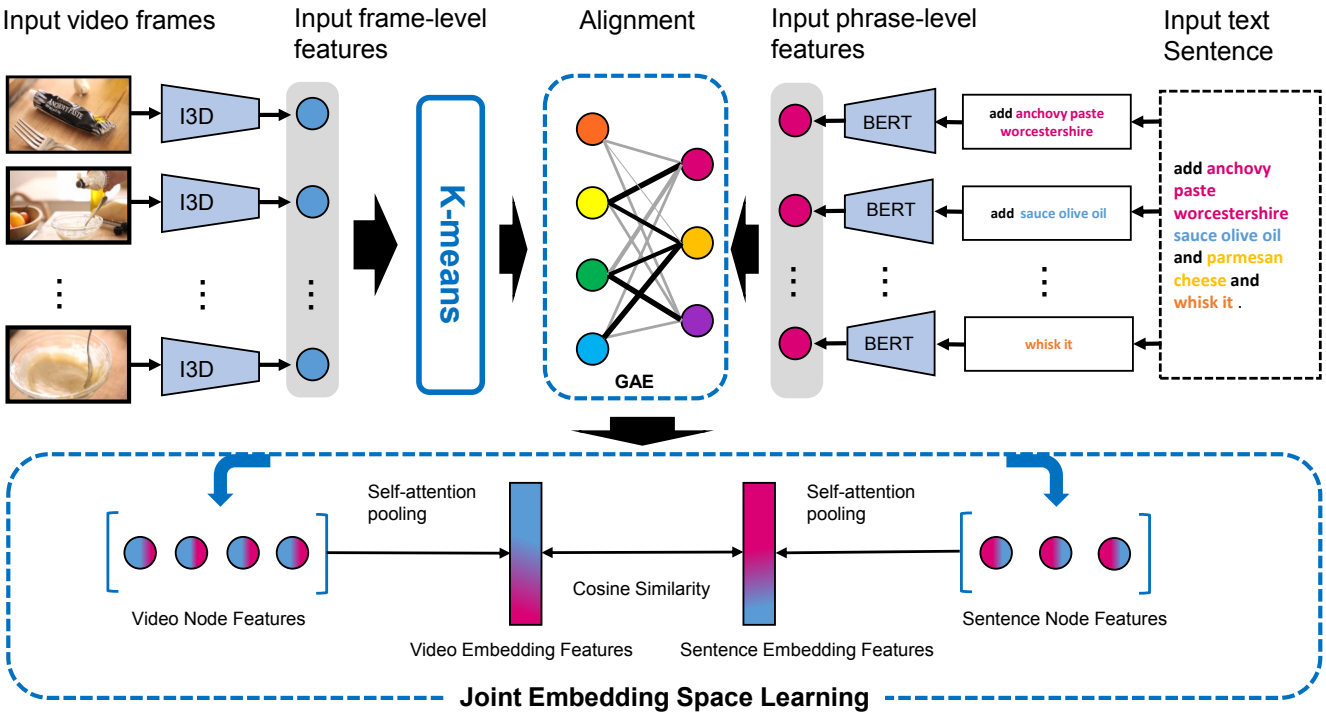

Figure 2: The overall framework of our proposed FCA-Net. We first decode the videos into visual semantic units based on an ordered clustering of the embedded framewise video features. We then parse textual sentences into phrases and extract phrase features by BERT. Finally, we implement a joint embedding learning with a link prediction strategy based on GAE and triplet ranking loss for cross-modal text-video retrieval.

attention mechanism in enhancing video representation and further performing text-to-video retrieval. Yu et al. [40] develop a joint sequence fusion model for the sequential interaction of video and text representations and further predict the similarities on the basis of fused features. Song et al. [34] present a polysemous instance embedding network to fuse globally and locally guided features of texts and videos for the polysemous problem. Chen et al. [7] propose an attention-based hierarchical graph reasoning model, which decomposes texts into events, actions, and entities and aligns texts with videos at three levels for video-text matching. Ging et al.[13] propose a cooperative hierarchical transformer to effectively encode texts and videos and a cross-modal cycle-consistency loss to perform semantic alignment between vision and text features. Other works [24, 44] extend the BERT model to feed visual and textual tokens as input to learn high-level semantic text-video representations for text-video retrieval.

Although these approaches [7, 40] consider the fine-grained modeling between frames and words or the three levels of videos and texts (e.g., events, actions, and entities) , they ignore the fine-grained modeling between video segments and phrases. Our work focuses on exploiting the fine-grained alignment between discriminative visual semantic units and phrases. The experimental results show that our fine-grained cross-modal interaction model benefits our retrieval accuracy, which significantly outperforms several state-of-the-art baselines.

