# Query-Adaptive Late Fusion for Hierarchical Fine-Grained Video-Text Retrieval

Wentao Ma, Qingchao Chen, Fang Liu, Tongqing Zhou, and Zhiping Cai

***Abstract*— Recently, a hierarchical fine-grained fusion mechanism has been proved effective in cross-modal retrieval between videos and texts. Generally, the hierarchical fine-grained semantic representations (video-text semantic matching is decomposed into three levels including global-event representation matching, action–relation representation matching, and local-entity representation matching) to be fused can work well by themselves for the query. However, in real-world scenarios and applications, existing methods failed to adaptively estimate the effectiveness of multiple levels of the semantic representations for a given query in advance of multilevel fusion, resulting in a worse performance than expected. As a result, it is extremely essential to identify the effectiveness of hierarchical semantic representations in a query-adaptive manner. To this end, this article proposes an effective query-adaptive multilevel fusion (QAMF) model based on manipulating multiple similarity scores between the hierarchical visual and text representations. First, we decompose video-side and text-side representations into hierarchical semantic representations consisting of global-event level, action-relation level, and local-entity level, respectively. Then, the multilevel representation of the video-text pair is aligned to calculate the similarity score for each level. Meanwhile, the sorted similarity score curves of the good semantic representation are different from the inferior ones, which exhibit a "cliff" shape and gradually decline (see Fig. 1 as an example). Finally, we leverage the Gaussian decay function to fit the tail of the score curve and calculate the area under the normalized sorted similarity curve as the indicator of semantic representation effectiveness, namely, the area of good semantic representation is small, and vice versa. Extensive experiments on three public benchmark video-text datasets have demonstrated that our method consistently outperforms the state-of-the-art (SoTA). A simple demo of QAMF will soon be publicly available on our homepage: https://github.com/Lab-ANT.**

***Index Terms*— Fine-grained fusion, Gaussian decay, query-adaptive, semantic representation.**

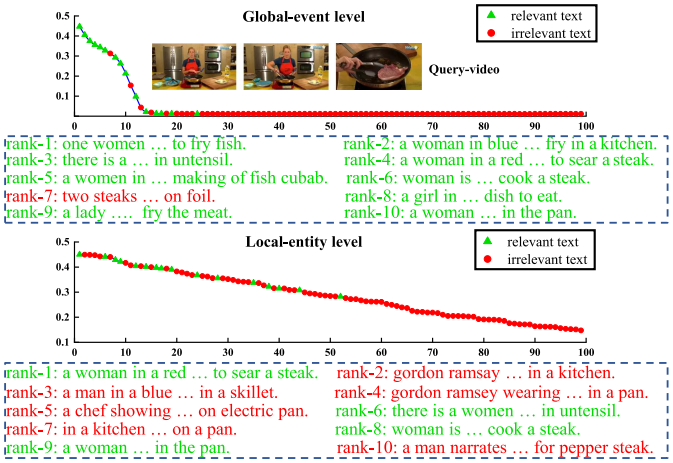


Fig. 1. Example of a multilevel semantic representations system. For a query video in the MSR-VTT dataset, the global-event representation level (top) and local-entity representation level (bottom) are employed to obtain two similarity score lists, respectively. There are 20 relevant texts for this query, where global-event representation produces good performance, but local-entity representation fails. And we plot the sorted scores for rank lists 1–100, and the corresponding ten top-ranked texts. Relevant texts are in green and irrelevant ones red. Note that the sorted score curve is cliff-shaped for global-event representation, but gradually descending for local-entity representation.

## I. Introduction

THIS article tackles the problem of joint video-text and text-video cross-modal retrieval. Given a query video (or natural language text queries), the aim of a cross-modal retrieval system is to search for all the semantically relevant and similar natural language texts (or videos) in a database. Hence, it is of significant theoretical and practical implications value to investigate effective cross-modal retrieval methods and apply them in real-world scenarios to promote the development of diversified retrieval techniques. Lately, to improve the performance of video-text cross-modal retrieval by exploiting both global and local semantic representations in text and video, various fine-grained cross-modal retrieval approaches have investigated the following strategies [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], including adopting the attention mechanism to gather valuable cross-modal and temporal cues [13], using temporal fusion mechanism to represent videos and texts, respectively, and aligns local components to compute overall similarities [2], [3], [4], parsing the video-text pairs into different semantic representation levels [1], [5], [7], [9], [10]. It is proven that fine-grained semantic detail representation and multilevel representation fusion have been extremely salutary for boosting cross-modal retrieval performances.

It is well acknowledged that the search accuracy would be very high if using a given query with good semantic representations. On the contrary, leveraging inferior ones will produce lower search quality. Ideally, if a to-be-fused semantic representation is effective that also complements existing ones,

Manuscript received 30 December 2021; revised 3 June 2022 and 9 August 2022; accepted 4 October 2022. Date of publication 25 October 2022; date of current version 3 May 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62072465, Grant 62172155, and Grant 62102425; in part by the Science and Technology Innovation Program of Hunan Province under Grant 2021RC2071; and in part by the Postgraduate Research and Innovation Project of Hunan Province under Grant CX20210080. *(Corresponding authors: Qingchao Chen; Fang Liu; Zhiping Cai.)*

Wentao Ma, Tongqing Zhou, and Zhiping Cai are with the College of Computer, National University of Defense Technology, Changsha 410005, China (e-mail: wtma@nudt.edu.cn; zhoutongqing@nudt.edu.cn; zpcai@nudt.edu.cn).

Qingchao Chen is with the National Institute of Health Data Science, Peking University, Beijing 100091, China (e-mail: qingchao.chen@pku.edu.cn).

Fang Liu is with the School of Design, Hunan University, Changsha 410082, China (e-mail: fangl@hnu.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2022.3214208

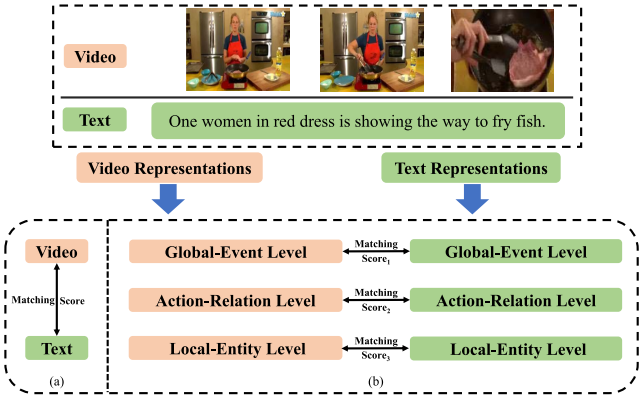


Fig. 2. Different cross-modal matching approaches. (a) Basic single vector-based similarity. (b) Multilevel representation matching.

a higher search performance may be achieved. However, in a realistic situation, an essential issue comes where we do not know in advance of fusion whether a complementary semantic representation is effective or not for a given query video (or text). Without identifying the risk of using or the quality of representation information may degrade the retrieval performance. Therefore, it is necessary to predict the semantic representation effectiveness with relatively satisfactory performance, and, in fact, more essential, to design a mechanism that is capable of promoting good representations and punishing inferior ones. It is a matter of concern that some state-of-the-art (SoTA) hierarchical fine-grained representation fusion methods [9], [10], [11] take the average of cross-modal similarity scores at all levels as final video-text matching similarity resulting in a retrieval performance worse than expected.

In light of the above analysis, this article proposes an effective query-adaptive multilevel fusion (QAMF) model for hierarchical fine-grained video-text retrieval which takes the advantage of global-to-local semantic representation and makes up their deficiencies. Similar to [1], [5], [9], [10], and [14], we decompose video-text matching into global-to-local three levels of representation matching, including global-event representation matching, action–relation representation matching, and local-entity representation matching, as shown in Fig. 2. Moreover, to capture the interaction between different levels of semantic representation, we propose a text transformer-graph inference module and align the cross-modal components of each representation level by the attention mechanism. For adaptive late fusion, our motivation is simple and effective: the similarity score curve for a good level of semantic representation is very steep, in a "cliff" shape, while that of an inferior one is gradually dropping, in a "hill slope" shape. Then, by fitting the sorted similarity score curve's tail with the Gaussian decay function, the area under the normalized similarity score curve can be regarded as the surrogate estimation of the semantic representation effectiveness. In the end, the late fusion weight of semantic representation at each level is assigned adaptively by the area under the normalization curve. To the best of our current knowledge, compared with the previous hierarchical fine-grained fusion methods, our QAMF can achieve SoTA performance on three public benchmark video-text datasets, including MSR-VTT [15], TGIF [16], and VATEX [17]. The main contributions of this article are as follows.

1) To tackle the hierarchical matching framework that fails to realize adaptive fusion, we propose a query-adaptive fusion mechanism (*called query-adaptive fusion*) to enable differential fusion of multilevel semantic representation based on representation merit estimation, rescoring, and attention assignment.
2) To capture the semantic interactions between the text graph nodes under different representation levels, we propose a transformer-based graph inference mechanism (*called text graph-transformer*) and embed it in the text encoding procedure.
3) Extensive experimental results on three public benchmark video-text datasets have demonstrated that our QAMF consistently outperforms the SoTA ones with a preferable margin.

