Audio-Visual Mismatch-Aware Video Retrieval via Association and Adjustment

Sangmin Lee, Sungjune Park, and Yong Man Ro

Image and Video Systems Lab, KAIST, South Korea
{sangmin.lee, sungjune-p, ymro}@kaist.ac.kr

Abstract. Retrieving desired videos using natural language queries has attracted increasing attention in research and industry fields as a huge number of videos appear on the internet. Some existing methods attempted to address this video retrieval problem by exploiting multimodal information, especially audio-visual data of videos. However, many videos often have mismatched visual and audio cues for several reasons including background music, noise, and even missing sound. Therefore, the naive fusion of such mismatched visual and audio cues can negatively affect the semantic embedding of video scenes. Mismatch condition can be categorized into two cases: (*i*) Audio itself does not exist, (*ii*) Audio exists but does not match with visual. To deal with (*i*), we introduce audio-visual associative memory (AVA-Memory) to associate audio cues even from videos without audio data. The associated audio cues can guide the video embedding feature to be aware of audio information even in the missing audio condition. To address (*ii*), we propose audio embedding adjustment by considering the degree of matching between visual and audio data. In this procedure, constructed AVA-Memory enables to figure out how well the visual and audio in the video are matched and to adjust the weighting between actual audio and associated audio. Experimental results show that the proposed method outperforms other state-of-the-art video retrieval methods. Further, we validate the effectiveness of the proposed network designs with ablation studies and analyses.

Keywords: Video retrieval, audio-visual mismatch, audio association, embedding adjustment, memory

1 Introduction

Video retrieval is to find corresponding videos from natural language queries made by humans. Given the huge number of videos on the internet, it is highly time-consuming and labor-intensive for people to find desired video scenes manually. Thus, automatic video retrieval has attracted increasing attention in research and industry fields due to its high practicality.

Video retrieval methods utilizing deep neural networks (DNNs) have been proposed to address arising issues in video retrieval. Some works focused on hierarchical feature matching [46, 61] for video retrieval. These methods tried to perform video-text matching in both local (*e.g.*, word-scene) and global (*e.g.*,

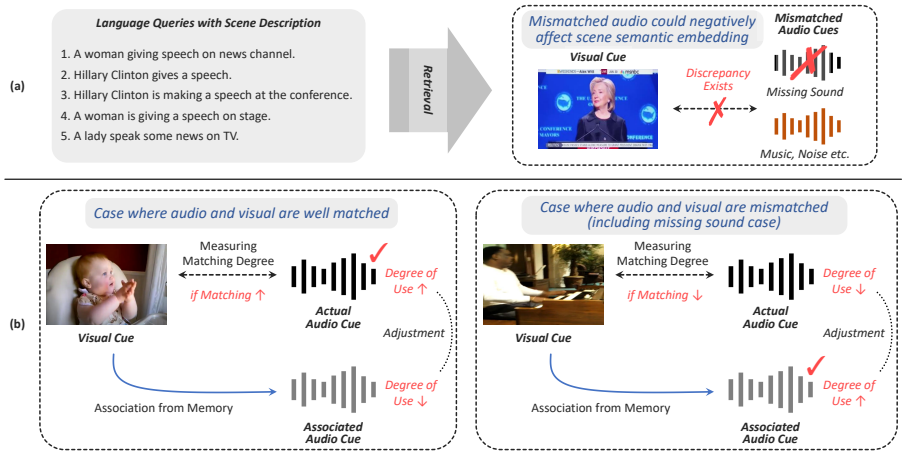

Fig. 1: (a) shows the examples of mismatching between visual and audio cues. (b) describes the concept of the proposed method for video embedding. To cope with the mismatch condition, the model adjusts the proportion of using actual audio and associated audio considering the degree of audio-visual matching.

sentence-video) levels. There exist metric learning-based video retrieval methods [45, 52, 53]. They formulated training metric criterion by considering similarities and relationships among samples. Several works addressed architectural aspects [4, 10]. They investigated effective architectures such as multi-level embedding and pooling strategies. These works mainly focused on visual and text matching.

However, natural language queries vary greatly and often contain details related to audio cues. For example, language queries 'A woman giving speech on news channel' and 'A woman is singing while men are playing on guitars' include audio information and it is worth to give guiding the model with audio information. Some existing works tried to address such video retrieval problem using multi-modal information, especially audio-visual data of videos [17, 36, 50, 54]. They showed that audio cues can contribute to the performance of video retrieval. Nonetheless, these works did not take into account the mismatch condition of visual and audio. In many videos, visual and audio cues are often not matched due to several reasons such as background music, noise, and even missing sound. Therefore, the naive fusion of visual and audio can negatively affect matching video with text queries. When fusing the visual and audio in the video, mismatched audio can guide the video embedding to have distracted semantics.

Our work addresses audio-visual mismatch issues for retrieving video from text, which have not been properly dealt with in previous video retrieval works. The audio-visual mismatch in video can be categorized into two cases: (*i*) Audio itself does not exist, (*ii*) Audio exists but does not match with visual (*e.g.*, background music, noise). Figure 1-(a) shows that such mismatched audio cues do not help to obtain appropriate embeddings for matching language semantics.

In this paper, we introduce a novel mismatch-aware associative transformer (MA-Transformer) with association and adjustment processes to deal with the aforementioned issues. To address the issue (*i*), we propose audio-visual associative memory (AVA-Memory) in MA-Transformer to associate audio cues even from videos without actual audio data. AVA-Memory is composed of visual and audio memories to store information of two modalities and enables to associate different modal cues from the other modal input. The associated audio cues from visual data can guide the video embedding feature to be aware of audio information for video-text matching even in the missing audio condition.

In addition, to deal with the issue (*ii*), we propose audio embedding adjustment by considering the degree of matching between visual and audio data (See Figure 1-(b)). Through AVA-Memory, the degree of audio-visual matching is determined by performing mutual associations from audio to visual and from visual to audio. It is possible to figure out how well the visual and audio of the video are matched and to adjust the weighting between actual audio and associated audio. If the visual and audio are matched well, the actual audio cue is mainly used for video embedding. Conversely, if the visual and audio are mismatched, the model adaptively lowers the use of the actual audio and performs video embedding mainly by using the associated audio. In the case of missing audio, only associated audio can be used for embedding.