2.2 Multimodal Information Alignment

Multimodal information alignment is one of the key technologies of cross-modal retrieval. It aims to find the correspondence between instance subcomponents from different modalities, and it has been demonstrated to be effective in different retrieval tasks, such as video moment retrieval [20, 42], text-image retrieval [19, 27, 33], and text-video retrieval [12, 28]. A straightforward approach to multimodal information alignment is to learn a similarity or distance metric for correlating cross-modal data. A classical method series includes the CCA-based approaches [1, 17], which adopt global alignment to allow the mapping of different modalities. The attention mechanism has been regarded as an effective method for aligning the semantic spaces between multiple modalities as it uncovers valuable components and avoids noise. Lee et al. [19] present a stacked cross-attention mechanism for aligning image regions with words in a sentence. Peng et al. [33] propose the multilevel adaptive visual-textual alignment to explore the visual-textual alignments from multiple levels. Messina et al. [27] propose the transformer encoder reasoning and alignment network, which forces a fine-grained word-region alignment for cross-modal information retrieval. Unlike these efforts, our work uses a link prediction strategy that is based on a Graph Auto-Encoder (GAE)[2, 15] for multimodal information alignment.

3 METHOD

As shown in Figure 2, the proposed framework consists of three modules: 1) video embedding learning, which involves extracting visual semantic unit features at multiple segment levels; 2) text embedding learning, which involves extracting a set of phrase features from the text query through the use of BERT; and 3) joint embedding learning, which involves integrating a link prediction strategy based on GAE to align key visual semantic units with phrases and optimize textual and video features with triplet ranking loss.

3.1 Video Embedding Learning

Video Preprocessing. Given a video clip $\mathcal{V}$, we sample uniformly a sequence of video frames $\{v_1, \cdots, v_N\}$ from $\mathcal{V}$ with a pre-specified video temporal resolution; here, N is the video length. We extract the framewise RGB features by using pre-trained I3D [4]. In this process, we utilize a video clip with frames $[t-4, t+4]$ as a feature representing frame t-th. The video clip features are represented as $F = \{f_t\}_{t=1}^{N}$, where $f_t \in \mathbb{R}^{d_v}$ denotes the feature vector of the t-th frame.

Clustering. To decode the video into coherent visual semantic units, we cluster the original framewise features of the video into K clusters by adopting k-means [25]. K is defined as the maximum number of possible sub-actions/sub-events as they occur in the ground truth. Specifically, we evaluate the clusters obtained at each video feature clustering and for each possible number of clusters by using the Calinski-Harbasz score [3]; large values indicate higher-quality clustering. Furthermore, we temporally segment each video feature $\{f_t\}_{t=1}^{N}$ separately, i.e., we assign each frame feature f_t to one of the clusters $C = \{C_i\}_{i=1}^{K}$. Following [18], we utilize the time stamp of each frame to order the clusters and obtain the ordered clusters for the decoding of each video.

To transform each cluster feature representation C_i into vectors of the same dimension for cross-modal alignment. We employ a mean pooling operation on each cluster feature and treat the pooling result as each cluster feature $f_{segment}^{i} \in \mathbb{R}^{d_v}$. Then, the projection is learned through transformation matrices W_v, as following:

$$\phi_v^i = W_v f_{segment}^i + b_v, \tag{1}$$

where $\phi_v^i \in \mathbb{R}^{d_*}$ is the embedding features of each cluster of video, $W_v \in \mathbb{R}^{d_* \times d_v}$ is the learned transformation matrix and $b_v \in \mathbb{R}^{d_*}$ is the bias term. Finally, the embedding feature representation of the video clip is a combination of all visual semantic units' representations: $\Psi = \{\phi_v^i\}_{i=1}^{K}$.

3.2 Text Embedding Learning

Given a text query Q consisting of M words $\{w_1, \cdots, w_M\}$, we employ the StanfordCoreNLP toolkit [26] to obtain a set of phrases $S = \{s_j\}_{j=1}^{J}$ from query Q, in which the corresponding verb is added to the phrase without the verb by preprocessing. To learn the contextual relations between the words in each phrase s_j, we feed the phrase consisting of several words into a pre-trained BERT [8] language representation model and take the hidden state of the first token in the last layer to represent the information of the entire input phrase. The process yields 768-D features $f_{phrase}^{j} \in \mathbb{R}^{d_t}$. Following previous work [29], we transform each phrase representation f_{phrase}^{j} into an embedding feature $\phi_t^j \in \mathbb{R}^{d_*}$ by using a gated embedding module [29]. The embedding feature representation of the text sentence is also a combination of all phrases' representations: $\Phi = \{\phi_t^j\}_{j=1}^{J}$.