The remainder of this article is organized as follows. First, we briefly review the related works in Section II and Section III introduces the design of our QAMF for fine-grained video-text retrieval. Then, we present the experimental settings and results in Sections IV and V. Finally, conclusions are given in Section VI.

## II. Related Works

The work related to this article includes visual-text matching, fine-grained cross-modal matching, and multimodal fusion, which are discussed in Sections II-A–II-C.

### A. Visual-Text Matching

Given a set of query images/videos (or natural language texts), our goal is to search the most relevant natural language texts (or images/videos) [18], [19].

Image-text matching retrieval has long been tackled via encoding images and sentences into fix-dimensional vectors and mapping them to a common latent space for similarity matching [6], [7], [20], [21], [22], [23]. Lee et al. [6] propose the stacked cross-attention mechanism to align each region of the image with the word, which greatly improved the alignment performance. Gu et al. [23] merge image and caption generation in a multitask framework to enrich the global representation. However, this method can hardly cover complex semantic representation information by a single fixed-dimensional vector. Wu et al. [7] parse text into objects, attributes, relationships, and sentences for multilevel alignment matching. Although image-text matching and video-text matching are both visual-text matching, the spatial–temporal evolution properties of video-text matching make it more complicated [2], [24]. To improve the alignment of the video–text, Yu et al. [2] estimate video-text similarity by the dense pairs between each word of the text and each frame of the video. Zhang et al. [24] implement video-to-text retrieval by parsing a hierarchical decomposition of the video-text.

The above methods have achieved gratifying performance for visual-text matching; nonetheless, these methods ignore the

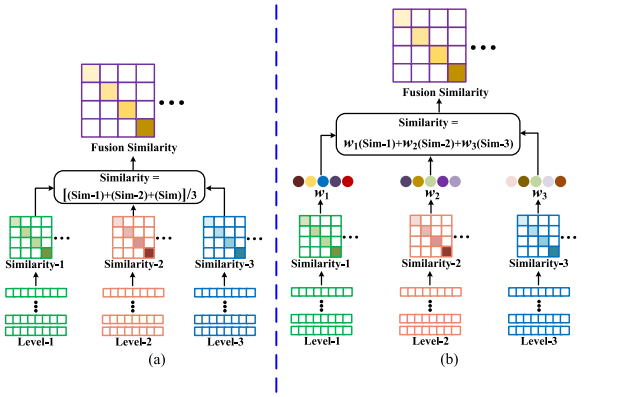


Fig. 3. Late fusion structure comparison. Different from the existing structure of late fusion (such as (a) HGR [10]), our proposed (b) QAMF realizes adaptive fusion by (12)–(17).

fine-grained representation information such as barely describing the image from global-to-local semantic representations and hardly capturing semantic interaction in a visual–text pair.

### B. Fine-Grained Cross-Modal Matching

In order to capture the visual–textual semantic detail representations, the problem of fine-grained cross-modal matching has been extensively investigated, leading to various methods [1], [2], [4], [5], [6], [9], [10], [13], [20], [25], [26]. Among them, these approaches have achieved better performance in fine-grained image-text retrieval [6], but learning semantic alignments between video-text pairs is more challenging. Furthermore, the image-text sequence representations ignore the topological structures, which makes them hard to capture the relationship between the local components within a global event. To enrich the representation of videos, Mithun et al. [13] and Liu et al. [26] adopt multimodal representation information from videos such as speech contents, action cues, and scene description sentences for video encoding. Yu et al. [2] and Song and Soleymani [4] adopt a sequence of video frames and text words to represent the fine-grained semantic representation information of the video-text pair and calculate the overall similarity by aligning the local components.

In terms of capturing hierarchical fine-grained semantic representations, our work is most similar to Wray et al. [9] and Chen et al. [10]. For the former, this method parses action phrases into different part-of-speech such as verbs and nouns for fine-grained retrieval. However, the sentences of text are more complicated than action phrases and global events will be ignored if only action phrases are considered. For the latter, it disentangles video-text matching into three hierarchical semantic representations matching, which is responsible for capturing global events, local actions, and entities, respectively. The semantic matching scores of these three levels are fused together as the final video-text similarity to enhance fine-grained semantic coverage.

### C. Multimodal Fusion

Various works have been investigated toward deep multimodal fusion [11], [27], [28]. The two main streams for multimodal fusion can be specified as early and late fusion, depending on steps to fuse, which have been discussed in image search [29], [30], [31], cross-modal retrieval [2], [9], [10], [13], [23], [26], multiview clustering [27], [32], [33], [34], and multimodal machine learning [11], [13], [35], [36]. In early fusion, descriptors employ a certain operation (*e.g.,* averaging, concatenation, self-attention, and compression) at the feature level or pixel level for fusion. Then, the fused features are processed together through the learning methods. While late fusion refers to fusion at the score or decision level. In late fusion, a good tradeoff can be made between the feature representation content and the efficiency in fusion.

In cross-modal retrieval, Yu et al. [2] propose the JSFusion model for estimating video-text hierarchical semantic similarity by dense pairwise comparisons between each word of the text and each frame of the video. Wray et al. [9] and Chen et al. [10], respectively, disentangle video-text pairs into different semantic representations and fuse video-text matching at different levels. Yet, these two approaches are not appropriate to assign fixed weight to all levels of semantic representations for fusion: for a given query, we should estimate the effectiveness of a level in a query-adaptive manner, so there is fall back if an ineffective level is integrated. Therefore, in our work, we propose a QAMF via similarity, as shown in Fig. 3. Our QAMF model estimates the effectiveness of each fused semantic representation in a query-adaptive manner. This allows the effectiveness of different semantic representation levels based on the similarity score it shares with each query (a video or a text), so that those "good" semantic representation levels are endowed with larger weights for providing greater contributions, while the "bad" ones are punished, thus attaining differential fusion of hierarchical semantic detail representation.

## III. Design of QAMF

To implement our QAMF model, we choose hierarchical graph reasoning (HGR) [10] as the basic model and integrate the proposed components (*i.e.*, text graph–transformer and query-adaptive fusion) into it to attain our QAMF, for multilevel fusion video-text matching. Fig. 4 illustrates the overview of our QAMF that consists of three modules: 1) text encoding module; 2) video encoding module; and 3) query-adaptive fusion module. Video–text pair alignment matching that from global-to-local levels and adaptive late-fusion to calculate the overall cross-modal similarity.

### A. Video-Text Hierarchical Encoding

Each overall sentence about the video is a summary description of the global event, which usually includes action–relation and local-entity (*e.g.,* agent and patient of the interactive relationship). The adoption of global-to-local levels for video descriptions can comprehensively understand the interaction between various entities, thereby enabling better semantic coverage of text representations.

*1) Text Encoding:* Transforming text descriptions into semantic graphs has been extensively investigated; in our

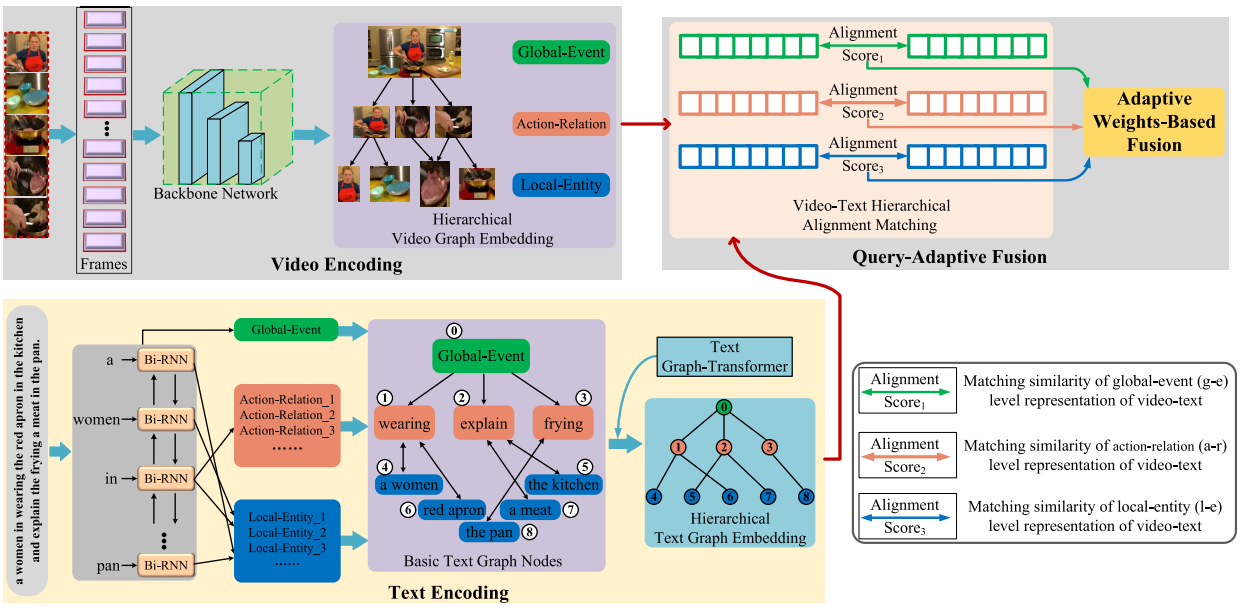


Fig. 4. Overview of our effective QAMF model based on manipulating multiple similarity scores between the hierarchical visual and text representations for video–text retrieval. First, we decompose the video-side and text-side into hierarchical semantic representations composed of global-event level, action–relation level, and local-entity level, respectively. Second, the multilevel is aligned to calculate the similarity score for each level. Finally, the query-adaptive fusion module is leveraged to effectively fuse the similarity of all levels.