The major contributions of the paper are as follows.

- We introduce a novel MA-Transformer with AVA-Memory for video retrieval. We can associate the audio cues from visual cues in the video. It enables to guide the visual embedding to be aware of audio context jointly for video-text matching even in the missing audio condition.
- We propose audio embedding adjustment which enables to address the mismatch between existing visual and audio cues. We can figure out the degree of audio-visual matching through the constructed AVA-Memory and adjust the use of audio cues based on it. To the best of our knowledge, it is the first attempt to deal with audio-visual mismatch issues in video retrieval.

2 Related Works

2.1 Video Retrieval

Video retrieval is to find corresponding videos from natural language queries made by humans. Based on the high practicality, video retrieval attracts increasing attention. Compared with image retrieval [13, 24, 28], video retrieval is more challenging because it necessitates thorough comprehension of temporal dynamics as well as complicated text semantics.

Early video retrieval works extended the image retrieval approach by spatio-temporal aggregation of frames for each video [8, 43, 47, 57]. There have been methods which develop feature aggregation for video retrieval. For example, average pooling [36, 42] and max pooling [41, 54] were used as aggregation methods for embedding feature in retrieval. Chen *et al.* [4] proposed a generalized

pooling operator to automatically use the best pooling strategy for each feature for retrieval. Recently, Dong *et al.* [10] proposed a dual-encoding network with multiple levels of feature capabilities for video retrieval. In [10], multi-level features are from average pooling, bi-directional GRU, and convolutional layer. Some works focused on hierarchical feature matching in terms of global and local matching. [60] estimates similarity through comparison between each word of text description and video frame. Zhang *et al.* [61] performs a paragraph-based video search using hierarchical decomposition of videos and paragraphs. Song *et al.* [46] proposed varied representations by combining global context with local features to consider polysemous videos. There were attempts to effectively pretrain the model with uncurated instruction [39] or image-level caption data [1].

These approaches do not take advantage of the further diverse information related to videos, including voice and other background sounds, which in practice can affect human-made natural language descriptions. [40,42] proposed a method using a pretrained model for audio and motion recognition. [36] investigated additional cues such as faces, OCR, speech. They proposed a more effective use of multi-modal features through a collaborative gating method. In addition to this, [12,16,17,35,38,50] introduced a video retrieval method by fusing the multi-modalities with transformer [48] structures. [50] proposed video-text matching method in terms of global and local alignment with multi-modal transformer.

However, the previous works did not take into account the visual and audio mismatch conditions. Many videos often have mismatched visual and audio cues, which can negatively affect matching video with text queries. When fusing the visual and audio in the video, mismatched audio can guide the video embedding to have distracted semantics. We introduce a novel MA-Transformer with AVA-Memory for addressing such mismatch issues.

2.2 Memory-Augmented Network

A memory-augmented network represents the neural network that includes external memory components for reading and writing historical information. Memory-augmented networks have been proposed to handle a variety of challenges in the deep learning field. There were several tasks to exploit the memory such as anomaly detection [18,44], few-shot learning [2,23,62], object tracking/detection [15, 25, 58], future prediction [29, 37], and representation learning [19, 26, 30]. There exist methods that exploit the memory-augmented network for cross-modal retrieval which is image and sentence matching [22]. It utilizes memorization of shared semantic representations to address the few-shot condition.

Unlike the existing memory-augmented networks, we introduce a novel AVA-Memory for learning audio-visual correspondences with audio and visual sub-memories. Through proposed audio-visual associative learning with the memory, it is possible to recall audio cues from the videos without audio and further to measure how well visual and audio are matched. The degree of matching from AVA-Memory is utilized to adjust audio embeddings afterward.

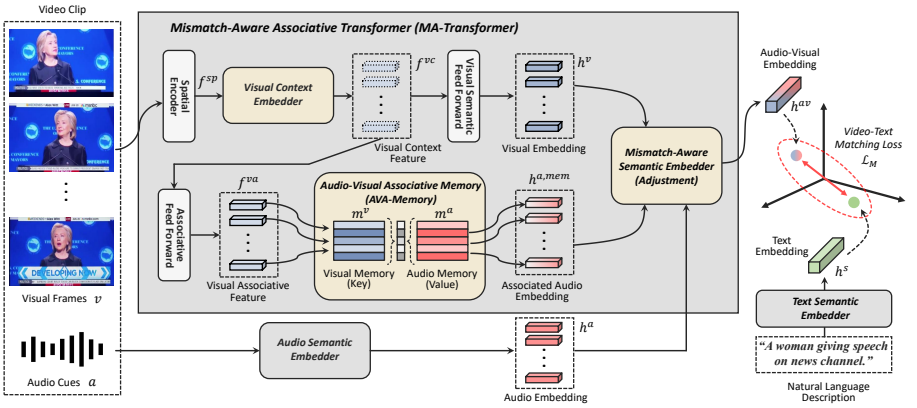

Fig. 2: Overall framework containing a proposed MA-Transformer for video retrieval. MA-Transformer mainly consists of 3 parts: a visual context embedder, an audio-visual associative memory, and a mismatch-aware semantic embedder. Each of them is for extracting spatio-temporal context of visual data, associating audio cues from visual features, and encoding video semantics jointly with visual and audio with being aware of mismatch condition, respectively.

3 Proposed Approach

Video retrieval task can be formulated as follows. Let $v = \{v_t\}_{t=1}^{n}$ denote the n frames in video clip while $a = \{a_t\}_{t=1}^{n}$ indicates audio cues with n partial audio samples. a_t is VGGish [36] feature of 1s audio and v_t is middle frame in 1s. Let s denote text sentence. A video mapping function $\mathcal{F}_v$ and a text mapping function $\mathcal{F}_s$ are optimized to make video embedding $\mathcal{F}_v(v, a)$ and text embedding $\mathcal{F}_s(s)$ be matched. As a result, it is possible to retrieve videos from text by measuring similarity between the embeddings. The goal of this work is to make video embedding to be aware of the auditory context even in the case of audio-visual mismatch conditions (*e.g.*, missing sound, background music).