3.3 GAE for Cross-Modal Alignment.

As described in Sections 3.1 and 3.2, video embedding learning and text embedding learning yield many-to-many visual and textual semantic units. The aforementioned attention mechanism exploits the cross-modal interactions by calculating the dot product similarities between queries and videos. However, the projected queries and videos may contain noisy or meaningless information. To adaptively discover valuable information and restrain the useless ones, we model the interactions as a link prediction problem. Specifically, we design a GAE-based undirected bipartite graph for the alignment of many-to-many visual and textual semantic units. The designed graph can help discover the explicit relations between visual and textual semantic units and further capture cross-modal semantic alignments accurately. Formally, we define our undirected bipartite graph as $G = \{V, E, X\}$, where $V = \{x_1, \cdots, x_n\}$ is a set of nodes with all visual and textual semantic units and E is the set of link weights among each node that can be represented by an adjacency matrix $A = [a_{ij}] \in \mathbb{R}^{n \times n}$. X is the feature matrix of all nodes, i.e., $X = \Psi \oplus \Phi \in \mathbb{R}^{(K+J) \times d_*}$, where $\oplus$ is a concatenation operation and n and d_* are the number of nodes and dimensions of X, respectively. In previous link prediction problems, the weights of the observed links are given by the data. By contrast, our weights are initially estimated by semantic similarity. Our target is to learn the latent representations of nodes and obtain accurate link weights in graph $G = \{V, E, X\}$. GAE consists of a two-layer graph convolutional encoder and an inner product decoder. It can align the semantic spaces between visual and textual semantic units and obtain enhanced video and text features. Next, we explain in detail the two components.

Graph Convolutional Encoder. The graph convolutional encoder aims to transform raw video and text features to augmented video and text features with the structure of the constructed graph. The Graph Convolutional Network (GCN) [16] takes a graph as input, performs computations over the structure, and returns the updated features of each object node as the output. In the bipartite graph G, we project the input features X into an interaction space by adopting a nonlinear transformation operation (in Section 3.1 and 3.2). To dynamically uncover related phrases of visual semantic units (or related visual semantic units of phrases), we construct the edge relationship by calculating the inner product similarity [35]:

$$A = \varphi(X)\,\varphi(X)^T \odot M', \tag{2}$$

where A is defined as the initialized adjacency matrix, φ is a nonlinear transformation operation for learning the edge link weights between heterogeneous nodes, $M' \in \{0,1\}^{n \times n}$ is the mask matrix for constructing bipartite graph, and $\odot$ is the element-wise product. The GCN can be compounded of several layers stacked together. A single graph convolutional layer is defined as:

$$Z = \sigma\left(\widetilde{D}^{-\frac{1}{2}} \widetilde{A D}^{-\frac{1}{2}} X W\right), \tag{3}$$

where $\widetilde{A} = A + I$ is the adjacency matrix of the graph with added self-loops, $\widetilde{D}^{-\frac{1}{2}}$ is its diagonal degree matrix with $\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$, $W \in \mathbb{R}^{d_* \times d_*}$ is the layer-specific learnable weight matrix, σ is a nonlinear activation function, and $Z \in \mathbb{R}^{n \times d_*}$ is the output features matrix.

To extract multi-layer graph features, we stack multiple graph convolution layers (3) as follows:

$$Z^{l+1} = \sigma\left(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}Z^l W^l\right), \tag{4}$$

where $Z^0 = X$, $Z^l \in \mathbb{R}^{n \times d_*^l}$ is the output of the l^{th} graph convolution layer, d_*^l is the number of output channels of layer l, and $W^l \in \mathbb{R}^{d_*^l \times d_*^{l+1}}$ maps d_*^l channels to d_*^{l+1} channels. Since the graph convolution operation can be stacked into multiple layers, we add a layer to horizontally concatenate the output Z^l, $l = 1, \cdots, L$. The final result is output as $Z^{1:L} := \left[Z^1, \cdots, Z^L\right]$, where the layer number L of our graph convolutional network is set to 2. Finally, we adopt the last graph convolution layer's output $Z_L \in \mathbb{R}^{n \times d_*}$ as node representations.

Inner Product Decoder. The inner product decoder is used to reconstruct the adjacency matrix and dynamically discover valuable potential link weights in visual semantic units and phrases. Considering that our latent embedding already contains content and structure information and flexibility, we choose to adopt a simple inner product decoder [15] to predict the link weights between visual semantic units and phrases by reconstructing the adjacency matrix, the reconstructed adjacency matrix $\hat{A}$ can be presented as follows:

$$\hat{A} = sigmoid\left(Z_L Z_L^T\right) \odot M', \tag{5}$$

where $\hat{A}$ is the reconstructed adjacency matrix of the graph, $M' \in \{0, 1\}^{n \times n}$ is the mask matrix for constructing bipartite graph, and $\odot$ is the element-wise product.

Reconstruction Loss. To effectively represent the enhanced video and text features Z_L, we need to ensure that the reconstructed adjacency matrix $\hat{A}$ is consistent with the initialized adjacency matrix A. Therefore, we adopt the cross-entropy loss [15] to measure the autoencoder approximation error. This approach minimizes the reconstruction loss by measuring the difference between A and $\hat{A}$, formulated as

$$\mathcal{L}_r = -\frac{1}{n^2} \sum_{(i,j) \in V \times V)} [A_{ij} \log \hat{A}_{ij} + (1 - A_{ij}) \log(1 - \hat{A}_{ij})]. \tag{6}$$

To sum up, Algorithm 1 summarize the whole process of GAE. Finally, we adopt the updated adjacency matrix $\hat{A}$ to obtain enhanced video and text feature representations $\hat{Z} \in \mathbb{R}^{n \times d_*}$.