study, we follow the works [10] and [14] that disentangle the video description into three levels of the semantic representation, which are global-event representation level, action–relation representation level, and local-entity representation level. To be specific, given a video description $T$ with $N$ words $\{T_1, \ldots, T_n\}$, where $n \in [1, N]$. For the global-event representation, we mainly capture the event described in the sentence as a whole. Hence, we first utilize word2vec to convert $T$ that consists of into word vector embeddings $\{t_1, \ldots, t_n\}$, where $n \in [1, N]$

$$t_n = \text{word2vec}(T_n), \quad n \in [1, N]. \tag{1}$$

Then, we aggregate the word embedding vectors by an attention mechanism to focus on important events in the sentence. Thus, the global-event sentence representation $c_g$ is given by

$$c_g = \sum_{i=1}^{N} \alpha_{g,i}\ t_i \tag{2}$$

$$\alpha_{g,i} = \frac{\exp\left(u_g t_i\right)}{\sum_{j=1}^{N} \exp\left(u_g t_j\right)} \tag{3}$$

where $u_g$ is the parameter to be learned by the model that focuses on important hidden layer representations. Meanwhile, to obtain more fine-grained semantic information, we adopt an off-the-shelf semantic role parsing toolkit [37] to get verbs, noun phrases, and their semantic role relations in the sentence $T$. As a global event is composed of different action–relations, the second level of the text graph is the action–relation representation level, whose nodes are verb phrases. Then the remaining third level is naturally the local-entity level, whose nodes are noun phrases. For action–relation and local-entity nodes, we employ max pooling over words in each node as action–relation node representations $c_a = \{c_{a,1}, \ldots, c_{a,N_a}\}$ and local-entity node representations $c_l = \{c_{l,1}, \ldots, c_{l,N_l}\}$, where $N_a$ and $N_l$ are numbers of action–relation and local-entity nodes, respectively (following the HGR [10] and hierarchical cross-modal graph consistency (HCGC) [14]).

In terms of edge connections, the verb phrases are considered action–relation nodes and connected to the sentence node with direct edges. Since the sentence nodes contain global-event semantic representation information, the contextual relationships between action–relation nodes can be implicitly learned from the sentence nodes in the graph reasoning. While the noun phrases are local-entity nodes that are connected with different action–relation nodes. As a result, the edge type between the local-entity nodes and the action–relation nodes is determined by the semantic role of the local-entity in reference to the action–relation. In particular, similar to [10], if a local-entity node provides multiple semantic roles to different action–relation nodes, we will duplicate the local-entity nodes for each semantic role. In Fig. 4, we give an example of the constructed hierarchical text graph.

After constructing the text graph, the graph–transformer is leveraged to learn the semantic interactions between the nodes of different levels. As a result, given the initialized node representation $c_i \in \{c_g, c_a, c_l\}$, the graph–transformer is utilized to select relevant context information from neighbor nodes to enhance the representation for each node

$$\tilde{\beta}_{ij} = \left(u_a^q c_i^l\right)^T \left(u_a^p c_j^l\right) / \sqrt{D} \tag{4}$$

$$\beta_{ij} = \frac{\exp\left(\tilde{\beta}_{ij}\right)}{\sum_{j \in \mathcal{N}_i} \exp\left(\tilde{\beta}_{ij}\right)} \tag{5}$$

where $\mathcal{N}_i$ represents of neighborhood nodes of node $i$, $u_a^q$ and $u_a^p$ are parameters to compute the attention of graph–transformer, $c_i^l$ is the output representation of node $i$ at the

$l$th graph reasoning layer, and $D$ is the dimension of the node representation. Then, the shared $u_t$ is adopted to transform contexts from attended neighbor nodes to node $i$ with a residual connection

$$c_i^{l+1} = c_i^l + u_t^{l+1} \sum_{j \in \mathcal{N}_i} \left(\beta_{ij} c_j^l\right). \tag{6}$$

As a result, after the transformer-based graph reasoning process, we can obtain the final node representations of the hierarchical text graph, namely $c_g$ indicates the global-event node representation, $c_a$ for action–relation node representation, and $c_l$ for local-entity node representation.

*2) Video Encoding:* Different from the text, for video encoding, we adopt the pretrained ResNet model for hierarchical representation. Similar to [1] and [10], we obtain different levels of semantic representation by constructing three independent video embeddings. Specifically, given video $V$ containing a frame sequence $\{f_1, \ldots, f_m\}$ as input, where $m \in [1, M]$. We adopt three different linear transformation weights $u_g^v$, $u_a^v$ and $u_l^v$ to encode the video frame sequence into three levels of visual vector embedding

$$y_{k,i} = u_k^g f_m, \quad k \in \{g, a, l\}. \tag{7}$$

Moreover, we reuse the above attention mechanism similar to (2) and (3) to obtain a global representation vector to denote the overall event description in the video as $y_g$. While the action–relation representation and local-entity representation, the video fine-grained semantic representations are a frame-wise features of time segmentation sequence $y_a = \{y_{a,1}, \ldots, y_{a,m}\}$ and $y_l = \{y_{l,1}, \ldots, y_{l,m}\}$, respectively. Therefore, the final parsing video hierarchical coding representation is: $y_g$ indicates the global-event representation, $y_a$ for action–relation representation, and $y_l$ for local-entity representation. These semantic detail features will be sent to the query-adaptive fusion module to match with their corresponding textual feature representations at different levels to align videos and texts.

### B. Query-Adaptive Fusion

To improve the performance of video-text pair matching, we adaptively fuse results from the three levels of semantic representation for the overall cross-modal similarity.

*1) Differential Merits of the Representations:* In information retrieval, for a specific query, a good semantic representation means that its search accuracy is high. In contrast, the semantic representation with low search quality is called the inferior one. When the adopted semantic representation information is good and complementary to existing ones, a higher performance is expected. Yet, due to the low discriminability of semantic representation, many irrelevant results have high similarity scores. Specifically, the formal description of the extreme case in the video–text cross-modal retrieval is as follows: the best and worst semantic representations for a given query video $v_q$ (or text $t_q$). In a video–text dataset containing $V$ videos with $T$ text descriptions for each video, for simplicity, we assume that: 1) there is only one relevant text description $t^*$ to query video (there is only one relevant video $v^*$ to query text) that 2) the text description (or video) similarity scores are normalized with a maximum value of 1. Intuitively, the best level of semantic representation satisfies the following requirements:

$$s_{t,v_q}^{(\text{best})} = \begin{cases} 1, & \text{if } t = t^* \\ 0, & \text{otherwise} \end{cases}, \quad t = 1, 2, \ldots, T \tag{8}$$

$$s_{v,t_q}^{(\text{best})} = \begin{cases} 1, & \text{if } v = v^* \\ 0, & \text{otherwise} \end{cases}, \quad v = 1, 2, \ldots, V \tag{9}$$

where $s_{t,v_q}^{(\text{best})}$ is the similarity score of text description $t$ to query $v_q$ ($s_{v,t_q}^{(\text{best})}$ is the similarity score of text description $v$ to query $t_q$) *with respect to* the best semantic representation. Only the similarity score of the relevant text $t$ or video $v$ is 1, and all other irrelevant texts (or videos) are 0. In contrast, the worst semantic representation has completely different results, that is,

$$s_{t,v_q}^{(\text{worst})} = \begin{cases} 0, & \text{if } t = t^* \\ 1, & \text{otherwise} \end{cases}, \quad t = 1, 2, \ldots, T \tag{10}$$

$$s_{v,t_q}^{(\text{worst})} = \begin{cases} 0, & \text{if } v = v^* \\ 1, & \text{otherwise} \end{cases}, \quad v = 1, 2, \ldots, V. \tag{11}$$

Therefore, (8)–(11) are defined by the discrimination ability of semantic representation to obtain the similarity score curve, once sorted, exhibit a "cliff" shape and a "hill slope" shape, respectively. Then, we find that the effectiveness of semantic representation is estimated as negatively related to the area under the normalized similarity score curve.

*2) Rescoring Ranking List With Decay Function:* As mentioned above, we estimate the effectiveness of semantic representation by the area of the sorted similarity score curve. However, the global-event curve drops sharply into a "cliff" shape, and it is quite easy to tell that global-event is a good representation level. But the effectiveness of action–relation and local-entity is not so obvious: both sorted similarity score curves have a relatively "high tail," and scores of the top-ranked videos/texts are not remarkably higher than the tail, for example, many irrelevant results have high ranks due to the low discriminability of inferior semantic representation.