To this end, we introduce a mismatch-aware associative transformer (MA-Transformer) for video embedding. MA-Transformer consists of three major parts: a visual context embedder, an audio-visual associative memory (AVA-Memory), and a mismatch-aware semantic embedder. The visual contexts are processed with self-attention mechanism to focus on the parts suitable for associating audios and embedding the semantics of visual data. AVA-Memory includes visual and audio sub-memories and is able to associate audio cues from visual cues. Finally, in the mismatch-aware semantic embedder, the degree of use between actual audio and associated audio is adjusted in consideration of the degree of matching between visual and audio. Then, it encodes audio-visual semantic embeddings that effectively match text queries. In terms of training, we propose audio-visual associative learning which enables to learn the association between visual and audio modalities in a self-supervised way.

3.1 Mismatch-Aware Associative Transformer

Figure 2 shows the overall framework of the proposed MA-Transformer. First, each frame of the input video is independently fed to a spatial encoder (CNN) to extract spatial features $f^{sp}=\{f_t^{sp}\}_{t=1}^{n}\in\mathbb{R}^{n\times d_{sp}}$ (d_{sp} is channel dim).

Visual context embedder. The extracted spatial features are received by the visual context embedder to encode the relationship between features. The visual context embedder mainly has the form of transformer [11, 48] and can encode the overall spatio-temporal context of visual features through the self-attention mechanism. The overall process of the context embedder is similar to that of the transformer [48]. However, the last part of it consists of multi-head attention, not feed-forward. It can be formulated as follows.

$$\begin{aligned} e_0 &= [f_1^{sp}W_{vc}; f_2^{sp}W_{vc}; \ldots; f_n^{sp}W_{vc}] + E_{\mathrm{POS}}, \\ e'_l &= \mathrm{MHA}(\mathrm{LN}(e_{l-1})) + e_{l-1}, \qquad l = 1, \ldots, L_v \\ e_l &= \mathrm{MLP}(\mathrm{LN}(e'_l)) + e'_l, \qquad l = 1, \ldots, L_v \\ f^{vc} &= \mathrm{MHA}(\mathrm{LN}(e_{L_v})) + e_{L_v}, \end{aligned} \tag{1}$$

where $W_{vc}\in\mathbb{R}^{d_{sp}\times d_v}$ indicates a fc layer and E_{POS} is positional encoding. MHA and LN represent multi-head attention and layer normalization, respectively. L_v indicates the number of layers. As a result, we obtain visual context feature $f^{vc}=\{f_t^{vc}\}_{t=1}^{n}\in\mathbb{R}^{n\times d_v}$. We apply multi-head attention as the last layer to aggregate f_t^{vc} differently with separate feed forward layers later. Note that the feed forward layer indicates $y = \mathrm{MLP}(\mathrm{LN}(x)) + x$ operation. f^{vc} separately goes through two paths. One (upper path of f^{vc} in Figure 2) is for embedding semantic-related visual features. The other (lower path of f^{vc} in Figure 2) is for associating audio features from the visual context.

In the upper path, visual semantic feed forward is further applied to f^{vc} to aggregate for attending the semantic-related positions of the video sequence. Through the procedure, we obtain visual embedding $h^v=\{h_t^v\}_{t=1}^{n}\in\mathbb{R}^{n\times d_v}$ that contains semantic characteristics of visual cues. The visual embedding is directly utilized as input of the mismatch-aware semantic embedder.

In the lower path, another feed forward layer, associative feed forward is further applied to f^{vc} to aggregate the audio-related characteristics of the video sequence. It is possible because we construct the audio-visual association based on the output of this feed forward. We get visual associative feature $f^{va}=\{f_t^{va}\}_{t=1}^{n}\in\mathbb{R}^{n\times d_v}$ which is used to recall audio cues from AVA-Memory.

AVA-Memory. The extracted visual associative features are utilized as memory queries for accessing a visual memory m^v and an audio memory m^a in AVA-Memory. The visual and audio memories are constructed as $m^v = \{m_r^v\}_{r=1}^{k}\in\mathbb{R}^{k\times d_v}$ and $m^a = \{m_r^a\}_{r=1}^{k}\in\mathbb{R}^{k\times d_a}$, respectively with k slots and (d_v, d_a) channels. A vector $m_r^v\in\mathbb{R}^{d_v}$ indicates the r-th memory component of m^v. AVA-Memory has key-value memory structure to map one modal space to another. In this case, the visual memory is the key and the audio memory is the value. Addressing vectors $w_t^v = \{w_{t\_r}^v\}_{r=1}^{k}\in\mathbb{R}^k$ is individually obtained from each time component f_t^{va} of visual associative features f^{va}. Note that each addressing

vector is used to access the components of audio memory m^a. The addressing procedure with input f_t^{va} can be written as

$$w_{t\_r}^v = \frac{\exp(d(f_t^{va}, m_r^v)/\tau_m)}{\sum_{r=1}^{k} \exp(d(f_t^{va}, m_r^v)/\tau_m)}, \tag{2}$$

$$d(f_t^{va}, m_r^v) = \frac{f_t^{va} \cdot m_r^v}{||f_t^{va}||\,||m_r^v||}, \tag{3}$$

where $d(\cdot,\cdot)$ indicates cosine similarity, $\exp(\cdot)/\sum\exp(\cdot)$ denotes softmax, and τ_m is a memory temperature. Each component $w_{t\_r}^v$ of w_t^v can be considered as an attention weight for audio memory slot m_r^a at time-step t. m^a outputs an associated audio feature $f_t^{a,mem} \in \mathbb{R}^{d_a}$ as follows.

$$f_t^{a,mem} = \sum\nolimits_{r=1}^{k} w_{t\_r}^v m_r^a. \tag{4}$$

Repeating this for each time, we obtain associated audio features $f^{a,mem} = \{f_t^{a,mem}\}_{t=1}^n \in \mathbb{R}^{n\times d_a}$. It passes through a fc layer and we obtain associated audio embedding $h^{a,mem} = \{h_t^{a,mem}\}_{t=1}^n \in \mathbb{R}^{n\times d_a}$. The learning scheme of AVA-Memory is addressed in Section 3.2. In the meantime, we acquire the actual audio embedding $h^a = \{h_t^a\}_{t=1}^n \in \mathbb{R}^{n\times d_a}$ from audio data via audio semantic embedder which is typical transformer [48] (bottom path of Figure 2).