3.4 Joint Embedding Learning

The purpose of joint embedding learning between video and text features is to perform similarity comparisons. As mentioned in the previous section, the cross-modal alignment method yields the aligned video and text feature representations $\hat{Z}$. For ultimate video and text embedding representations, we use self-attention pooling ρ consisting of two fully connected layers, a tanh activation function, and a softmax activation function on $\hat{Z}$ to obtain the attended weight video and text features. The generated video features $O \in \mathbb{R}^{K \times d_*}$ and text features $T \in \mathbb{R}^{J \times d_*}$ are denoted as

$$O = \sum^{K} \rho\left(\{\hat{Z}^i\}_{i=1}^{K}\right) \odot \{\hat{Z}^i\}_{i=1}^{K}, \tag{7}$$

$$T = \sum^{J} \rho\left(\{\hat{Z}^j\}_{j=K+1}^{K+J}\right) \odot \{\hat{Z}^j\}_{j=K+1}^{K+J}, \tag{8}$$

where K is the number of visual semantic units, J is the number of phrases, $\odot$ is the element-wise product. And then, a triplet ranking loss is employed to optimize the performance of the joint embedding learning.

Triplet Ranking Loss. Similar to other cross-modal retrieval methods [6, 30], we opt for the triplet learning method to optimize model parameters. Since we target for both video-to-text and text-to-video retrieval, the input of the loss function is composed of two triplets, namely (T, O, O_-) and (O, T, T_-). The first element of the triplet is either a video (O) or text (T) query, followed by a true positive and a negative example of a different modality as the second and third elements. The loss function is defined as:

$$\begin{aligned}\mathcal{L}_{rank} = &\max(0, \delta - cos(T, O) + cos(T, O_-)) \\ &+ \max(0, \delta - cos(O, T) + cos(O, T_-)),\end{aligned} \tag{9}$$

where $cos(\cdot, \cdot)$ is the cosine similarity and $0 < \delta \leq 1$ is the margin.

Overall Loss. The overall loss function that is used to train FCA-Net is the summation of the triplet ranking loss ($\mathcal{L}_{rank}$) and reconstruction loss ($\mathcal{L}_r$) :

$$\mathcal{L}_{full} = \mathcal{L}_{rank} + \lambda \mathcal{L}_r, \tag{10}$$

where λ serves as the trade-off parameter.

Algorithm 1 Algorithm to optimize GAE

Require: The node feature matrix X, and number of iterations e.
  for $i = 1, 2, \cdots, e$ **do**
    Compute A via Eq.(2).
    Compute Z_L via Eq.(3-4).
    Compute $\hat{A}$ via Eq.(5).
    repeat
      Update GAE with Eq.(6) by gradient descent.
    until convergence or exceeding maximum iterations.
    Get the new node representations Z_L and reconstructed adjacency matrix $\hat{A}$.
  end for
  Perform graph convolutional on $\hat{A}$.
Ensure: The node representations $\hat{Z}$.

4 EXPERIMENT

4.1 Dataset and Evaluation Metrics

MSR-VTT. The dataset is constructed by [38]. It contains 10,000 unique video clips with 20 descriptive captions per clip. Following the settings in [29, 38, 40], we adopt three kinds of evaluation settings: 1) **Full test set [38]**, testing on the full 2,990 test clips; 2) **1k-Miech test set [29]**, testing on the 1,000 test clips; and 3) **1k-Yu test set [40]**, testing on the 1,000 test clips.

YouCook2. The dataset is constructed by [43]. It contains 2,000 cooking instructional videos from 89 cooking recipes; including

14,000 video clips with a natural language sentence per clip. Following the splits in HowTo100M [30], there are 3,350 clips with their corresponding annotated sentences for testing.

VATEX. The dataset is recently constructed by [36]. It is a multilingual video-text dataset, including 25,991 videos for training, 3,000 for validation, and 6,000 for testing. There are 10 sentences in English and Chinese languages to describe each video, where the captions are longer and the visual contents are richer. We only adopt English annotations in experiments. Following the splits in HGR [7], there are 1,500 clips for validation, 1,500 clips for testing.

Evaluation metrics. We employ the widely used median retrieval rank (MedR) and recall rate at top $\mathcal{K}$ (R@$\mathcal{K}$) for assessing retrieval accuracy. MedR measures the median rank position among where true positives are returned. R@$\mathcal{K}$ measures the fraction of true positives being ranked at top $\mathcal{K}$ returned results. Therefore, lower MedR scores and higher R@$\mathcal{K}$ scores indicate better performance.