To alleviate the impact of the "high tail," Zheng et al. [31] proposed to find a reference score curve in the irrelevant data for each query to approximate the tail of the initial ranking similarity score curve, and if subtracted, would highlight the protrusion of the top-ranked scores. However, the method to construct the reference curve is complicated and can hardly be generated online. Inspired by Zheng et al. [31], Bodla et al. [38], and Ma et al. [39], this article proposes to construct a decay function for rescoring the ranking list for each query. This function can be regarded as a penalty to the tail of the initial similarity score curve since the ranking score tails of both good and inferior semantic representations are almost always false positives. The "high tail" of inferior semantic representation means a higher likelihood of being false positives, and if penalized, it will highlight the top-ranked scores. We propose to fit and normalize the initial ranking similarity score list with Gaussian penalty function $s_j^{(i)}$, and

the equation is as follows:

$$g_j^{(i)} = s_j^{(i)} e^{-\frac{\left(s_j^{(i)} - s_1^{(i)}\right)^2}{\sigma}} \tag{12}$$

where $s_j^{(i)}$ is an initial ranking similarity score list obtained by semantic representation $\mathcal{F}^{(i)}$, $j = 1, 2, \ldots, and T/V$ ($T$ indicates video-to-text retrieval, $V$ denotes text-to-video retrieval). Moreover, $s_1^{(i)}$ denotes the rank-1 result, assuming that the good or inferior similarity score curves are those in which rank-1 is a true match. Specifically, the "high tail" of inferior semantic representation is fit by the Gaussian function, which is penalized significantly. Next, the Gaussian function fitting curve $g_j^{(i)}$ is subtracted from the initial ranking similarity score curve of the query

$$\hat{s}_j^{(i)} = s_j^{(i)} - g_j^{(i)}. \tag{13}$$

Then, the fitting curve of the Gaussian function for the "high tail" penalty closely approximates the profile of the initial ranking similarity score curve, thus scores of the top-ranked texts or videos can be highlighted in the resulting curve $\hat{s}_j^{(i)}$.

*3) Adaptive Weights Estimation:* Similarities of different representation levels usually have different influences on the overall similarity, so we should adaptively assign the similarities with different attention. To improve the efficiency of the model and eliminate the influence of outliers, $\hat{s}_j^{(i)}$ necessary to undergo min–max normalization

$$\overline{s}_j^{(i)} = \frac{\hat{s}_j^{(i)} - \min \hat{s}_j^{(i)}}{\max \hat{s}_j^{(i)} - \min \hat{s}_j^{(i)}}. \tag{14}$$

After min–max normalization, $\overline{s}_q^{(j)}$ is the normalized similarity score curve that can be used to estimate the effectiveness of semantic representation. For a given query video (or text) with $K$ levels of semantic representation, we have $K$ similarity score ranking lists $\{s_j^{(i)}\}_{i=1}^K$. After normalization to $\{\overline{s}_j^{(i)}\}_{i=1}^K$, the query-adaptive weight of semantic representation $\mathcal{F}^{(i)}$ to query video (or text) can be calculated as

$$w_q^{(i)} = \frac{\frac{1}{A_i}}{\sum_{k=1}^K \frac{1}{A_k}} \tag{15}$$

where $A_i$ ($i = 1, \ldots, K$) denotes the $i$th level semantic representation under the score curve of area.

*4) Hierarchical Similarity Fusion:* Suppose that $K$ levels of semantic representation are fused, given the query $v_q$ (or $t_q$) and a dataset $d$ contains $V$ videos with $T$ text descriptions for each video, the similarity score of $d$ to $v_q$ (or $t_q$) *with respect to* semantic representation $\mathcal{F}^{(i)}$, $i = 1, \ldots, K$ is represented as $s_{d,v_q}^{(i)}$ (or $s_{d,t_q}^{(i)}$). Let $w_q^{(i)}$, $i = 1, \ldots, K$ encode the weight of semantic representation $\mathcal{F}^{(i)}$ for query $v_q$ (or $t_q$), and has a sum of 1. Then, we take the adaptive fusion of cross-modal similarity at all levels as the final matching similarity

$$\text{sim}(V, T) = \prod_{i=1}^K \left(s_{d,v_q}^{(i)}\right) w_q^{(i)}, \quad \text{where} \quad \sum_{i=1}^K w_q^{(i)} = 1 \tag{16}$$

or

$$\text{sim}(V, T) = \prod_{i=1}^K (s_{d,t_q}^{(i)}) w_q^{(i)}, \quad \text{where} \quad \sum_{i=1}^K w_q^{(i)} = 1. \tag{17}$$

Note that the weight $w_q^{(i)}$ of query-adaptive fusion is determined by (15).

## IV. Experimental Settings and Baselines

In this section, we briefly introduce the three public benchmark video–text datasets adopted in our work, experimental evaluation metrics, implementation details, and baselines.

### A. Datasets

The **MSR-VTT** [15] dataset contains 10 000 videos with 20 text descriptions for each video. We follow the standard split with 6573 videos for training, 497 for validation, and 2990 for testing.

The **TGIF** [16] dataset is composed of GIF format videos, where there are 79 451 videos for training, 10 651 for validation, and 11 310 for testing in the official split. Each video corresponds to one to three descriptive sentences.

The **VATEX** [17] dataset includes 25 991 videos for training, 3000 for validation, and 6000 for testing. There are ten sentences in English and Chinese languages to describe each video. In our work, we only utilize English annotations.

### B. Evaluation Metrics

We adopt three common metrics to measure our QAMF model: recall at $K$ ($R@K$), median rank (MedR), and mean rank (MnR). $R@K$ is the fraction of queries that correctly retrieve desired items in the top $K$ of the ranking list. And, we set $K = 1, 5, 10$, following the tradition [9], [10], [14]. MedR and MnR measure the median and average rank of correct items in the retrieved ranking list, respectively. Additionally, we also use the sum of all $R@K$ as rsum to measure the overall retrieval performance. For $R@K$ and rsum, a higher score indicates better performance, and for MedR and MnR, a lower score indicates better performance. It is the rsum of all the evaluation indicators in the model for $R@1$, $R@5$, and $R@10$

$$\text{sum} = \underbrace{R@1 + R@5 + R@10}_{\text{Text}\to\text{Video} \quad \text{or} \quad \text{Video}\to\text{Text}} \tag{18}$$

$$\text{rsum} = \underbrace{R@1 + R@5 + R@10}_{\text{Text}\to\text{Video}} + \underbrace{R@1 + R@5 + R@10}_{\text{Video}\to\text{Text}}. \tag{19}$$

### C. Implementation Details

The experiments are conducted with Ubuntu18.04, Intel[1] Core[2] i7-9700KF CPU@3.60 GHz, 64.00 GB RAM, and Nvidia GeForce RTX-2080Ti GPU. Similar to [10]: for the video encoding, we utilize the pretrained ResNet [42] to extract the visual semantic features of MSR-VTT and TGIF,

[1]Registered trademark.
[2]Trademarked.

TABLE I
VIDEO–TEXT RETRIEVAL COMPARISON WITH STATE-OF-THE-ART METHODS ON MSR-VTT DATASETS

| Methods | Text-to-Video | | | | | Video-to-Text | | | | | rsum |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | MedR | MnR | R@1 | R@5 | R@10 | MedR | MnR | |
| VSE [36] | 5.0 | 16.4 | 24.6 | 47 | 215.1 | 7.7 | 20.3 | 31.2 | 28 | 185.8 | 105.2 |
| VSE++ [22] | 5.7 | 17.1 | 24.8 | 65 | 300.8 | 10.2 | 25.4 | 35.1 | 25 | 228.1 | 118.3 |
| Mithum *et al.* [13] | 5.8 | 17.6 | 25.2 | 61 | 296.6 | 10.5 | 26.7 | 35.9 | 25 | 266.6 | 121.7 |
| W2VV [20] | 6.1 | 18.7 | 27.5 | 45 | - | 11.8 | 28.9 | 39.1 | 21 | - | 132.1 |
| DualEn [5] | 7.7 | 22.0 | 31.8 | 32 | - | 13.0 | 30.8 | 43.3 | 15 | - | 148.6 |
| HGR [10] | 9.2 | 26.2 | 36.5 | 24 | 164.0 | 15.0 | 36.7 | 48.8 | 11 | 90.4 | 172.4 |
| HCGC [14] | 9.7 | 28.0 | 39.2 | 19 | 129.5 | 17.1 | 40.5 | 53.2 | 9 | 58.2 | 187.7 |
| Our QAMF | **11.00** | **28.38** | **39.22** | **22** | **150.28** | **16.17** | **37.91** | **50.95** | **11** | **89.34** | **183.64** |

and adopt the Inflated 3D ConvNets (I3D) [43] video features provided by the VATEX dataset. For the text encoding, we set the word embedding size as 300 and initialize with pretrained Glove embeddings [44], and the dimension of embedding space for each level is 1024. In particular, if not specified, the $\sigma$ parameter in the Gaussian decay function is set to 0.5.

### D. Baselines

We conduct a comparison with some SoTA approaches.