Mismatch-aware semantic embedder. The last part of MA-Transformer, mismatch-aware semantic embedder receives the visual embedding h^v, the actual audio embedding h^a, and the associated audio embedding $h^{a,mem}$. The positional encoding is applied separately to h^v, h^a, and $h^{a,mem}$. In addition, [CLS] token is applied for aggregating the feature afterward as [7]. The integrated input of the mismatch-aware semantic embedder is formulated as follows.

$$\begin{aligned} e_0^{av} =& [[CLS]; [h_1^v; \ldots; h_n^v]; [h_1^a; \ldots; h_n^a]; [h_1^{a,mem}; \ldots; h_n^{a,mem}]] \\ &+ [0; E_{\text{POS}}; E_{\text{POS}}; E_{\text{POS}}]. \end{aligned} \tag{5}$$

The mismatch-aware semantic embedder has a similar structure to the visual context embedder except for the last part (it ends with feed forward). To adjust the weighting between the associated audio and the actual audio, we define matching index α (0∼1) which indicates how well visual and audio are matched. If α is high, the degree of using actual audio h^a is increased, and vice versa, it is lowered. This matching index α is obtained by exploiting AVA-Memory and it is described in Section 3.2. We apply α according to each index c of e_0^{av}. In multi-head attention of the embedder, we adjust the attention with α as follows.

$$\beta_c = \begin{cases} 1 & \text{if } c \in \text{index of } [CLS] \textit{ and } h^v \\ \alpha & \text{if } c \in \text{index of } h^a \\ 1-\alpha & \text{if } c \in \text{index of } h^{a,mem} \end{cases} \tag{6}$$

$$\text{Attention}(Q, K_c) = \text{Softmax}(QK_c^T/\sqrt{dim} + \log\beta_c), \tag{7}$$

where c indicates the index of e_0^{av}. Q and K are query and key in multihead attention [48] while dim indicates their dimension. This attention is applied to

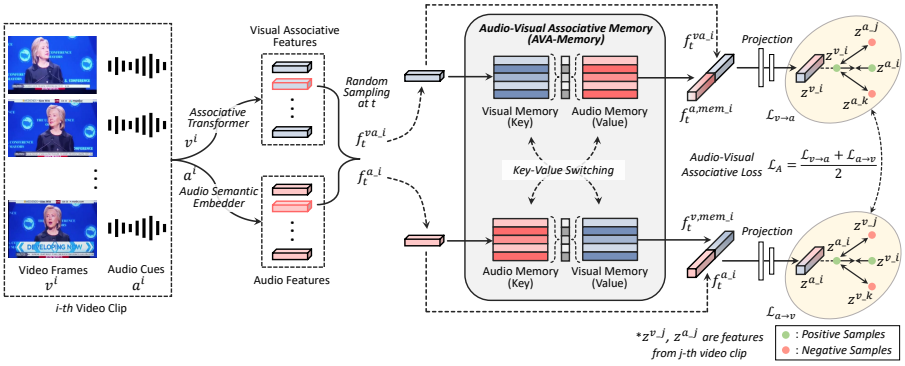

Fig. 3: Proposed audio-visual associative learning with AVA-Memory. Through the associative learning, it is possible to store audio-visual features in sub-memories and naturally associate with one another. As a result, we can obtain the corresponding audio cues from the visual features with AVA-Memory and further determine whether the visual and audio are well matched.

value V from e_0^{av} to focus on parts of h^v, h^a, and $h^{a,mem}$. By applying α in log form, each attention weight is adjusted proportionally considering exponential scale in softmax function. Based on this self-attention scheme, the embedder integrally attends to information considering the reliability of audio cues. If the audio cue is not reliable (not matched with visual), it decreases the proportion of actual audio h^a while it increases the proportion of associated audio $h^{a,mem}$. Finally, we get an audio-visual embedding h^{av} by aggregating features of the semantic embedder with [CLS] output. h^{av} is used to match text embedding h^s.

3.2 Associative Learning with AVA-Memory

AVA-Memory is trained with audio-visual data as shown in Figure 3. The goal of the associative learning is to store visual and audio features in their sub-memories and to build the link between sub-memories. As a result, we can associate different modality from the other modal input and further determine whether visual and audio are well matched or not.

With the video frames v^i and audio cues a^i from i-th video, we extract visual associative features f^{va_i} and audio features f^{a_i} from MA-Transformer and audio semantic embedder, respectively. Note that f^a does not pass through the last fc layer of the audio semantic embedder and thus is different from audio embedding h^a. We randomly sample paired features ($f_t^{va\_i}$, $f_t^{a\_i}$) at time t. We perform sampling a pair to construct associative learning (attract and repel) in a computationally efficient way. In addition, the combination of training data can be diversified with randomness. In terms of the visual feature $f_t^{va\_i}$, the visual memory is used as a key while the audio memory is used as a value. In contrast for the audio feature $f_t^{a\_i}$, those memories are switched at this time. Each feature changes its modality space through key-value memory addressing as described