4.2 Experimental Settings

Implementation details. For MSR-VTT and YouCook2 datasets, we sample each video at 10 fps, while for VATEX we upsampled the features from 10 fps to 25 fps. For all datasets, we extract I3D [4] RGB features for the video frames and use these features as input to our model. For each semantic phrase in a sentence, we use pre-trained BERT [8] to extract 768-D word embedding and use gated embedding module [29] to obtain 768-D each semantic phrase representation. Ultimately, the dimensionality of the join embedding space is fixed to be in 768-D.

Training Details. Our model is implemented on the PyTorch[1] platform and trained with a Nvidia 2080Ti GPU machine. In all experiments, we use Adam [14] to optimize models with a minibatch of 64 text-video pairs and a learning rate of 0.0001. In the summarization module, the maximum number of video feature clusters K is set to 4, 5, and 4 for MSR-VTT, YouCook2, and VATEX datasets, respectively. The margin δ in Eq. (9) is set to 0.2 and the trade-off parameter λ in Eq.(10) is set as 0.1.

4.3 Performance Comparison

To demonstrate the effectiveness of our proposed FCA-Net model, we compare it with recently proposed state-of-the-art methods: (1) RNN-based methods: VSE++ [11], DualEncoding [10], TCE [39], (2) Transformer-based methods: ActBERT [44], MMT [12], COOT [13], (3) Multimodal Fusion methods: Mithun et al. [31], MEE [29], JPoSE [37], CE [21], and (4) other methods: JSFusion [40], HGLMM FV CCA [17], Miech et al. [30], HGR [7], SUPPORT-SET [32]. Note that for fair comparisons, we directly cited the results from their original papers without pre-training on HowTo100M [30].

Table 1 summarizes the performance of FCA-Net for the MSR-VTT dataset. FCA-Net consistently yields the best performance in all three dataset splits and thus demonstrates high effectiveness with simple yet reasonable designs. In the MSR-VTT full test set [38], we can observe that HGR and our method performs much better than Mithun et al. [31] in most cases. This result suggests that performing intermodality fine-grained interaction and alignment could lead to much better results than the methods that use common latent spaces only. Moreover, our method performs better than HGR [7] and CE [21]. This result proves the advantages of introducing fine-grained cross-modal alignment for text-video retrieval. For the second split [29] and third split [40] of the MSR-VTT dataset, we observe that the proposed FCA-Net still substantially outperforms prior models. SUPPORT-SET [32] is a state-of-the-art method

[1] http://www.pytorch.org

Table 1: Text-to-video retrieval comparison with state-of-the-art methods on MSR-VTT dataset (Section 4.3). It is worth noting that on the 1k-Yu test set [40], the symbol asterisk (*) indicates that it was trained on the corresponding training set of 9,000 videos and the others on the training set of 7,010 videos.

Method	R@1	R@5	R@10	MedR
Full test set [38]				
Mithun et al. [31]	7.0	20.9	29.7	38
HGR [7]	9.2	26.2	36.5	24
CE [21]	10.0	29.0	41.2	16
DualEncoding [10]	11.6	30.3	41.3	17
FCA-Net	**12.5**	**32.6**	**45.8**	**13**
1k-Miech test set [29]				
MEE [29]	16.8	41.0	54.4	9
JPoSE [37]	14.3	38.1	53.0	9
TCE [39]	17.1	39.9	53.7	9
MMT [12]	**20.3**	49.1	63.9	6
FCA-Net	19.7	**49.6**	**66.4**	**5**
1k-Yu test set [40]				
JSFusion [40]	10.2	31.2	43.2	13
Miech et al. [30]	12.1	35.0	48.0	–
MMT* [12]	24.6	54.0	67.1	4
SUPPORT-SET* [32]	**27.4**	**56.3**	67.7	3
FCA-Net	23.2	55.6	**70.3**	**3**

Table 2: Text-to-video retrieval comparison with state-of-the-art methods on YouCook2 dataset (Section 4.3).

Method	R@1	R@5	R@10	MedR
HGLMM FV CCA [17]	4.6	14.3	21.6	75
Miech et al. [30]	4.2	13.7	21.5	65
COOT [13]	5.9	16.7	24.8	49.7
ActBERT [44]	9.6	26.7	38.0	19
FCA-Net	**12.2**	**31.5**	**42.6**	**14**

Table 3: Text-to-video retrieval comparison with state-of-the-art methods on VATEX dataset (Section 4.3).

Method	R@1	R@5	R@10	MedR
VSE++ [11]	33.7	70.1	81.0	2
HGR [7]	35.1	73.5	83.5	2
DualEncoding [10]	36.8	73.6	83.7	–
SUPPORT-SET [32]	**44.6**	81.8	89.5	1
FCA-Net	41.3	**82.1**	**91.2**	**1**

that leverages a generative objective to improve the instance discrimination limitations of contrastive learning. This method shows much a better performance than others on the MSR-VTT 1k-Yu test set [40]. SUPPORT-SET [32] and our method consider learning concept sharing in the common video-text space. In our work, concept sharing refers to fine-grained related video semantic units and phrases. Therefore, it points to the positive effect of exploiting the fine-grained semantic alignment of our approach. Notably, although those multimodal fusion methods [12, 21, 31] achieve performance improvements by fusing other modalities, such as motion and audio features in video embedding, they still follow the trend of comprehensively capturing the semantic meanings of different modalities and ignore the exploration of semantic alignment between valuable components in videos and sentences.