1) *VSE [36]:* It is a SoTA cross-modal retrieval model and is also regarded as a strong baseline in the text–video or text–image retrieval tasks.
2) *VSE++ [22]:* An improved version of visual-semantic embeddings (VSE), which utilizes a novel loss based on augmented data and fine-tuning to significantly improve cross-modal retrieval performance.
3) *Mithum et al. [13]:* Adopt multimodal cues from videos and a modified pairwise ranking loss to enhance the discrimination between feature representations.
4) *W2VV [20]:* W2VV can transform natural language statements into meaningful visual feature representations, that is, the relevant video–text pairs in feature representation space will be pulled closer, while irrelevant ones will be pushed apart.
5) *DualEn [5]:* Mean pooling, biGRU, and convolutional neural network (CNN) are leveraged to realize the visual–text pairs coarse-to-fine-grained and spatial–temporal feature representations.
6) *HGR [10]:* The graph convolutional network is used to model the hierarchical representations of video and text, respectively, and the alignment of video–text pairs is implemented at three levels of visual–text embedding common space.
7) *HCGC [14]:* Multilevel graph consistency learning is leveraged to bridge the semantic gap between video-text cross-modal retrieval.

## V. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we present experimental results with the corresponding analysis on three public benchmark datasets for video-text bi-directional cross-modal retrieval. For clarifying the evaluation logic, we elaborate six research questions (RQs) as our evaluation goals in experiments as follows.

TABLE II
TEXT-TO-VIDEO RETRIEVAL COMPARISON WITH STATE-OF-THE-ART METHODS ON THE TGIF AND VATEX DATASETS

| Datasets & Methods | | R@1 | R@5 | R@10 | MedR |
|---|---|---|---|---|---|
| TGIF | DeViSE [40] | 0.8 | 3.5 | 6.0 | 379 |
| | VSE++ [22] | 0.4 | 1.6 | 3.6 | 692 |
| | Corr-AE [41] | 0.9 | 3.4 | 5.6 | 365 |
| | PVSE [4] | 2.17 | 7.76 | 12.25 | 155 |
| | HGR [10] | 4.5 | 12.4 | 17.8 | 160 |
| | HCGC [14] | 6.3 | 16.2 | 22.9 | 79 |
| | Our QAMF | **6.72** | **14.85** | **20.74** | **159** |
| VATEX | VSE [36] | 28.0 | 64.2 | 76.9 | 3 |
| | VSE++ [22] | 33.7 | 70.1 | 81.0 | 2 |
| | DualEn [5] | 31.1 | 67.4 | 78.9 | 3 |
| | HGR [10] | 35.1 | 73.5 | 83.5 | 2 |
| | Our QAMF | **38.37** | **78.14** | **88.92** | **1** |

1) *RQ1:* Is the **overall performance** of our QAMF model superior to the SoTA methods?
2) *RQ2:* Are the **hierarchical encoding** (see Section III-A), **text graph–transformer** component (see Section III-A1), and **query-adaptive fusion** component (see Section III-B) in QAMF essential and effective?
3) *RQ3:* The **decay function** plays a significant role in query-adaptive fusion of our QAMF. How to choose the fitting decay function? Why the Gaussian function is adopted?
4) *RQ4:* What are the impact of **parameter** $l$ **and** $\sigma$ in the query-adaptive fusion of our QAMF?
5) *RQ5:* What is the **generalization capability** of the proposed query-adaptive weights estimation and similarity fusion mechanism?
6) *RQ6:* What is the **qualitative performance** of the proposed bi-directional retrieval model?

**For convenience**, in the tables of experimental results, we employ abbreviations "[l-e]," "[a-r]," and "[g-e]" to represent local-entity representation level, action–relation representation level, and global-event representation level, respectively.

### A. Comparison With State-of-the-Art (RQ1)

As shown in Tables I and II, we present performance comparisons with a group of the SoTA on three public benchmark datasets. On MSR-VTT, HGR [10] is superior to DualEn [5] and other methods for each metric, and the overall

TABLE III
RESULTS ON THE ABLATION STUDY OF THE REPRESENTATION LEVELS IN HIERARCHICAL ENCODING OF QAMF. THE HIGHEST SCORE IS SHOWN IN **BOLD**

| Datasets | Different Encoding Methods | Text-to-Video | | | | | | Video-to-Text | | | | | | rsum |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | sum | MedR | MnR | R@1 | R@5 | R@10 | sum | MedR | MnR | |
| MSR-VTT | [l-e] | 6.48 | 19.76 | 28.40 | 54.64 | 38 | 173.50 | 8.33 | 23.21 | 33.91 | 65.45 | 22 | 107.02 | 120.09 |
| | [a-r] | 6.53 | 19.49 | 27.85 | 53.87 | 56 | 374.85 | 11.74 | 28.76 | 40.23 | 80.73 | 19 | 214.01 | 134.60 |
| | [g-e] | 7.15 | 21.72 | 31.72 | 60.59 | 32 | 234.50 | 11.77 | 29.93 | 41.10 | 82.80 | 17 | 231.68 | 143.39 |
| | [l-e] + [a-r] | 10.15 | 25.77 | 35.70 | 71.62 | 29 | 168.08 | 14.63 | 35.53 | 47.61 | 97.77 | 14 | 89.89 | 169.39 |
| | [l-e] + [g-e] | 10.87 | 27.40 | 38.17 | 76.44 | 23 | 145.22 | 15.23 | 36.47 | 49.05 | 100.75 | 12 | 88.37 | 177.19 |
| | [a-r] + [g-e] | 10.11 | 26.41 | 36.82 | 73.34 | 27 | 207.01 | 15.50 | 36.54 | 48.08 | 100.12 | 13 | 159.45 | 173.46 |
| | [l-e] + [a-r] + [g-e] | **11.00** | **28.38** | **39.22** | **78.60** | 22 | 150.28 | **16.17** | **37.91** | **50.95** | **105.03** | 11 | 89.34 | **183.63** |
| TGIF | [l-e] | 2.69 | 10.14 | 13.27 | 26.10 | 218 | 587.49 | 2.73 | 10.39 | 13.54 | 26.66 | 257 | 578.91 | 52.76 |
| | [a-r] | 2.91 | 9.97 | 12.84 | 25.72 | 274 | 679.47 | 3.14 | 9.86 | 13.23 | 26.23 | 293 | 669.75 | 51.95 |
| | [g-e] | 3.42 | 10.84 | 13.91 | 28.17 | 197 | 658.54 | 3.60 | 10.92 | 14.02 | 28.54 | 199 | 674.31 | 56.71 |
| | [l-e] + [a-r] | 4.17 | 11.63 | 16.17 | 31.97 | 212 | 434.85 | 4.32 | 11.72 | 16.02 | 32.06 | 195 | 378.69 | 64.03 |
| | [l-e] + [g-e] | 4.90 | 12.07 | 18.04 | 35.01 | 197 | 378.53 | 4.91 | 12.27 | 18.48 | 35.66 | 188 | 174.59 | 70.67 |
| | [a-r] + [g-e] | 5.14 | 12.71 | 18.59 | 36.44 | 154 | 399.62 | 4.94 | 12.89 | 18.81 | 36.64 | 155 | 349.72 | 73.08 |
| | [l-e] + [a-r] + [g-e] | **6.72** | **14.85** | **20.74** | **42.31** | 159 | 365.82 | **6.84** | **14.64** | **21.07** | **42.55** | 161 | 378.94 | **84.86** |
| VATEX | [l-e] | 27.95 | 54.84 | 67.36 | 150.15 | 5 | 31 | 29.03 | 56.69 | 70.74 | 156.46 | 7 | 40 | 306.61 |
| | [a-r] | 30.87 | 58.72 | 71.92 | 161.51 | 4 | 31 | 33.96 | 60.74 | 73.34 | 168.04 | 5 | 34 | 329.55 |
| | [g-e] | 32.45 | 65.94 | 76.49 | 174.88 | 4 | 26 | 35.83 | 67.62 | 79.43 | 182.88 | 3 | 24 | 357.76 |
| | [l-e] + [a-r] | 34.32 | 64.7 | 76.07 | 175.09 | 3 | 19 | 37.79 | 67.63 | 78.30 | 183.72 | 3 | 21 | 358.81 |
| | [l-e] + [g-e] | 35.42 | 69.71 | 81.84 | 186.97 | 3 | 18 | 38.94 | 71.84 | 86.15 | 196.93 | 2 | 19 | 383.90 |
| | [a-r] + [g-e] | 36.62 | 75.79 | 88.31 | 200.72 | 2 | 17 | 39.67 | 77.56 | 88.79 | 206.02 | 1 | 19 | 406.74 |
| | [l-e] + [a-r] + [g-e] | **38.37** | **78.14** | **88.92** | **205.43** | 1 | 16 | **41.25** | **80.19** | **90.27** | **211.71** | 1 | 13 | **417.14** |

retrieval quality reflected by the rsum metric is also boosted by a large margin (+23.8). In particular, our QAMF model performance is further enhanced to 183.64 in rsum, which is a significant improvement compared to DualEn and HGR. We believe the major gain comes from text graph–transformer and query-adaptive fusion, which enhances the complementarity of semantic representations at global-to-local levels. Although HGR implements hierarchical fine-grained video-text matching, it ignores semantic interactions between the text graph nodes of different levels and also fails to realize the adaptive differential fusion of multilevel semantic representation, thus not as outstanding as our QAMF in the video–text pairs for cross-modal retrieval.