in the previous section. We can obtain concatenated features $[f_t^{a,mem\_i}; f_t^{va\_i}]$ and $[f_t^{a\_i}; f_t^{v,mem\_i}]$. They pass through the projection head with a fc layer to make z^{v_i} and z^{a_i}, respectively. Actually, they should be $z_t^{v\_i}$ and $z_t^{a\_i}$. However, z represents the feature for each video, so it is denoted as z^{v_i} and z^{a_i} here. If z^{v_i} and z^{a_i} are from a pair, we consider them as a positive set (*e.g.*, z^{v_i}, z^{a_i}). Otherwise, we regard them as a negative set (*e.g.*, z^{v_i}, z^{a_j}). The memory can associate one to another modality by making such a positive set distinctly close. However, the training data also includes videos in which visual and audio are not matched. Therefore, we introduce a new audio-visual associative loss which is the transformed version of contrastive loss [6] to reduce the influence of these mismatched samples. This is inspired by the work [21] that uses $\log(1-p)$ loss rather than the typical $-\log(p)$ (cross-entropy) to reduce the effect of noisy class labels. In case of $-\log(p)$, the gradient on the hard samples ($p \downarrow$) is higher, whereas in the case of $\log(1-p)$, the gradient on the easy samples ($p \uparrow$) is higher. Therefore, when $\log(1-p)$ is used, learning can proceed in a direction that does not force optimization on difficult mismatched samples. As a result, our audio-visual associative loss can be formulated as follows.

$$\mathcal{L}_{v\to a} = \frac{1}{N}\sum_{i=1}^{N}\log(1 - \frac{\exp(d(z^{v\_i}, z^{a\_i})/\tau_l)}{\sum_{j=1}^{N}\exp(d(z^{v\_i}, z^{a\_j})/\tau_l)}), \tag{8}$$

$$\mathcal{L}_{a\to v} = \frac{1}{N}\sum_{i=1}^{N}\log(1 - \frac{\exp(d(z^{a\_i}, z^{v\_i})/\tau_l)}{\sum_{j=1}^{N}\exp(d(z^{a\_i}, z^{v\_j})/\tau_l)}), \tag{9}$$

$$\mathcal{L}_A = \frac{\mathcal{L}_{a\to v} + \mathcal{L}_{v\to a}}{2}, \tag{10}$$

where N and τ_l indicate a batch size and a loss temperature parameter, respectively. By minimizing $\mathcal{L}_A$, we can attract one another within the positive set and repel each other within the negative set to properly associate distinct audio cues from visual and vice versa. During the training phase, the weights of m^v and m^s are updated via backpropagation as [18, 29].

Furthermore, we can determine whether the visual and audio are aligned or not by utilizing AVA-Memory. When a specific i-th video (v^i, a^i) is given, we formulate the matching index α as follows.

$$\alpha = max(0, \frac{1}{n}\sum_{t=1}^{n} d(z_t^{v\_i}, z_t^{a\_i})), \tag{11}$$

where $d(\cdot,\cdot)$ indicates cosine similarity function. $z_t^{v\_i}$ and $z_t^{a\_i}$ indicates the visual and audio projections of i-th video at time t. As a result, α has a value between 0 and 1. α has a high value for a high degree of matching between visual and audio. In case of missing audio condition, matching index α is set as 0.

3.3 Video-Text Matching

To encode the text embedding h^s in the side of natural language query, we adopt common language model, BERT [7] as [17, 50]. The networks are trained

to match h^s with h^{av} in latent space (See Figure 2). Based on them, triplet ranking loss [13] is applied to formulate video-text matching loss $\mathcal{L}_M$ as follows.

$$\begin{aligned}\mathcal{L}_M = & max(0, \delta + d(h^{av}, h^{s-}) - d(h^{av}, h^s)) \\ & + max(0, \delta + d(h^{av-}, h^s) - d(h^{av}, h^s)),\end{aligned} \tag{12}$$

where $d(\cdot, \cdot)$ indicates cosine similarity and δ is a margin parameter. h^{av-} and h^{s-} indicate the features from negative samples which are not paired with one another. Model training is conducted with both associative loss and matching loss concurrently. The total objective loss is defined as $\mathcal{L} = \mathcal{L}_A + \mathcal{L}_M$ for training.

At inference time, we can obtain the similarity between a video v and a text s with latent space similarity as follows.

$$sim(v, s) = d(h^{av}, h^s). \tag{13}$$

As a result, the videos with high similarities can be retrieved from text queries.

4 Experiments

4.1 Datasets

MSR-VTT. MSR-VTT dataset [56] consists of 10k web videos and corresponding 200k text descriptions. Each video has 20 text captions with diverse descriptions. About 10% of the videos in MSR-VTT do not contain audio data. The *MSR-VTT-Original* partition [56] of MSR-VTT employs 6,513 clips for training, 497 clips for validation, and 2,990 clips for testing. In another partition *MSR-VTT-Miech* [40], 6,656 and 1,000 clips are used for training and testing, respectively. We evaluated these partitions for comprehensive validation.
VATEX. VATEX [51] consists of YouTube videos with multilingual text descriptions. Chinese and English descriptions exist and we adopt English descriptions only. As following the data partition [5], we perform the retrieval experiments with 25,991 training videos, 1,500 validation videos, and 1,500 testing videos.
TGIF. TGIF [33] is a GIF format dataset which does not have audio data. It contains 100,000 GIF videos and corresponding 120,000 text descriptions about the GIFs' content. According to the data split [32], we perform the experiments with 78,799 training videos, 10,705 validation videos, and 11,351 testing videos.

4.2 Implementation

For MSR-VTT and TGIF, we adopt the spatial encoder as pretrained ResNeXt-101 [55] and ResNet-152 [20]. We concatenate the two features to make a 4,096-d feature as [10]. In terms of VATEX dataset, we use the I3D [3] features with 1,024-d offered by the dataset constructor [51]. Audio data is firstly processed with pretrained VGGish as [36]. The proposed model is trained by Adam optimizer [27] with an initial learning rate of 0.00005 and a batch size of 128. Memory slot size s is fixed as 500 for all experiments. Memory and loss temperature parameters (τ_m, τ_l) are both set as 0.1 for all experiments according to temperature setup [6]. The margin parameter δ is set as 0.2. The overall detailed network structures are described in the supplementary material.

Table 1: Video retrieval performance results on *MSR-VTT-Original* dataset.