Table 2 shows the experimental results with the latest contributions in the YouCook2 dataset. As shown in the last column, FCA-Net consistently outperforms all baseline methods w.r.t all evaluation metrics. The improvement can be as high as 4.6% relative to the recent ActBERT [44] w.r.t R@10. The reason is that FCA-Net performs finer-grained text-video interactions for in-depth relationship modeling between visual semantic units and phrases. In contrast to our work, these methods [13, 17, 30] encode different modalities by using their corresponding backbone architectures and learn a joint embedding space. However, these methods neglect to perform fine-grained interactions between different modalities.

Analogously, we compare our approach to recent works on the VATEX dataset (Table 3). FCA-Net also achieves the best retrieval performance with significant margins from baselines. Our method achieves 8.6% and 1.7% improvement relative HGR [7] and SUPPORT-SET [32] in terms of R@5 and R@10, respectively. The results verify that FCA-Net is effective in modeling fine-grained interactions between visual semantic units and phrases and thus leads to satisfactory text-video retrieval performance.

4.4 Study of FCA-Net

We experiment with different model variants to verify the effectiveness of diverse components and alignment based on GAE. The following is a brief introduction to these different model variant methods:

- **FCA-MeanP and FCA-GRU and FCA-SelfP.** Instead of using GAE for semantic alignment, we apply mean pooling, GRU, and self-attention pooling to many-to-many embedding to represent video and text features as cross-modal joint representations.
- **FCA-CoA.** We eliminate the GAE module. That is, we utilize co-attention to align visual semantic units and phrases.

We explore these model variants on the MSR-VTT full test set [38]. The experiment results of the component-wise comparison are displayed in Table 4. We omit the results on YouCook2 and VATEX because of space limitations, but they show similar trends to MSR-VTT. We have three main observations: 1) Relative to FCA-meanP, FCA-GRU helps to slightly improve performance because of the intramodality context modeling based on phrases and visual semantic units. FCA-SelfP slightly outperforms some baselines, including FCA-MeanP and FCA-GRU, which exploit self-attention pooling operations to automatically capture valuable contextual information. 2) FCA-CoA significantly outperforms FCA-meanP, FCA-GRU, and FCA-SelfP by a large margin, thus verifying the effectiveness of inferring fine-grained latent alignments. 3) By comparing FCA-Net and FCA-CoA models, we show that FCA-Net can further improve the performance significantly. This indicates the

Table 4: Text-to-video retrieval comparison with different variants on MSR-VTT dataset using the full test set [38] (Section 4.4).

Method	R@1	R@5	R@10	MedR
FCA-MeanP	4.7	16.2	23.5	63
FCA-GRU	4.8	17.6	25.3	52
FCA-SelfP	5.1	18.5	26.2	47
FCA-CoA	8.6	25.3	36.2	25
FCA-Net	**12.5**	**32.6**	**45.8**	**13**

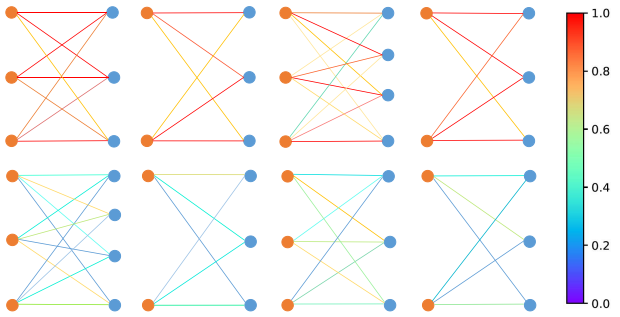

Figure 3: Testing the edge link weights of the adjacency matrix in the positive and negative text-video retrieval examples for YouCook2. Top 4 and bottom 4 are positive and negative text-video retrieval examples with the higher and lower link weights, respectively. In each subgraph, red nodes in the left are visual semantic units, blue nodes in the right are phrases. We visualize the edge link weights using the color map shown in the right. Higher weights are redder.

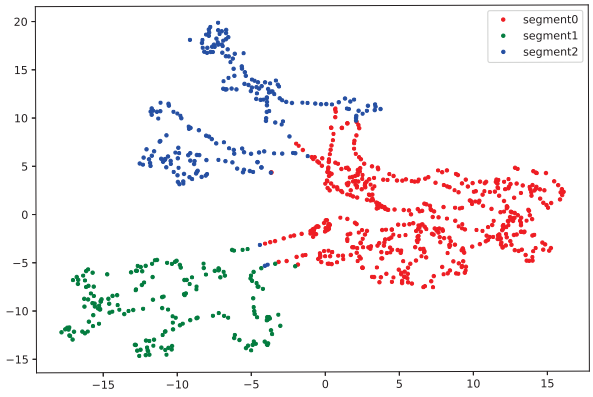

Figure 4: Framewise video features clustering visualization using t-SNE.