To further demonstrate the strength of the QAMF on different datasets, we provide quantitative results on TGIF and VATEX in Table II. Similar to the HGR, the model utilizes ResNet image features on the TGIF and I3D video features on the VATEX. We can see that the performance of our is superior to SoTA methods. On TGIF, our QAMF yields the $R@K$ ($K = 1, 5, 10$) of 6.72, 14.85, and 20.74 respectively, when equipping the baseline HGR with text graph–transformer and query-adaptive fusion. On VATEX, the QAMF keeps the performance of 38.37, 78.14, and 88.92 in $R@K$ ($K = 1, 5, 10$), compared to 35.1, 73.5, and 83.5 of the HGR baseline. The results illustrate that it is beneficial to improve the cross-modal retrieval accuracy by combining the global-to-local in an adaptive manner and the rich semantic interactions between the text graph nodes of different levels.

### B. Ablation Study (RQ2)

Compared with the HGR baseline, our QAMF involves hierarchical encoding for deliberate representation, text graph–transformer, and query-adaptive fusion components. In this part, we conduct ablation studies on these factors.

*1) Different Combinations of Representation in Hierarchical Encoding:* The ablation study results are shown in Table III. From the evaluation results, we draw the following conclusions.

1) It is worth noting that for the top three rows of each dataset, video–text matching results via three levels (including [l-e], [a-r], and [g-e]) on three public benchmark datasets. It shows that the [g-e] representation achieves the best performance, obtaining 143.4, 56.71, and 357.76 in rsum on MSR-VTT, TGIF, and VATEX, respectively. By contrast, the [l-e] representation leads to inferior performance, which yields 120.09 and 306.61 in rsum on MSR-VTT and VATEX, respectively. Moreover, [a-r] representation results in moderate accuracy on the three datasets.
2) Under our QAMF (equipping the baseline with text graph–transformer and query-adaptive fusion) to evaluate the performance of different combinations of representation (including [l-e] + [a-r], [l-e] + [g-e], [a-r] + [g-e], and [l-e] + [a-r] + [g-e]). We conduct a series of multilevel fusion ablation studies on MSR-VTT, TGIF, and VATEX, the bottom four rows of each dataset in Table III show the detailed results. On MSR-VTT, by combining [a-r] and [l-e], the [g-e] performance is boosted to 73.34 and 76.44 in sum, respectively. Note that for [a-r] and [l-e] which have moderate performance, their combination achieves a sum of 71.62. When the three representation levels ([g-e] + [a-r] + [l-e]) are merged, the fusion still yields stable improvement. Similar results can be observed on TGIF and VATEX. It is evident that our QAMF brings consistent benefits to combinations of various semantic representation levels.

*2) On the Involvement of Text Graph–Transformer and Query-Adaptive Fusion:* The ablation study results of com-

TABLE IV

UNDER MULTILEVEL COMBINATION ([L-E] + [A-R] + [G-E]), RESULTS ON THE ABLATION STUDY OF THE TEXT GRAPH–TRANSFORMER COMPONENT AND THE QUERY-ADAPTIVE FUSION COMPONENT IN QAMF. THE HIGHEST SCORE IS SHOWN IN **BOLD**

| Datasets | Methods | Text-to-Video | | | Video-to-Text | | | rsum |
|---|---|---|---|---|---|---|---|---|
| | | sum | MedR | MnR | sum | MedR | MnR | |
| MSR-VTT | HGR | 71.90 | 24 | 164.00 | 100.50 | 11 | 90.40 | 172.40 |
| | HGR + Text Graph-Transformer | 73.46 | 23 | 178.24 | 103.56 | 11 | 99.42 | 177.02 |
| | HGR + Query-Adaptive Fusion | 75.37 | 22 | 159.04 | 104.07 | 11 | 90.29 | 179.44 |
| | HGR + Text Graph-Transformer + Query-Adaptive Fusion | **78.60** | 22 | 150.28 | **105.03** | 11 | 89.34 | **183.63** |
| TGIF | HGR | 34.70 | 160 | 364.57 | 35.64 | 158 | 148.58 | 70.34 |
| | HGR + Text Graph-Transformer | 36.29 | 178 | 336.51 | 37.19 | 152 | 167.82 | 73.48 |
| | HGR + Query-Adaptive Fusion | 38.46 | 164 | 359.42 | 39.02 | 160 | 194.78 | 77.48 |
| | HGR + Text Graph-Transformer + Query-Adaptive Fusion | **42.31** | 159 | 365.82 | **42.55** | 161 | 378.94 | **84.86** |
| VATEX | HGR | 192.10 | 2 | 19 | 202.00 | 3 | 18 | 394.10 |
| | HGR + Text Graph-Transformer | 196.52 | 2 | 19 | 206.31 | 3 | 16 | 402.83 |
| | HGR + Query-Adaptive Fusion | 200.57 | 2 | 17 | 208.41 | 2 | 15 | 408.98 |
| | HGR + Text Graph-Transformer + Query-Adaptive Fusion | **205.43** | 1 | 16 | **211.71** | 1 | 13 | **417.14** |

TABLE V

UNDER OUR QAMF TO EVALUATE THE PERFORMANCE OF MULTILEVEL FUSION (INCLUDING [L-E] + [A-R], [L-E] + [G-E], [A-R] + [G-E], AND [L-E] + [A-R] + [G-E]) VIA EMPLOYING DIFFERENT FITTING DECAY FUNCTIONS ON THE MSR-VTT DATASET. THE HIGHEST SCORE IS SHOWN IN **BOLD**

| Level Combinations | Fitting Function | Text-to-Video | | | Video-to-Text | | |
|---|---|---|---|---|---|---|---|
| | | sum | MedR | MnR | sum | MedR | MnR |
| [l-e] + [a-r] | Boltzmann Function | 70.48 | 29 | 219.74 | 97.64 | 14 | 173.46 |
| | Gompertz Function | **71.88** | 29 | 181.52 | 96.47 | 14 | 169.82 |
| | Gaussian Function | 71.62 | 29 | 168.08 | **97.77** | 14 | 89.89 |
| [l-e] + [g-e] | Boltzmann Function | **76.81** | 23 | 128.79 | 100.42 | 12 | 173.53 |
| | Gompertz Function | 76.02 | 23 | 186.52 | **101.17** | 12 | 69.27 |
| | Gaussian Function | 76.44 | 23 | 145.22 | 100.75 | 12 | 88.37 |
| [a-r] + [g-e] | Boltzmann Function | **73.49** | 27 | 226.82 | 99.73 | 13 | 119.78 |
| | Gompertz Function | 72.87 | 27 | 193.38 | 98.95 | 13 | 209.44 |
| | Gaussian Function | 73.34 | 27 | 207.01 | **100.12** | 13 | 159.45 |
| [l-e] + [a-r] + [g-e] | Boltzmann Function | 78.53 | 22 | 169.53 | **105.21** | 11 | 205.73 |
| | Gompertz Function | 78.16 | 22 | 193.27 | 104.88 | 11 | 179.80 |
| | Gaussian Function | **78.60** | 22 | 150.28 | 105.03 | 11 | 89.34 |

ponents are shown in Table IV, under multilevel combination ([l-e] + [a-r] + [g-e]) to evaluate the contributions of each component. We evaluate the performance of the proposed components (*i.e.*, text graph–transformer and query-adaptive fusion) for video–text retrieval. To extensively investigate the contributions of each component, we compare our QAMF with its four counterparts on three datasets, which are the HGR baseline and three variations of our QAMF: QAMF with text graph–transformer only (namely, HGR+Text Graph-Transformer), QAMF with query-adaptive fusion only (namely, HGR+Query-Adaptive Fusion), and full QAMF (namely, equipping the HGR with Text Graph-Transformer and Query-Adaptive Fusion). From the results, one can see that the performance of the HGR baseline without text graph–transformer or query-adaptive fusion is worse than two variations of QAMF on three datasets, which indicates that both components contribute to the structure of global-to-local hierarchical fusion video-text matching.

### C. Decay Function (RQ3)

The decay function plays a significant role in the query-adaptive fusion of our QAMF. In this section, we discuss the decay function selection and the benefits of using decay functions, respectively.

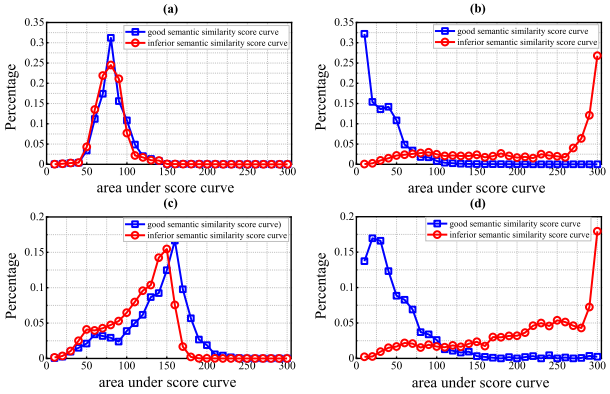


Fig. 5. Impact of Gaussian decay function. We calculate the proportion of good and inferior similarity score curves against the area under the score curve. Without the Gaussian decay function, for (a) global-event representation and (c) action–relation representation, good and inferior similarity score curves cannot be distinguished. Yet, when Gaussian fitting is subtracted, for (b) global-event representation and (d) action–relation representation, good and inferior similarity score curves are clearly separated.