Method	Video Retrieval Performance				
	R@1↑(%)	**R@5↑(%)**	**R@10↑(%)**	**MedR↓**	**mAP↑(%)**
W2VV [8]	1.1	4.7	8.1	236	3.7
Francis [14]	6.5	19.3	28.0	42	-
VSE++ [13]	8.7	24.3	34.1	28	16.9
W2VV++ [31]	11.1	29.6	40.5	18	20.6
TCE [59]	7.7	22.5	32.1	30	-
HGR [5]	9.2	26.2	36.5	24	-
UWML [53]	10.9	30.4	42.3	-	-
HSL [10]	11.6	30.3	41.3	17	21.2
PSM [34]	12.0	31.7	43.0	16	21.9
T2VLAD [50]	12.7	34.8	47.1	12	-
Proposed Method	**14.7**	**37.0**	**48.6**	**11**	**25.6**

Table 2: Video retrieval performance results on *MSR-VTT-Miech* dataset.

Method	Video Retrieval Performance				
	R@1↑(%)	**R@5↑(%)**	**R@10↑(%)**	**MedR↓**	**mAP↑(%)**
W2VV [8]	2.7	12.5	17.3	83	7.9
VSE++ [13]	17.0	40.9	52.0	10	16.9
W2VV++ [31]	21.7	48.6	60.9	6	34.4
TCE [59]	17.1	39.9	53.7	9	-
HGR [5]	22.9	50.2	63.6	5	35.9
MMT [17]	20.3	49.1	63.9	6	-
HSL [10]	23.0	50.6	62.5	5	36.1
PSM [34]	24.2	53.0	65.3	5	37.9
T2VLAD [50]	26.1	54.7	68.1	4	-
Proposed Method	**27.8**	**57.3**	**68.7**	**4**	**41.2**

4.3 Performance Evaluation

Video retrieval performance is measured by rank-based metrics such as recall at K (R@K), median rank (MedR), and mean average precision (mAP). R@K (K=1, 5, 10) indicates the percentage of queries that find correct samples among the top K results. MedR indicates the median rank of the first correct sample in the retrieved results. mAP represents the mean of the average precision scores for each query. Higher R@K, mAP and lower MedR indicate better performances.

Video Retrieval Performance Comparison. We perform performance comparisons according to the training and testing protocol [5, 10], not with the pretraining-based methods using a large amount of additional visual-text data [1, 12, 38]. We conduct the experiments on MSR-VTT with different types of data partitions *MSR-VTT-Original* [56] and *MSR-VTT-Miech* [40]. Table 1 and

Table 3: Video retrieval performance results on VATEX dataset.

Method	Video Retrieval Performance				
	R@1↑(%)	R@5↑(%)	R@10↑(%)	MedR↓	mAP↑(%)
W2VV [8]	14.6	36.3	46.1	-	-
VSE++ [13]	31.3	65.8	76.4	-	-
CE [36]	31.1	68.7	80.2	-	-
W2VV++ [31]	32.0	68.2	78.8	-	-
Dual Encoding [9]	31.1	67.4	78.9	3	-
HGR [5]	35.1	73.5	83.5	2	-
HSL [10]	36.8	73.6	83.7	2	52.0
HGR (+GPO) [4]	37.3	73.4	82.4	-	-
Proposed Method	**39.0**	**75.6**	**84.1**	**2**	**55.3**

Table 4: Video retrieval performance results on TGIF dataset.

Method	Video Retrieval Performance				
	R@1↑(%)	R@5↑(%)	R@10↑(%)	MedR↓	mAP↑(%)
W2VV++ [31]	9.4	22.3	29.8	48	16.2
Dual Encoding [9]	9.1	21.3	28.6	50	15.7
HGR [5]	4.5	12.4	17.8	160	-
CF-GNN [49]	10.2	23.0	30.7	44	-
SEA (BERT) [32]	10.7	24.4	31.9	37	17.9
SEA (BERT+biGRU) [32]	11.1	25.2	32.8	35	18.5
Proposed Method	**11.5**	**26.3**	**34.9**	**31**	**19.2**

2 show the text-to-video retrieval performances on MSR-VTT dataset. In case of MSR-VTT, we use (ResNeXt101+ResNet152) for visual and (VGGish) for audio. [8,10,13,31,34] in tables use the same visual feature as us. While, multimodal models [17, 50] use (VGGish+Speech) audio features in addition to many visual features together (DenseNet161+SENet154 +ResNet50+S3D+OCR). Ours outperforms these even with simple and fewer features. We perform the video retrieval experiments on VATEX according to the data split [5]. Table 3 shows the comparison results on VATEX dataset. In case of VATEX, we use (I3D) for visual and (VGGish) for audio. All the other methods use the same visual feature as us. The proposed method surpasses the other state-of-the-art methods in terms of all evaluation metrics. These MSR-VTT and VATEX include both mismatching cases (*i.e.*, missing audio or existing audio but not matched). To validate our method on fully missing audio condition, we conduct the experiments on TGIF which does not have audio data at all. We utilize MSR-VTT for training AVA-Memory concurrently. In terms of TGIF experiment, MSR-VTT is additional data but not supervision data because we do not use any text labels of MSR-VTT. Thus, we use the same video-text pairs with the other methods. At testing time, only associated audio cues are used dominantly (α=0). We conduct

Table 5: Effects of the network designs on the performances. Performance evaluations are conducted on *MSR-VTT-Miech*.

Method	Video Retrieval Performance				
	R@1↑(%)	R@5↑(%)	R@10↑(%)	MedR↓	mAP↑(%)
w/o Audio Association	26.5	56.4	68.1	4	40.3
w/o Mismatch-Aware Adjustment	25.0	55.0	66.6	4	39.1
Proposed Method	**27.8**	**57.3**	**68.7**	**4**	**41.2**

Table 6: Effects of using different dataset in associative learning procedure to validate the generalizability of audio-visual association on TGIF dataset.

Method	Associative Learning	Video Retrieval Performance	
		R@1↑(%)	mAP↑(%)
w/o AVA-memory	✗	10.8	18.5
Proposed Method	✓ (w/ MSR-VTT)	**11.5**	**19.2**

the video retrieval experiments on TGIF as follows the split [32]. We use only (ResNeXt101+ResNet152) for visual and state-of-the-art [32] uses the same one. As a result, ours outperforms the other methods on TGIF as shown in Table 4.