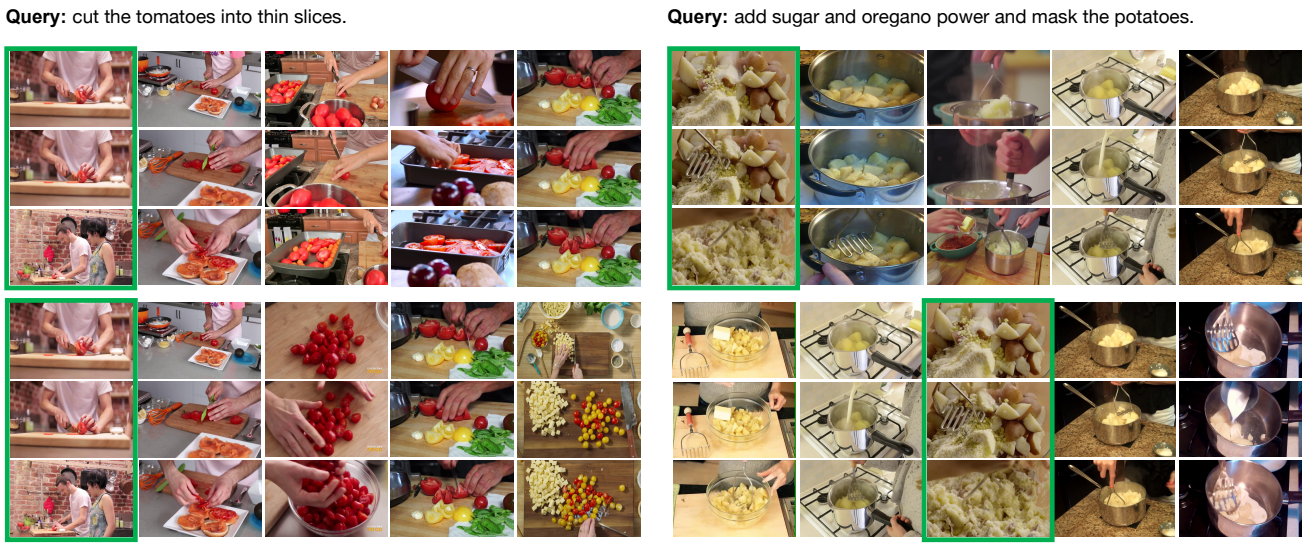

Figure 5: Qualitative examples of the text-video tasks: In (a), (b), we show retrieval ranks of FCA-Net (top) and FCA-SelfP (bottom) on YouCook2 dataset. Given a textual description as a query, we retrieve the most relevant video ranked from left to right. True positives are bounded in green boxes.

effectiveness of GAE in aligning key visual semantic units and phrases by predicting valuable link weights.

4.5 Visualization and Qualitative Analysis

Visualization. To demonstrate the effectiveness of the GAE-based link prediction strategy, we visualize 4 testing the edge link weights of the adjacency matrix in the positive and negative text video retrieval instances for YouCook2 (Figure 3). We observe substantially different patterns between the high-weight and low-weight subgraphs. This result explains why FCA-Net can predict the edge link weights merely from these subgraphs. Furthermore, this result validates the capability of FCA-Net to predict high link weights for positive examples and thereby achieve feature enhancement between videos and texts to further improve retrieval accuracy. As shown in Figure 4, we use t-SNE to visualize the clustering results of framewise video features in the example shown in Figure 5 (b). The proposed method can clearly cluster video embeddings according to different action classes.

Qualitative Analysis. We show the first example in Figure 5 (*a*). The simple sentence describes the action "cut the tomatoes into thin slices" in a short-term segment. We observe that both models are able to rank the true positive in the top 1 position because the actions of the video clip described in the sentence are visually easy to distinguish and match. In another example from Figure 5 (*b*), we select a complex sentence that involves three actions (i.e., "add sugar," "add oregano powder," and "mask the potatoes"). By comparing FCA-Net with its variant FCA-SelfP, we find that our model successfully retrieves the correct video containing all actions and entities described in the sentence. The other videos lacks the "add sugar" and "add oregano power" actions. In the bottom example, the FCA-SelfP model also retrieves similar scenes in the video. However, we observe that the videos involving related elements and few actions are only ranked as true positive in the top 3 position. The performance of FCA-SelfP indicates that removing the fine-grained cross-modal alignment hurts the expressiveness of the video and text representation and further degrades retrieval performance.

5 CONCLUSION

This work contributes to a new cross-modal learning method to model videos and texts jointly, which leverages the intrinsic semantic cues of both videos or texts. Specifically, we consider the correspondences of visual semantic units and phrases, and a link prediction strategy based on GAE is proposed to align visual semantic units and phrases. Evaluation results on public datasets demonstrate that our model exhibits highly competitive performance compared to state-of-the-art baselines.