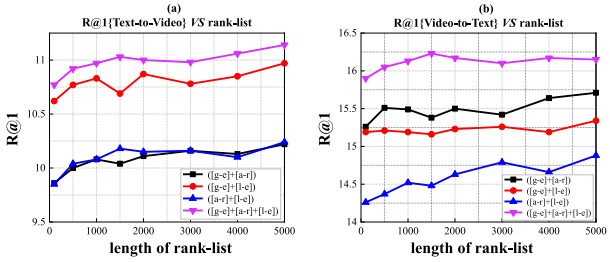


Fig. 6. Sensitivity of text-to-video retrieval and video-to-text retrieval to parameter $l$ on MSR-VTT. We test $l$ = 100, 500, 1000, 1500, 2000, 3000, 4000, and 5000 in distinct multilevel combinations (including [l-e] + [a-r], [l-e] + [g-e], [a-r] + [g-e], and [l-e] + [a-r] + [g-e]). (a) $R@1$ text-to-video versus ranklist. (b) $R@1$ text-to-video versus ranklist.

*1) Function Selection:* In Table V, under our QAMF to evaluate the performance of different levels of combination (including [l-e] + [a-r], [l-e] + [g-e], [a-r] + [g-e], and [l-e] + [a-r] + [g-e]) via employing different fitting decay

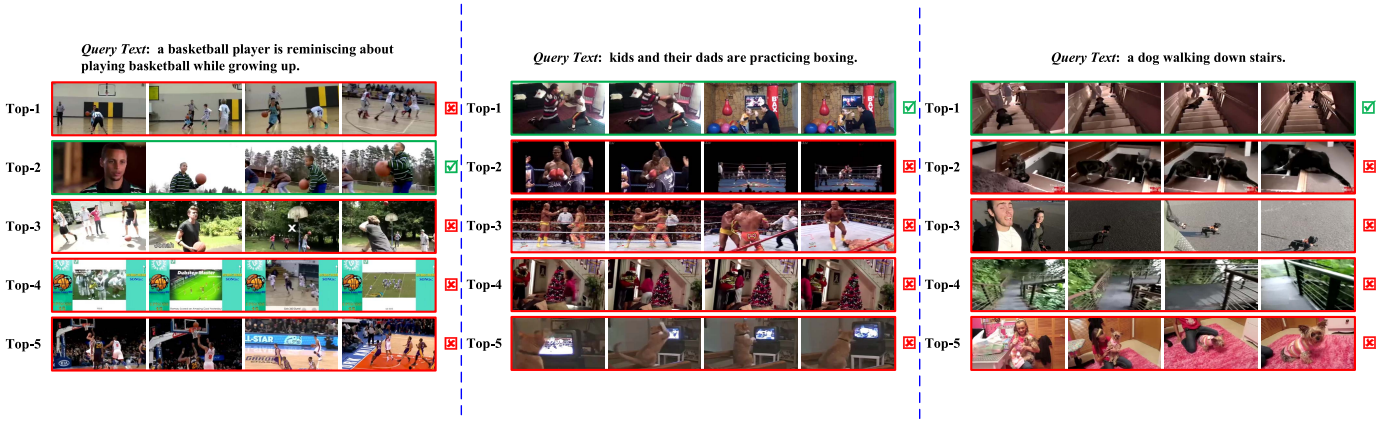


Fig. 7. Text-to-video retrieval examples on the MSR-VTT dataset. We visualize Top-5 retrieved videos. Truly matched videos are marked with a green marker, and falsely matched ones are red.

functions on MSR-VTT. Apart from the proposed Gaussian function, other functions with more parameters can also be adopted as penalty fitting functions, which fully consider removing the "high tail" problem. For example, the Boltzmann function and the Gompertz function can be used, but such functions would increase the number of parameters. As a result, we adopt the Gaussian function with fewer empirical parameter requirements, which can achieve similar results with the Boltzmann function and Gompertz function.

*2) Impact of Using Decay Function:* In order to illustrate the working mechanism of the Gaussian decay function in the query-adaptive fusion of QAMF, for global-event ([g-e]) representation and action–relation ([a-r]) representation, we have selected some good and inferior similarity score curves from MSR-VTT. Good similarity score curves are those in which rank-1 text (or video) is a true match, and inferior similarity score curves are those in which top rank is a false match. We calculate the proportion of good and inferior similarity score curves against the area under the score curve in Fig. 5. We find that after Gaussian decay fitting normalization, good queries tend to have a small area under the score curve, and vice versa. In this way, we can roughly judge the effectiveness of a level of semantic representation after Gaussian decay fitting subtraction.

### D. Impact of Parameters (RQ4)

In this section, we discuss the influence of parameters $l$ and $\sigma$ used in the decay function. Wherein, $l$ represents the length of the initial similarity score list (*i.e.*, the fitting length of the Gaussian decay function), and $\sigma$ implies the rate of decay. We test different $l$ and $\sigma$ on MSR-VTT, and the detailed experimental results are demonstrated in Fig. 6 and Table VI.

When evaluating parameter $l$, presented in Fig. 6, one can see the accuracy increases steadily with $l$. As a matter of fact, when we select the longer list of initial similarity scores (large $l$), it is more likely to find a better fit to the "high tail." Yet, the computational complexity of the area under the curve also increases with $l$. Considering this, we choose $l = 2000$ in our experiments as a tradeoff between speed and accuracy. From Table VI, compared with $l$, the performance

TABLE VI

SENSITIVITY OF TEXT-TO-VIDEO AND VIDEO-TO-TEXT TO PARAMETER $\sigma$ UNDER OUR QAMF ON MSR-VTT. WE TEST $\sigma$ IN MULTILEVEL FUSION ([L-E] + [A-R] + [G-E])

| Cross-modal Retrieval & $\sigma$ | | R@1 | R@5 | R@10 | MedR | MnR |
|---|---|---|---|---|---|---|
| Text-to-Video | $\sigma$ = 0.1 | 10.85 | 27.67 | 37.83 | 22 | 152.09 |
| | $\sigma$ = 0.3 | **11.04** | 28.07 | 39.27 | 22 | 151.26 |
| | $\sigma$ = 0.5 | 11.00 | **28.38** | 39.22 | 22 | 150.28 |
| | $\sigma$ = 0.7 | 10.97 | 28.21 | **39.69** | 22 | 149.92 |
| | $\sigma$ = 0.9 | 10.64 | 28.94 | 38.73 | 22 | 152.35 |
| | $\sigma$ = 1.1 | 10.21 | 27.79 | 37.42 | 23 | 156.73 |
| Video-to-Text | $\sigma$ = 0.1 | 15.92 | 36.82 | 50.17 | 11 | 91.36 |
| | $\sigma$ = 0.3 | 16.11 | **38.23** | 50.82 | 11 | 90.16 |
| | $\sigma$ = 0.5 | 16.17 | 37.91 | **50.95** | 11 | 89.34 |
| | $\sigma$ = 0.7 | **16.20** | 37.64 | 50.91 | 11 | 89.04 |
| | $\sigma$ = 0.9 | 16.04 | 37.20 | 50.65 | 11 | 89.59 |
| | $\sigma$ = 1.1 | 15.37 | 36.96 | 49.98 | 11 | 91.83 |

of video–text retrieval is almost less sensitive to $\sigma$. Note that recall and MnR are stable between 0.3 and 0.7 and we can always find the best performance for $R@1$, $R@5$, and $R@10$ in the range of 0.3–0.7. In all our experiments about MSR-VTT, we set $\sigma$ to 0.5, even though a $\sigma$ value of 0.7 seems to give better performance. This is because we conducted comprehensive sensitivity analysis experiments and a difference of tiny is not significant.

### E. Generalization of the Components (RQ5)

We evaluate the generalization capacity of the proposed components (text graph–transformer and query-adaptive fusion) by integrating with an SoTA baseline DualEn [5] under different fusion inputs (*i.e.*, different semantic representation combinations), including [l-e] + [a-r], [l-e] + [g-e], [a-r] + [g-e], and [l-e] + [a-r] + [g-e]. Table VII shows the performance of retrieval on MSR-VTT. Equipping the text graph–transformer and the query-adaptive fusion into DualEn can also bring improvements. The results suggest that the proposed two components (*i.e.*, text graph–transformer and the query-adaptive fusion) can better learn the alignment of local components and global event structures, which improves the generalization ability.