4.4 Ablation Study

Effects of network designs. We analyze the effects of the network design by conducting ablation studies as shown in Table 5. We investigate the effectiveness of audio association and mismatch-aware adjustment. In the table, the model 'w/o Audio Association' indicates the MA-Transformer without using associated audio features at mismatch-aware semantic embedder. It means the video embedding is made with only visual and actual audio cues (adjustment is still conducted for only actual audio). The second model 'w/o Mismatch-Aware Adjustment' indicates the model without adjusting weights between actual audio and associated audio. As shown in the table, the final proposed method clearly outperforms the base models 'w/o Audio Association' and 'w/o Mismatch-Aware Adjustment'. The results show the superiority of the proposed network designs in terms of both association and adjustment.

Generalizability of AVA-Memory. To validate the generalizability of AVA-memory in terms of audio-visual association, we observe the effect of exploiting different dataset in associative learning. Table 6 shows the results about the generalizability of audio-visual association on TGIF dataset. The first baseline model is trained without AVA-Memory. The second one is the model that utilizes a different dataset, MSR-VTT in associative learning. The training set of TGIF

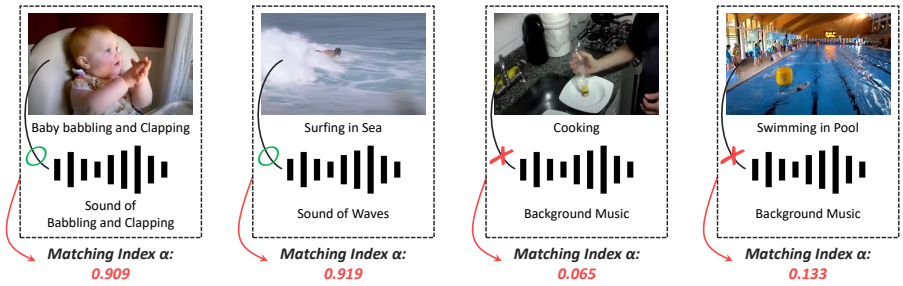

Fig. 4: Matching index examples obtained by AVA-Memory on VATEX dataset.

is mainly used for learning with the video-text matching loss $\mathcal{L}_M$ while MSR-VTT is used for learning with the associative loss $\mathcal{L}_A$. Then, the models are evaluated on the test set of TGIF. As shown in the table, the associative learning with the different dataset (*i.e.*, MSR-VTT) also can fairly contribute to the retrieval performances. Note that TGIF does not include actual audio at all and fully exploits the associated audio cues at searching time. This result shows the generalizability of audio-visual association in terms of training data. Since the proposed audio-visual associative learning does not require any labels at all, any videos with audio-visual information are available for this learning scheme.

4.5 Qualitative Results on Matching Index

Figure 4 shows the matching index examples on VATEX test dataset. Each matching index is obtained by AVA-Memory according to equation (11). As can be seen in the figure, matching indexes are high for video samples in which visual and audio are well matched. Contrary, the matching index is low for samples where visual and audio are not matched (*e.g.*, music). Such matching index values are convincingly obtained and used in the adjustment process.

5 Conclusion

The objective of the proposed work is to address the audio-visual mismatch condition when retrieving videos from the text semantics. To this end, we propose MA-Transformer with AVA-Memory which enables to associate audio cues from visual and further to adjust the audio embeddings considering the degree of matching between visual and audio cues. As a result, the proposed method outperforms the state-of-the-art video retrieval methods on various datasets including audio-visual mismatch conditions. Further, we validate the network designs by conducting ablation studies and qualitative analyses.

Acknowledgement. This work was supported by IITP grant(No. 2020-0-00004).

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1728–1738 (2021)
2. Cai, Q., Pan, Y., Yao, T., Yan, C., Mei, T.: Memory matching networks for one-shot image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4080–4088 (2018)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6299–6308 (2017)
4. Chen, J., Hu, H., Wu, H., Jiang, Y., Wang, C.: Learning the best pooling strategy for visual semantic embedding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15789–15798 (2021)
5. Chen, S., Zhao, Y., Jin, Q., Wu, Q.: Fine-grained video-text retrieval with hierarchical graph reasoning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10638–10647 (2020)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (ICML). pp. 1597–1607. PMLR (2020)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT) (2019)
8. Dong, J., Li, X., Snoek, C.G.: Predicting visual features from text for image and video caption retrieval. IEEE Transactions on Multimedia **20**(12), 3377–3388 (2018)
9. Dong, J., Li, X., Xu, C., Ji, S., He, Y., Yang, G., Wang, X.: Dual encoding for zero-example video retrieval. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9346–9355 (2019)
10. Dong, J., Li, X., Xu, C., Yang, X., Yang, G., Wang, X., Wang, M.: Dual encoding for video retrieval by text. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2020)
12. Dzabraev, M., Kalashnikov, M., Komkov, S., Petiushko, A.: Mdmmt: Multidomain multimodal transformer for video retrieval. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3354–3363 (2021)
13. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. In: British Machine Vision Conference (BMVC) (2018)
14. Francis, D., Anh Nguyen, P., Huet, B., Ngo, C.W.: Fusion of multimodal embeddings for ad-hoc video search. In: IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) (2019)
15. Fu, Z., Liu, Q., Fu, Z., Wang, Y.: Stmtrack: Template-free visual tracking with space-time memory networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13774–13783 (2021)