While encouraging, the current work still has limitations in retrieval efficiency by considering fine-grained interactions between texts and videos, making it difficult to apply to real-world scenarios. Hence, how to improve retrieval efficiency is our future direction.

6 ACKNOWLEDGEMENTS

The work is supported by the National Key R&D Program of China (No. 2018YFB1402600), the National Natural Science Foundation of China (No. 62072116, No. 61772190), the Shanghai Pujiang Program (No. 20PJ1401900), and the Natural Science Foundation of Hunan Province (No. 2020JJ4219).

REFERENCES

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *ICML*. 1247–1255.

[2] Rianne van den Berg, Thomas N Kipf, and Max Welling. 2017. Graph convolutional matrix completion. *arXiv preprint arXiv:1706.02263* (2017).

[3] Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* 3, 1 (1974), 1–27.

[4] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*. 6299–6308.

[5] Jingjing Chen and Chong-Wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *ACM MM*. 32–41.

[6] Jing-Jing Chen, Chong-Wah Ngo, Fu-Li Feng, and Tat-Seng Chua. 2018. Deep understanding of cooking procedure for cross-modal recipe retrieval. In *ACM MM*. 1020–1028.

[7] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-grained Video-Text Retrieval with Hierarchical Graph Reasoning. In *CVPR*. 10638–10647.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*. 4171–4186.

[9] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. 2019. Dual encoding for zero-example video retrieval. In *CVPR*. 9346–9355.

[10] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2021. Dual Encoding for Video Retrieval by Text. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[11] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* (2017).

[12] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *ECCV*. 214–229.

[13] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. 2020. Coot: Cooperative hierarchical transformer for video-text representation learning. In *NeurIPS*.

[14] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

[15] Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* (2016).

[16] Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

[17] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2015. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*. 4437–4446.

[18] Anna Kukleva, Hilde Kuehne, Fadime Sener, and Jurgen Gall. 2019. Unsupervised learning of action classes with continuous temporal embedding. In *CVPR*. 12066–12074.

[19] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *ECCV*. 201–216.

[20] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *ACM SIGIR*. 15–24.

[21] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487* (2019).

[22] Zhenguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, and Xun Wang. 2021. Deep Dual Consecutive Network for Human Pose Estimation. In *CVPR*. 525–534.

[23] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. 2019. Towards natural and accurate future motion prediction of humans and animals. In *CVPR*. 10004–10012.

[24] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. 2020. Univilm: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353* (2020).

[25] James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1. 281–297.

[26] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL*. 55–60.

[27] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2020. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *arXiv preprint arXiv:2008.05231* (2020).

[28] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-End Learning of Visual Representations From Uncurated Instructional Videos. In *CVPR*. 9876–9886.

[29] Antoine Miech, Ivan Laptev, and Josef Sivic. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *CoRR* abs/1804.02516 (2018). http://arxiv.org/abs/1804.02516

[30] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*. 2630–2640.

[31] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. 2018. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *ICMR*. 19–27.

[32] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F. Henriques, and Andrea Vedaldi. 2021. Support-set bottlenecks for video-text representation learning. In *ICLR*.

[33] Yuxin Peng, Jinwei Qi, and Yunkan Zhuo. 2019. MAVA: Multi-Level Adaptive Visual-Textual Alignment by Cross-Media Bi-Attention Mechanism. *IEEE Transactions on Image Processing* 29 (2019), 2728–2741.

[34] Yale Song and Mohammad Soleymani. 2019. Polysemous visual-semantic embedding for cross-modal retrieval. In *CVPR*. 1979–1988.

[35] Xiaolong Wang and Abhinav Gupta. 2018. Videos as space-time region graphs. In *ECCV*. 399–417.

[36] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In *ICCV*. 4581–4591.

[37] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. 2019. Fine-Grained Action Retrieval Through Multiple Parts-of-Speech Embeddings. In *ICCV*. 450–459.

[38] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*. 5288–5296.

[39] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. 2020. Tree-Augmented Cross-Modal Encoding for Complex-Query Video Retrieval. In *ACM SIGIR*. 1339–1348.

[40] Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. A Joint Sequence Fusion Model for Video Question Answering and Retrieval. In *ECCV*. 487–503.

[41] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017. End-to-end concept word detection for video captioning, retrieval, and question answering. In *CVPR*. 3165–3173.

[42] Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*. 9159–9166.

[43] Luowei Zhou, Nathan Louis, and Jason J Corso. 2018. Weakly-Supervised Video Object Grounding from Text by Loss Weighting and Object Interaction. In *BMVC*. 50.

[44] Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *CVPR*. 8746–8755.

[45] Yuan Zhuang, Zhenguang Liu, Peng Qian, Qi Liu, Xiang Wang, and Qinming He. 2020. Smart Contract Vulnerability Detection using Graph Neural Network. In *IJCAI*. 3283–3290.