TABLE VII

COMPARISON OF GENERALIZATION UNDER DIFFERENT COMBINATIONS. DUALEN‡ INDICATES THAT DUALEN [5] IS EQUIPPED WITH TEXT GRAPH–TRANSFORMER AND QUERY-ADAPTIVE FUSION COMPONENTS

| Level Combinations & Methods | | Text-to-Video | | | Video-to-Text | | |
|---|---|---|---|---|---|---|---|
| | | sum | MedR | MnR | sum | MedR | MnR |
| [a-r] + [l-e] | DualEn | 52.37 | 45 | 289.60 | 78.51 | 22 | 248.75 |
| | DualEn‡ | 53.58 | 43 | 174.61 | 80.29 | 21 | 209.76 |
| [g-e] + [l-e] | DualEn | 53.64 | 43 | 264.76 | 78.90 | 22 | 294.56 |
| | DualEn‡ | 56.13 | 42 | 224.81 | 80.59 | 21 | 200.74 |
| [g-e] + [a-r] | DualEn | 52.17 | 45 | 305.51 | 79.25 | 22 | 264.93 |
| | DualEn‡ | 54.56 | 44 | 205.54 | 81.04 | 20 | 239.61 |
| [g-e] + [a-r] + [l-e] | DualEn | 61.50 | 32 | 219.74 | 87.10 | 15 | 134.94 |
| | DualEn‡ | 63.70 | 32 | 186.52 | 89.60 | 14 | 189.65 |

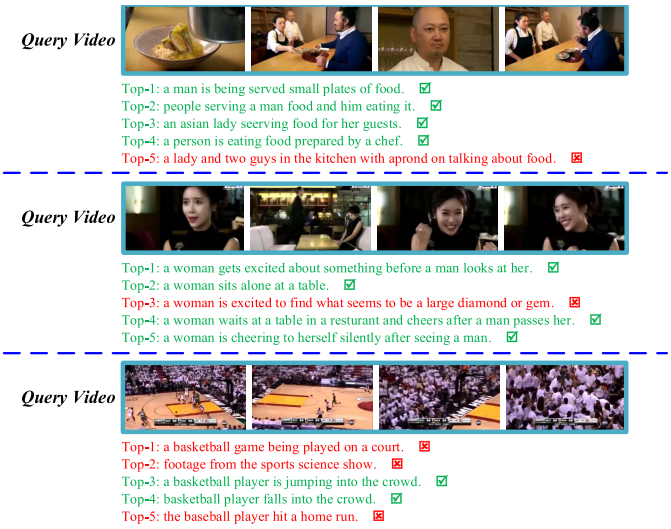


Fig. 8. Video-to-text retrieval examples on the MSR-VTT dataset and we visualize Top-5 retrieved texts. Truly matched texts are marked with a green marker, and falsely matched ones red.

### F. Qualitative Results (RQ6)

We show a few qualitative results in Figs. 7 and 8 for text-to-video and video-to-text by visualizing the retrieval results with examples from MSR-VTT. For each query, its top-5 ranked texts (or videos) resulted from our QAMF. As we can see from Fig. 7, QAMF successfully retrieves the correct video that contains all actions and entities described in the sentence in the middle and right examples. Although the left example shows a fail case, where the top-1 retrieved videos are largely relevant to the text query though are not ground-truth, QAMF can still search for the correct video in top-2. In Fig. 8, we provide qualitative results on video-to-text retrieval as well, which demonstrate the effectiveness of QAMF for bi-directional cross-modal retrieval.

## VI. CONCLUSION

The most existing multilevel fine-grained semantic representation fusion video–text retrieval takes the average of cross-modal similarities at all levels as a final video–text similarity. However, simple averaging cannot distinguish the contribution of each level to the final performance, which makes the result less than expected. Hence, this article designed a QAMF model for hierarchical fine-grained video–text retrieval. First, our QAMF estimates the effectiveness of each to-be-fused semantic representation in a query-adaptive manner. This makes ineffective semantic representation unlikely to have a negative impact on overall accuracy. Meanwhile, we leverage a text graph–transformer inference model to capture the semantic interactions between the text graph nodes of different levels. Second, our QAMF provides no extra knowledge about querying video–text pairs and evaluates the effectiveness of semantic representation only by scoring the similarity of cross-modal components at each semantic representation. Experiments on three benchmark video–text datasets demonstrate the strength of the QAMF model, and we report competitive results compared with the SoTA methods.

In terms of fusion technique, our QAMF verifies the feasibility of query-adaptive late fusion in the cross-modal video–text retrieval. In the future, we will further explore the probability distribution properties of similarity scores and the semantic representation selection strategies in fusion. In addition, our research is a two-stream model technology, namely, applying separate video and text encoders and matching video–text pairs on the final embedding space. Although this proposal achieves promising performance, however, they only achieve suboptimal results due to the lack of closer video–text interactions. Therefore, in future work, we will also explore how to employ advanced two-stream vision–language pretraining (*e.g.*, CLIP [45] and ALIGN [46]) in video–text retrieval tasks.

## REFERENCES

[1] J. Qi, Y. Peng, and Y. Yuan, "Cross-media multi-level alignment with relation attention network," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 892–898.
[2] Y. Yu, J. Kim, and G. Kim, "A joint sequence fusion model for video question answering and retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 471–487.
[3] Y. Zhang, W. Zhou, M. Wang, Q. Tian, and H. Li, "Deep relation embedding for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 30, pp. 617–627, 2021.
[4] Y. Song and M. Soleymani, "Polysemous visual-semantic embedding for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1979–1988.
[5] J. Dong et al., "Dual encoding for zero-example video retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 9346–9355.
[6] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 201–216.
[7] H. Wu et al., "Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 6609–6618.
[8] Y.-W. Zhan, X. Luo, Y. Wang, and X.-S. Xu, "Supervised hierarchical deep hashing for cross-modal retrieval," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3386–3394.
[9] M. Wray, G. Csurka, D. Larlus, and D. Damen, "Fine-grained action retrieval through multiple parts-of-speech embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 450–459.
[10] S. Chen, Y. Zhao, Q. Jin, and Q. Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10638–10647.
[11] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep multimodal fusion by channel exchanging," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1–11.

[12] B. Liu, Q. Zheng, Y. Wang, M. Zhang, J. Dong, and X. Wang, "FeatInter: Exploring fine-grained object features for video-text retrieval," *Neurocomputing*, vol. 496, pp. 178–191, Jul. 2022.
[13] N. C. Mithun, J. Li, F. Metze, and A. K. Roy-Chowdhury, "Learning joint embedding with multimodal cues for cross-modal video-text retrieval," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2018, pp. 19–27.
[14] W. Jin, Z. Zhao, P. Zhang, J. Zhu, X. He, and Y. Zhuang, "Hierarchical cross-modal graph consistency learning for video-text retrieval," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 1114–1124.
[15] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5288–5296.
[16] Y. Li et al., "TGIF: A new dataset and benchmark on animated GIF description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4641–4650.
[17] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, "VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4581–4591.
[18] S. R. Dubey, "A decade survey of content based image retrieval using deep learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2687–2704, May 2022.
[19] Y. Wang, "Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 1s, pp. 1–25, Mar. 2021.
[20] J. Dong, X. Li, and C. G. M. Snoek, "Predicting visual features from text for image and video caption retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3377–3388, Dec. 2018.
[21] L. Jin, K. Li, Z. Li, F. Xiao, G.-J. Qi, and J. Tang, "Deep semantic-preserving ordinal hashing for cross-modal similarity search," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1429–1440, May 2019.
[22] F. Faghri, D. J. Fleet, J. Ryan Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," 2017, *arXiv:1707.05612*.
[23] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7181–7189.
[24] B. Zhang, H. Hu, and F. Sha, "Cross-modal and hierarchical modeling of video and text," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 374–390.
[25] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, and H. T. Shen, "Cross-modal attention with semantic consistence for image–text matching," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5412–5425, Feb. 2020.
[26] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman, "Use what you have: Video retrieval using representations from collaborative experts," 2019, *arXiv:1907.13487*.
[27] X. Liu et al., "Late fusion incomplete multi-view clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2410–2423, Oct. 2019.
[28] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 10, pp. 1863–1883, Oct. 2019.
[29] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1224–1244, May 2017.
[30] K. Nandakumar, Y. Chen, S. C. Dass, and A. K. Jain, "Likelihood ratio-based biometric score fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 342–347, Feb. 2008.
[31] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian, "Query-adaptive late fusion for image search and person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1741–1750.
[32] J. Liu, X. Liu, S. Wang, S. Zhou, and Y. Yang, "Hierarchical multiple kernel clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 10, pp. 8671–8679.
[33] S. Wang et al., "Multi-view clustering via late fusion alignment maximization," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 3778–3784.
[34] W. Tu et al., "Deep fusion clustering network," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 11, pp. 9978–9987.
[35] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.
[36] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," 2014, *arXiv:1411.2539*.
[37] P. Shi and J. Lin, "Simple BERT models for relation extraction and semantic role labeling," 2019, *arXiv:1904.05255*.
[38] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS—Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5561–5569.
[39] W. Ma, T. Zhou, J. Qin, Q. Zhou, and Z. Cai, "Joint-attention feature fusion network and dual-adaptive NMS for object detection," *Knowl.-Based Syst.*, vol. 241, Apr. 2022, Art. no. 108213.
[40] A. Frome et al., "DeViSE: A deep visual-semantic embedding model," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.
[41] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 7–16.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[43] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.
[44] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[45] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[46] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.