16. Gabeur, V., Nagrani, A., Sun, C., Alahari, K., Schmid, C.: Masking modalities for cross-modal video retrieval. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1766–1775 (2022)
17. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: European Conference on Computer Vision (ECCV) (2020)
18. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M., Venkatesh, S., Hengel, A.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1705–1714 (2019)
19. Han, T., Xie, W., Zisserman, A.: Memory-augmented dense predictive coding for video representation learning. In: European Conference on Computer Vision (ECCV). pp. 312–329. Springer (2020)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
21. Hu, P., Peng, X., Zhu, H., Zhen, L., Lin, J.: Learning cross-modal retrieval with noisy labels. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5403–5413 (2021)
22. Huang, Y., Wang, L.: Acmm: Aligned cross-modal memory for few-shot image and sentence matching. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5774–5783 (2019)
23. Kaiser, Ł., Nachum, O., Roy, A., Bengio, S.: Learning to remember rare events. In: International Conference on Learning Representations (ICLR) (2017)
24. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3128–3137 (2015)
25. Kim, J.U., Park, S., Ro, Y.M.: Robust small-scale pedestrian detection with cued recall via memory learning. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3050–3059 (2021)
26. Kim, M., Hong, J., Park, S.J., Ro, Y.M.: Multi-modality associative bridging through memory: Speech sound recollected from face video. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 296–306 (2021)
27. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)
28. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
29. Lee, S., Kim, H.G., Choi, D.H., Kim, H.I., Ro, Y.M.: Video prediction recalling long-term motion context via memory alignment learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3054–3063 (2021)
30. Lee, S., Kim, H.I., Ro, Y.M.: Weakly paired associative learning for sound and image representations via bimodal associative memory. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10534–10543 (2022)
31. Li, X., Xu, C., Yang, G., Chen, Z., Dong, J.: W2VV++: fully deep learning for ad-hoc video search. In: ACM International Conference on Multimedia (ACM MM). pp. 1786–1794 (2019)
32. Li, X., Zhou, F., Xu, C., Ji, J., Yang, G.: Sea: Sentence encoder assembly for video retrieval by textual queries. IEEE Transactions on Multimedia **23**, 4351–4362 (2021)
33. Li, Y., Song, Y., Cao, L., Tetreault, J., Goldberg, L., Jaimes, A., Luo, J.: Tgif: A new dataset and benchmark on animated gif description. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4641–4650 (2016)

34. Liu, H., Luo, R., Shang, F., Niu, M., Liu, Y.: Progressive semantic matching for video-text retrieval. In: ACM International Conference on Multimedia (ACM MM). pp. 5083–5091 (2021)
35. Liu, S., Fan, H., Qian, S., Chen, Y., Ding, W., Wang, Z.: Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11915–11925 (2021)
36. Liu, Y., Albanie, S., Nagrani, A., Zisserman, A.: Use what you have: Video retrieval using representations from collaborative experts. In: British Machine Vision Conference (BMVC) (2019)
37. Marchetti, F., Becattini, F., Seidenari, L., Bimbo, A.D.: Mantra: Memory augmented networks for multiple trajectory prediction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7143–7152 (2020)
38. Miech, A., Alayrac, J.B., Laptev, I., Sivic, J., Zisserman, A.: Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9826–9836 (2021)
39. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9879–9889 (2020)
40. Miech, A., Laptev, I., Sivic, J.: Learning a text-video embedding from incomplete and heterogeneous data. arXiv preprint arXiv:1804.02516 (2018)
41. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2630–2640 (2019)
42. Mithun, N.C., Li, J., Metze, F., Roy-Chowdhury, A.K.: Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In: ACM International Conference on Multimedia Retrieval (ICMR). pp. 19–27 (2018)
43. Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J., Yokoya, N.: Learning joint representations of videos and sentences with web image search. In: European Conference on Computer Vision (ECCV). pp. 651–667. Springer (2016)
44. Park, H., Noh, J., Ham, B.: Learning memory-guided normality for anomaly detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14372–14381 (2020)
45. Patrick, M., Huang, P.Y., Asano, Y., Metze, F., Hauptmann, A.G., Henriques, J.F., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. In: International Conference on Learning Representations (ICLR) (2020)
46. Song, Y., Soleymani, M.: Polysemous visual-semantic embedding for cross-modal retrieval. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1979–1988 (2019)
47. Torabi, A., Tandon, N., Sigal, L.: Learning language-visual embedding for movie understanding with natural-language. arXiv preprint arXiv:1609.08124 (2016)
48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 5998–6008 (2017)
49. Wang, W., Gao, J., Yang, X., Xu, C.: Learning coarse-to-fine graph neural networks for video-text retrieval. IEEE Transactions on Multimedia **23**, 2386–2397 (2021)
50. Wang, X., Zhu, L., Yang, Y.: T2vlad: global-local sequence alignment for text-video retrieval. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5079–5088 (2021)

51. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4581–4591 (2019)
52. Wei, J., Xu, X., Yang, Y., Ji, Y., Wang, Z., Shen, H.T.: Universal weighting metric learning for cross-modal matching. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13005–13014 (2020)
53. Wei, J., Yang, Y., Xu, X., Zhu, X., Shen, H.T.: Universal weighting metric learning for cross-modal retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
54. Wray, M., Larlus, D., Csurka, G., Damen, D.: Fine-grained action retrieval through multiple parts-of-speech embeddings. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 450–459 (2019)
55. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1492–1500 (2017)
56. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5288–5296 (2016)
57. Xu, R., Xiong, C., Chen, W., Corso, J.: Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In: AAAI Conference on Artificial Intelligence (AAAI) (2015)
58. Yang, T., Chan, A.B.: Learning dynamic memory networks for object tracking. In: European Conference on Computer Vision (ECCV). pp. 152–167 (2018)
59. Yang, X., Dong, J., Cao, Y., Wang, X., Wang, M., Chua, T.S.: Tree-augmented cross-modal encoding for complex-query video retrieval. In: International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR). pp. 1339–1348 (2020)
60. Yu, Y., Kim, J., Kim, G.: A joint sequence fusion model for video question answering and retrieval. In: European Conference on Computer Vision (ECCV). pp. 471–487 (2018)
61. Zhang, B., Hu, H., Sha, F.: Cross-modal and hierarchical modeling of video and text. In: European Conference on Computer Vision (ECCV). pp. 374–390 (2018)
62. Zhu, L., Yang, Y.: Inflated episodic memory with region self-attention for long-tailed visual recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4344–4353 (2020)